Sequence Variation Observed in 27 Y-STR Markers with U.S. Population Samples

Carolyn R. Steffen, Katherine B. Gettings, Kevin M. Kiesler, Lisa A. Borsuk, and Peter M. Vallone

U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

Email: Becky.Steffen@nist.gov

P145

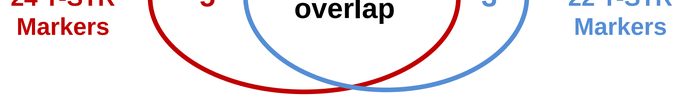
Sequencing short tandem repeat (STR) markers on the Y chromosome allows for characterization of repeat motif variations within the Y-STR marker that cannot be determined by length-based fragment analysis with capillary electrophoresis (CE). Two commercial sequence-based assays have been run at the U.S. National Institute of Standards and Technology (NIST) with U.S. population samples (Caucasian, Hispanic, African American, and Asian). The ForenSeq DNA Signature Prep Kit [1] and prototype PowerSeq 46GY System contain a total of 27 Y-STR loci between the two kits (DYF387S1, DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS460, DYS481, DYS505, DYS522, DYS533, DYS549, DYS570, DYS576, DYS612, DYS635, DYS643, and Y-GATA-H4). Data analysis was performed using an in-house pipeline based on open source software [2]. The resulting data reveals the degree of information gained through sequencing Y-STR loci and will be illustrated per population for the 27 loci. Allele calls from corresponding CE data were compared to sequence data for concordance. Discordant allele calls were further investigated to assess underlying causes such as null alleles, imbalanced peaks in multi-copy loci, flanking region insertions/deletions (indels), copy number variations, high stutter, and artifacts.

Samples Sequenced

NIST Male U.S. Population Samples

- **1032 males (ForenSeq DNA Signature Prep Kit)**
  - 359 Caucasians, 341 African Americans, 236 Hispanics, and 96 Asians
- **656 males (Prototype PowerSeq 46GY System)**
  - Subset of NIST 1036
  - 259 Caucasians, 256 African Americans, 138 Hispanics, and 3 Asians

27 Y-STR Markers Tested



- Complete CE profiles were generated for 29 autosomal STRs [3] + PowerPlex Y23 [4]
  - Examined with multiple kits and in-house primer sets enabling concordance
  - Yfiler Plus was typed across 656 male population samples
- Complete sequences with 27 autosomal STRs [5]

Methods Used for Sequencing

MiSeq FGx Instrument (Verogen)

ForenSeq DNA Signature Prep Kit (Verogen)

Prototype PowerSeq 46GY System (Promega)

ForenSeq Prep Kit	Tag and amplify targets
ForenSeq Prep Kit	Enrich and purify indexed libraries
ForenSeq Prep Kit	Normalize and pool libraries
MiSeq FGx Instrument	Perform automated cluster generation and DNA sequencing
ForenSeq Software	Analyze, view, and report data

Data Analysis was performed using an in-house pipeline based on open source software [2]

Data Analysis:

UNT HEALTH SCIENCE CENTER STRait Razor v2

27 Y-STR Loci Examined by Sequence

Characteristics for 27 Y-STR Loci

Probability of Identity (P_I) is displayed as percentage (%)

Y-STR Marker	Position (Mb) GRCh38 (Dec 2013)	Repeat Motif (forward strand)	# Observed Alleles from Sequencing	Probability of Identity (P_I)
*DYS393	3.1	$[AGAT]_n$	9	35.9
DYS505	3.6	$[TCCT]_n$	9	32.3
*DYS456	4.3	$[AGAT]_n$	8	31.1
DYS570	6.9	$[TTTC]_n$	21	21.5
DYS576	7.1	$[AAAG]_n$	10	19.4
DYS522	7.5	$[ATAG]_n$	8	33.5
*DYS458	7.9	$[GAAA]_n$	18	22.8
DYS481	8.4	$[CTT]_n$	23	15.9
DYS19	9.5	$[TCTA]_3CCTA[TCTA]_3$	9	30.2
DYS612	13.6	$[CCT]_3CTT[TCT]_4CCT[TCT]_n$	33	14.4
DYS391	14.1	$[TCTA]_n$	8	47.5
DYS635	14.4	$[TAGA]_3[TACA]_3[TAGA]_2[TACA]_2[TAGA]_4$	28	20.3
DYS437	14.5	$[TCTA]_3[TCTG]_2[TCTA]_4$	16	36.8
DYS439	14.5	$[GATA]_n$	7	35.1
DYS389 I	14.6	$[TAGA]_4[CAGA]_3$	6	42.5
DYS389 II	14.6	$[TAGA]_4[CAGA]_3N_{48}[TAGA]_4[CAGA]_{11}$	58	10.9
DYS438	14.9	$[TTTTC]_n$	9	28.6
DYS390	17.3	$[TAGA]_4CAGA[TAGA]_3[CAGA]_n$	26	20.8
DYS643	17.4	$[CTTTT]_n$	11	23.8
DYS533	18.4	$[TATC]_n$	8	36.7
GATA-H4	18.7	$[TCTA]_n$	7	39.8
DYS460	18.9	$[CTAT]_n$	8	41.4
DYS385 a/b	20.8	$[TTTC]_n$	30	13.9
DYS549	21.5	$[GATA]_n$	9	31.1
DYS392	22.6	$[ATA]_n$	10	38.0
DYS448	24.4	$[AGAGAT]_2N_{32}[AGAGAT]_2N_{14}[AGAGAT]_n$	30	20.3
DYF387S1	25.9	$[AAAG]_2GTAG[GAAG]_4[AAAG]_2GAAG[AAAG]_2[GAAG]_4[AAAG]_n$	59	7.7

