Data analysis options for genotyping the Precision ID GlobalFiler NGS STR Panel v2 sequencing data from four U.S. populations

Tunde I. Huszar[1], Kevin M. Kiesler[1], Sarah Riman[1], Lisa A. Borsuk[1], Robert Lagacé[2], Katherine B. Gettings[1], Peter M. Vallone[1]

(1) National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA; (2) Thermo Fisher Scientific, 6065 Sunol Blvd., Pleasanton, CA 94556, USA

Abstract

The Precision ID GlobalFiler™ NGS STR Panel v2 (Thermo Fisher, Waltham, MA) amplifies 35 forensic markers including 31 autosomal STRs, a Y-STR and three other sex determining markers in a multiplex designed for massively parallel sequencing (MPS) applications. Here, the generated data for 519 samples across the main four U.S. populations [1] are processed through three different data analysis options following initial data acquisition via the Ion Torrent Suite. The default Applied Biosystems™ Converge software output is compared to two agnostic academic software analyses (FDSTools v1.1.1 and STRait Razor v3.0). The concordance is reported here across the different data analysis options, highlights are provided for the observed discrepancies, imbalances and locus specific artifacts related to the Precision ID GlobalFiler™ NGS STR Panel v2 sequencing panel detected in this sample set, and characteristics in data reported as observed by the individual tools.