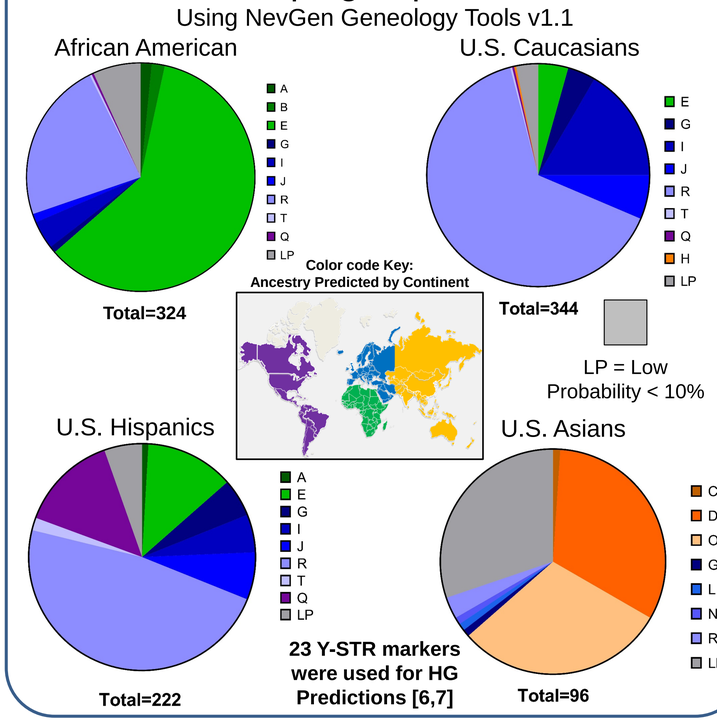
Y-STR Locus Variability

Y-STR Locus	Alleles Observed	Total (P_I) Listed Lowest to Highest	PPY23 Rank	African American (P_I)	U.S. Asian (P_I)	U.S. Caucasian (P_I)	U.S. Hispanic (P_I)
DYF387S1	59	7.7	N/A	9.0	7.0	10.7	9.3
DYS389 II	58	10.9	9	11.9	7.3	22.1	11.3
DYS385 a/b	30	13.9	1	13.1	11.3	21.6	15.2
DYS612	33	14.4	N/A	14.9	14.1	16.5	15.1
DYS481	23	15.9	2	13.2	16.7	25.9	18.7
DYS576	10	19.4	3	19.3	24.4	23.2	20.2
DYS635	28	20.3	8	25.2	24.1	29.4	21.7
DYS448	30	20.3	10	22.9	15.2	36.7	23.2
DYS390	26	20.8	6	30.4	19.2	29.3	33.7
DYS570	21	21.5	4	20.9	19.3	25.7	19.7
*DYS458	18	22.8	5	24.6	55.6	24.1	21.0
DYS643	11	23.8	7	21.4	26.4	38.3	32.9
DYS438	9	28.6	11	43.4	41.6	39.4	30.3
DYS19	9	30.2	12	27.7	24.3	49.0	32.9
DYS549	9	31.1	13	33.4	36.1	36.5	33.5
*DYS456	8	31.1	14	35.9	44.0	26.6	30.1
DYS505	9	32.3	N/A	33.7	27.5	40.6	34.7
DYS522	8	33.5	N/A	40.5	37.1	36.3	32.4
DYS439	7	35.1	15	35.4	40.9	37.1	31.3
*DYS393	9	35.9	22	32.1	37.5	40.0	40.2
DYS533	8	36.7	16	37.4	42.3	41.8	38.6
DYS437	16	36.8	18	49.3	55.7	42.9	37.9
DYS392	10	38.0	17	55.8	36.4	40.7	32.0
Y-GATA-H4	7	39.8	19	38.6	42.0	43.3	39.8
DYS460	8	41.4	N/A	41.4	34.5	46.3	43.4
DYS389 I	6	42.5	20	48.6	33.0	46.0	38.4
DYS391	8	47.5	21	56.2	69.7	44.1	44.1