Comparative data analysis

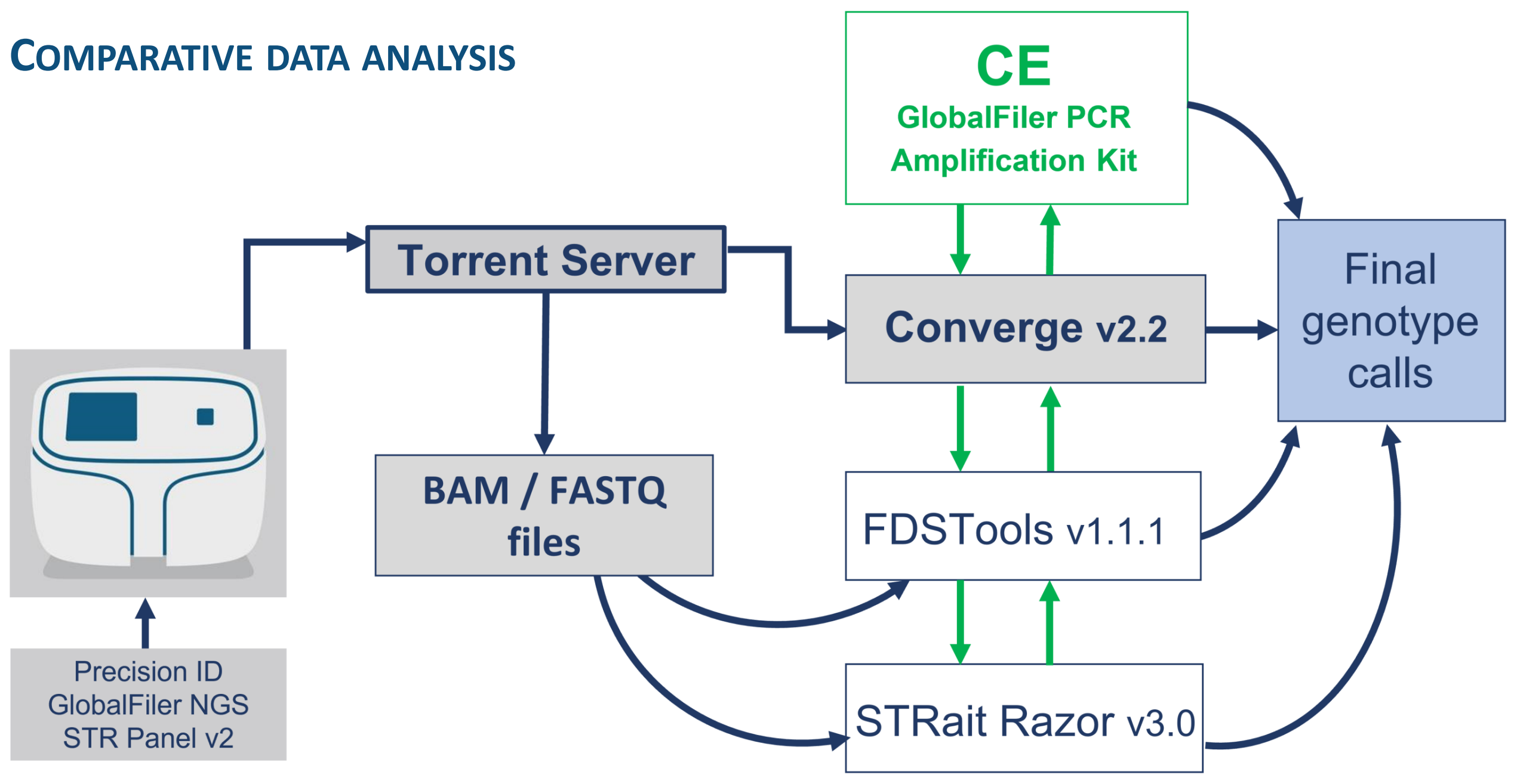


Figure 1
Schematic representation of the data analysis used for the Precision ID GlobalFiler™ NGS STR Panel v2 data. Parallel analyses to the Converge Forensic Analysis Software (Thermo Fisher) are the agnostic software FDSTools [2] and STRait Razor[3].

Samples

African American	169
Asian	123
Caucasian	129
Hispanic	98

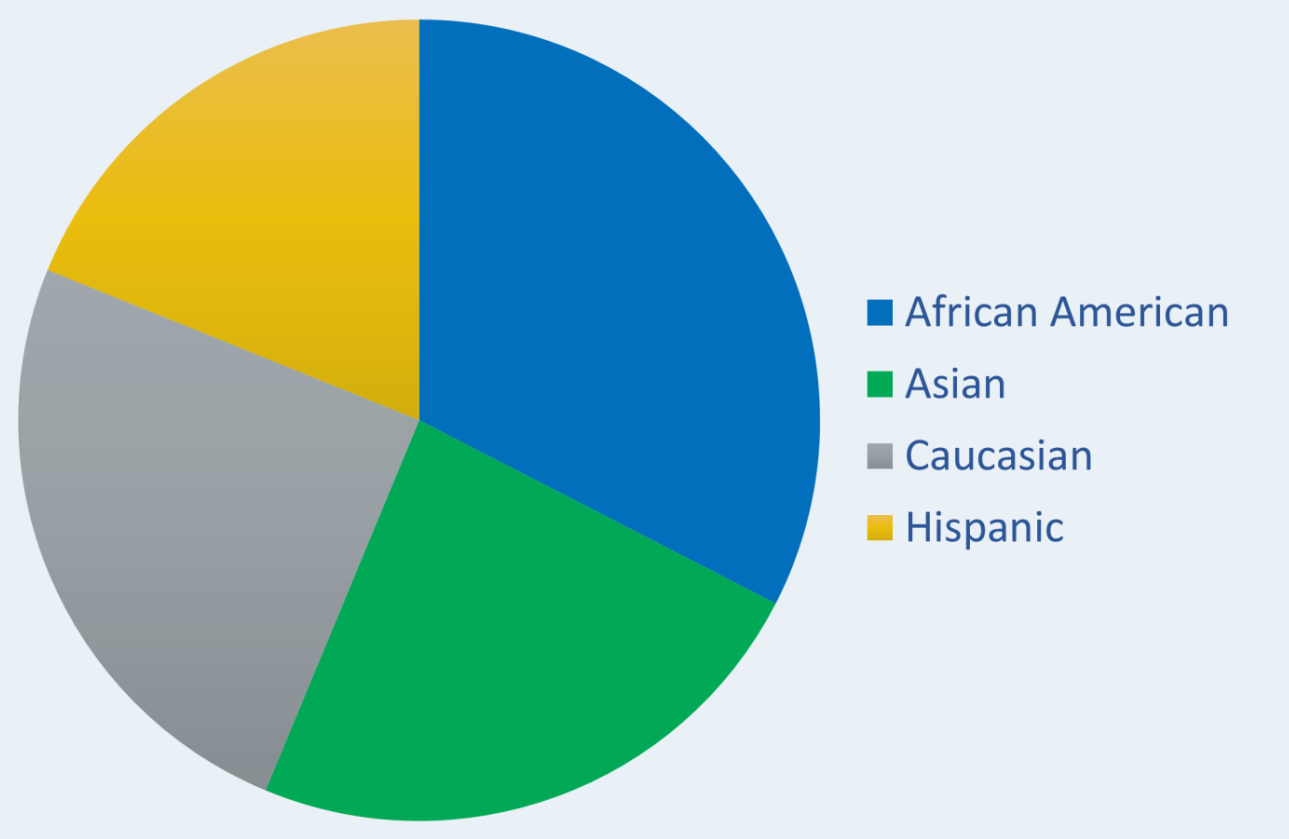
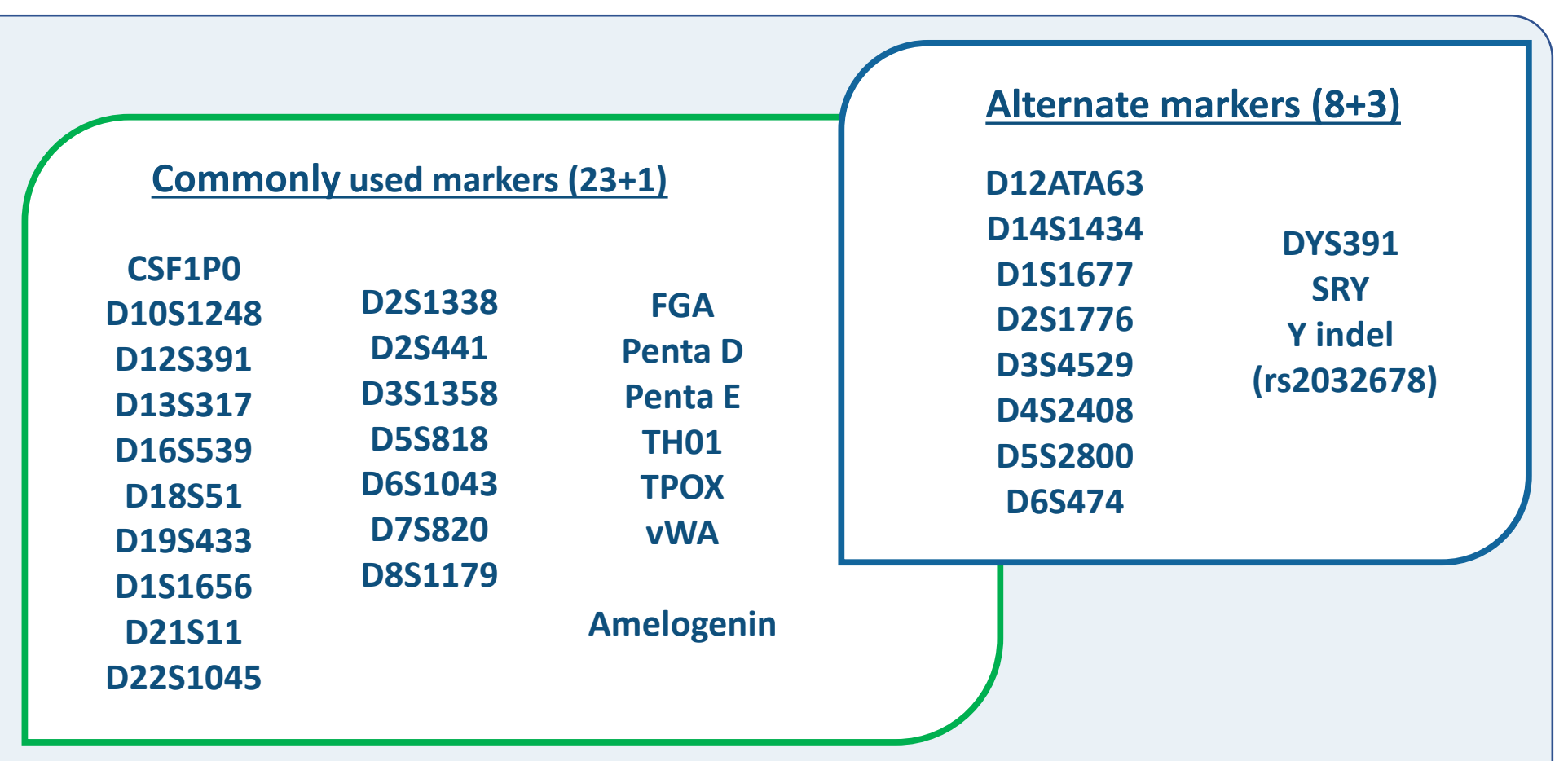


Figure 2
Distribution of the analyzed samples (n=519) across four main U.S. populations.

Markers



Commonly used markers (23+1)

CSF1P0
D10S1248
D12S391
D13S317
D16S539
D18S51
D19S433
D1S1656
D21S11
D22S1045

D2S1338
D2S441
D3S1358
D5S818
D6S1043
D7S820
D8S1179

FGA
Penta D
Penta E
TH01
TPOX
vWA

Amelogenin

Alternate markers (8+3)

D12ATA63
D14S1434
D1S1677
D2S1776
D3S4529
D4S2408
D5S2800
D6S474

DYS391
SRY
Y indel
(rs2032678)

Figure 3
The targeted markers in the Precision ID GlobalFiler™ NGS STR Panel v2 can be divided into a commonly used and an alternate set markers [4].

Discordance due to sequencing data type

Sequencing forensic markers on the Ion S5 platform generates its own artifact profile beyond the expected structure-derived stutter products and the occasional PCR or sequencing errors. The method generates sequence reads by processing changes in voltage levels due to the measurable pH change with the incorporation of nucleotides. This method is however prone to *homopolymer errors*, i.e. when the nucleotide pattern is monotonous, the accuracy of the detection of the number of individual nucleotides in a homopolymer chain is decreased. Due to the repetitive structure of the STRs this effect is excessive in markers like Penta D, Penta E or FGA, where it can impair accurate genotyping of these loci.

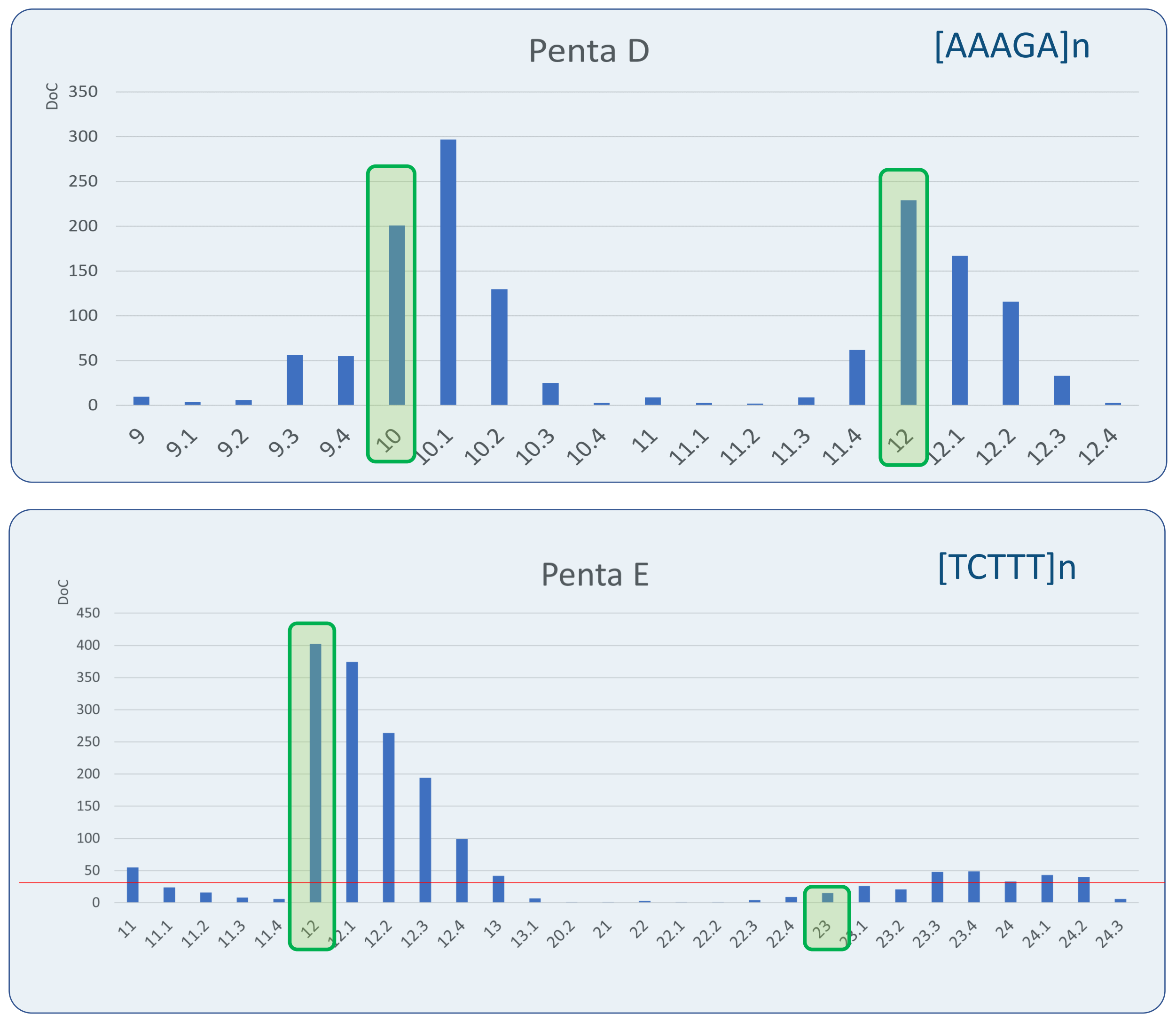


Figure 4
Representative examples of disproportionate amount of reads with homopolymer errors affecting the success of genotyping with either analysis method. The true genotypes are marked with green boxes, main repeat structure is noted in the upper right corner of each example.
In these examples multiple alleles are observed with only a single nucleotide (0.1) difference. In the first example the highest allele detected (10.1) is an artifact, while the true alleles (10 and 12) are heavily masked by a range of such artifacts.
In the second example only one true allele is detected (12), with an artifact being second highest (12.1), while the true second allele (23) is obscured below the analytical threshold (red horizontal line).

Discordance due to data analysis software

Different sequence analysis methods may use distinct ranges [5] of the sequence string targeted and depending on the placement of the recognition sequences [6] may interpret the same biological allele differently.

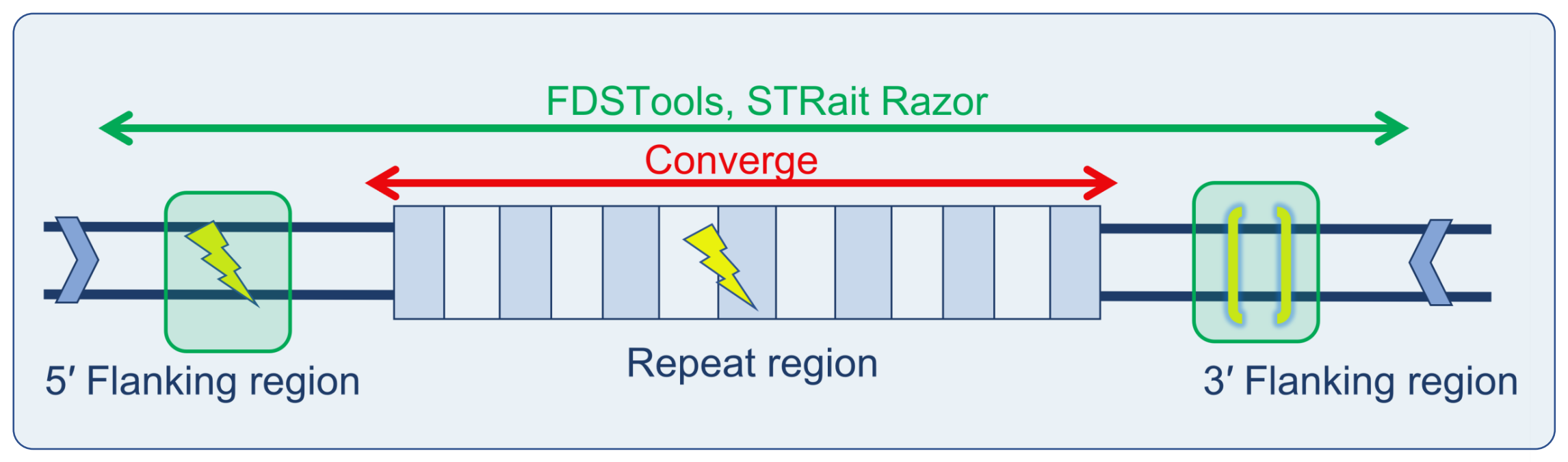


Figure 5
A schematic diagram, highlighting the source of discordances between the Converge and the agnostic software. Converge analysis range being limited to mainly the repeat region causes the software to miss indels and SNPs in the flanking regions.

Table 1
Summary of discordances between the Converge and the agnostic software in this data set.

	length					
marker	CE	sequ.	Converge	Discrepancy	Alternatives	pop(n)
D13S317	9	10	[TATC]10	4 bp deletion (3')	CE9_TATC[11]AATC[1]_+21GTCT>-	AA(2), HIS(2)
	10	11	[TATC]11	4 bp deletion (3')	CE10_TATC[12]AATC[1]_+21GTCT>-	AA(2)
	9	10	[TATC]10	1 bp insertion, 1 bp deletion (3')	CE9_TATC[11]_+0.1A_+4T>-	AA(1), CAU(1)
D14S1434	14	12	[CTGT]3 [CTAT]9	4 bp deletion (3')	CE12_CTGT[3]CTAT[10]_+8TCCA>-	HIS(1)
D18S51	14.2	14	[AGAA]14	2 bp indel, and flanking SNP (3')	CE14.2_AGAA[14]_+2A>G_+10.1->AG	AA(2)
	13.2	13	[AGAA]13	2 bp indel, and flanking SNP (3')	CE13.2_AGAA[13]_+2A>G_+10.1->AG	AA(2)
D19S433	15.2	16	AAGGTAGG [AAGG]14 *	2 bp deletion (5')	CE15.2_CCTT[14]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	CAU(1)
	13.2	14	AAGGTAGG [AAGG]12*	2 bp deletion (5')	CE13.2_CCTT[12]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	HIS(1)
D21S11	27.1	27	[TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA [TCTA]9	1 bp insertion (3')	CE27.1_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA [1]TCTA[2]TCCATA[1]TCTA[9]_+2.1->T	AA(1)
D2S1776	12	10	[AGAT]10	possible 8 bp insertion outside NGS range	CE10_AGAT[10]	HIS(1)
	14	11	[AGAT]11	possible 12 bp insertion outside NGS range	CE11_AGAT[11]	HIS(1)
	null	8	[AGAT]8	not amplified CE allele	CE8_AGAT[8]	HIS(1)
	null	9	[AGAT]9	not amplified CE allele	CE9_AGAT[9]	HIS(1)
D2S441	9.1	9	[TCTA]9	1 bp insertion in 5' flanking region,	CE9.1_TCTA[9]_0.1->A_-25G>A	ASI(4)
D3S4529	null	17	[ATCT]12 ATTT[ATCT]4*	possible variation impairing amplification of CE allele	CE17_AGAT[4]AAAT[1]AGAT[12]	ASI(1)
	null	16	[ATCT]11 ATTT[ATCT]4*	possible variation impairing amplification of CE allele	CE16_AGAT[4]AAAT[1]AGAT[11]	ASI(3)
D5S818	null	8	[AGAT]8*	SNPs in (5', 3')	CE8_ATCT[8]_-4C>A_+13A>G / [ATCT]8 rs73801920-A	AA(1)
D6S1043	18	19	[AGAT]13 ACAT[AGAT]5*	possible 4 bp insertion outside NGS range	CE19_ATCT[5]ATGT[1]ATCT[13]	ASI(1)
D6S474	14.3	15	[AGAT]5 [GATA]10	3 bp indel in repeat	CE14.3_AGAT[5]GATA[9]GAT[1]	HIS(1)
Penta D	2.2	null	null	allele not reported	CE2.2_AAAGA[5] rs1190908807	AA(36), HIS(4)
	3.2	null	null	allele not reported	CE3.2_AAAGA[6] rs1190908807	AA(6), HIS(1)
	null	12	[TCTTT]12*	not amplified CE allele	AAAGA[12]	ASI(1)
	13.4	14	[TCTTT]14	1 bp deletion (3')	CE13.4_AAAGA[14]_+9A>-	AA(1)
	9	11	[TCTTT]11	possible 10 bp deletion outside NGS range	AAAGA[11]	AA(1)
	12	14	[TCTTT]14	possible 10 bp deletion outside NGS range	AAAGA[14]	AA(1)
	14	15	[TCTTT]15	possible 5 bp deletion outside NGS range	AAAGA[15]	HIS(1)
Penta_E	19.4	20	[AAAGA]5 A[AAAGA]1 AAAA[AAAGA]13*	homopolymer artifact	allele not called	AA(1)
	19	20	[AAAGA]5 A[AAAGA]1 AAAA[AAAGA]13*	homopolymer artifact	allele not called	AA(1)
	null	20	[AAAGA]20*	not amplified CE allele	CE20_TCTTT[20]	AA(1)
	17	17.2	[AAAGA]2 A[AAAGA]1 AAAGGA[AAAGA]13*	homopolymer artifact	CE17_TCTTT[13]TCCTT[1] TCTTT[3]	HIS(1)

Discordant genotype calls, not originating from homopolymer error artifacts, were observed due to differences in reporting of flanking region variation. Indels not reported in Converge affected length-equivalent allele calls. Alternative analysis did consistently report flanking region variation between the two methods. Where available, rs# were provided. The (*) marks the loci where Converge do not report on the (+) strand [7].

NIST Disclaimer:
Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Commerce. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

All work has been reviewed and approved by the National Institute of Standards and Technology Research Protections Office. This study was determined to be "not human subjects research" (often referred to as research not involving human subjects) as defined in U. S. Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule (45 CFR 46, Subpart A), for the Protection of Human Subjects by the NIST Human Research Protections Office and therefore not subject to oversight by the NIST Institutional Review Board.

Acknowledgement

TIH was funded by the NIST Special Programs Office (Forensic Genetics Focus Area).
The authors wish to thank Thermo Fisher, Robert Lagacé, and Claus Børsting and Marie-Louise Kampmann for the assistance with the data files.

Summary

Discordances of genotypes of sequencing data compared to the length-based CE detection are unavoidable due to possibly different primer placement between kits. A novel source of discordance, the bioinformatically derived difference in genotypes, could be detected using parallel independent analysis methods.

References:

1. Gettings et al.2018. Forensic Sci. Int. Genet. 2018, 37, 106-115.
2. Hoogenboom et al. 2017. Forensic Sci. Int. Genet. 2017, 27, 27–40.
3. Woerner et al. 2017. Forensic Sci. Int. Genet. 2017, 30, 18–23.
4. Gettings et al. 2017. Forensic Sci. Int. Genet. 2017, 31, 111–117.
5. Gettings et al.2019. Forensic Sci. Int. Genet. 2019, 43, 102165.
6. Huszar et al. 2021. Genes, 2021, 12, 12111739.
7. Parson et al. 2016. Forensic Sci. Int. Genet. 2016, 22, 54-63.