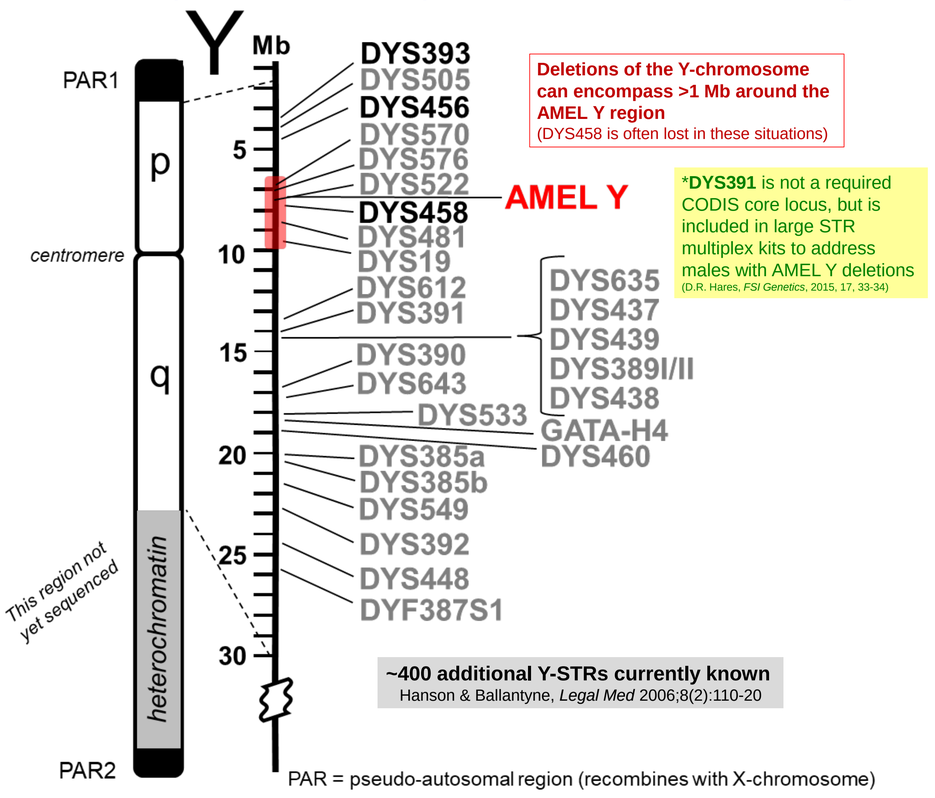
***Markers in PowerSeq 46GY only have data for 656 male samples**

Y-STR Haplogroup Predictions

Using NevGen Geneology Tools v1.1



Relative positions of 27 Y-STR loci available in ForenSeq and PowerSeq 46GY kits for ChrY testing



Deletions of the Y-chromosome can encompass >1 Mb around the AMEL Y region
(DYS458 is often lost in these situations)

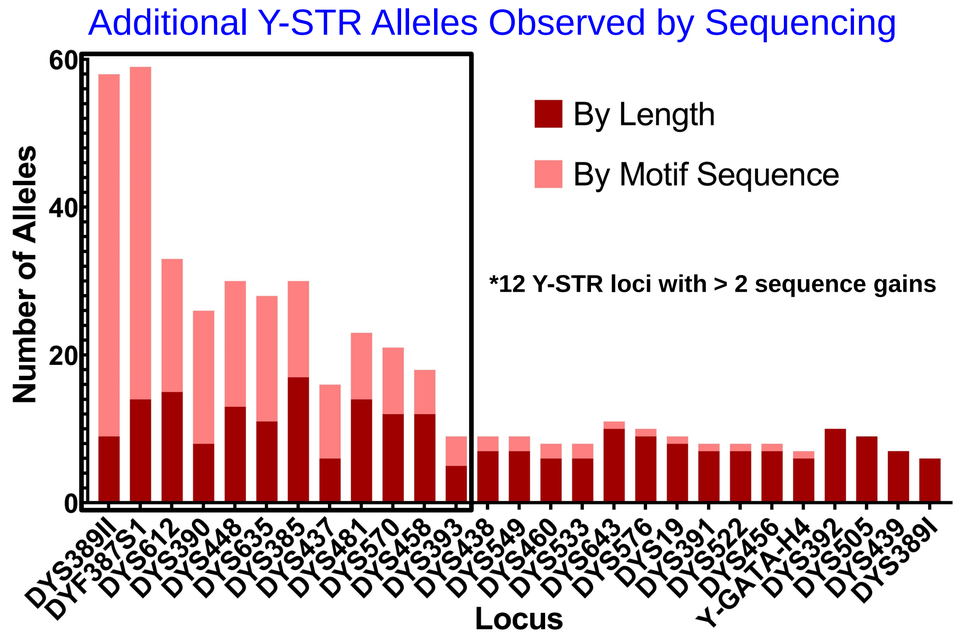
***DYS391** is not a required CODIS core locus, but is included in large STR multiplex kits to address males with AMEL Y deletions
(D.R. Hares, *FSI Genetics*, 2015, 17, 33-34)

~400 additional Y-STRs currently known
Hanson & Ballantyne, *Legal Med* 2006;8(2):110-20

PAR = pseudo-autosomal region (recombines with X-chromosome)

Chromosomal location of each Y-STR marker was determined by using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) and the Dec 2013 assembly of the human genome (GRCh38/hg38).

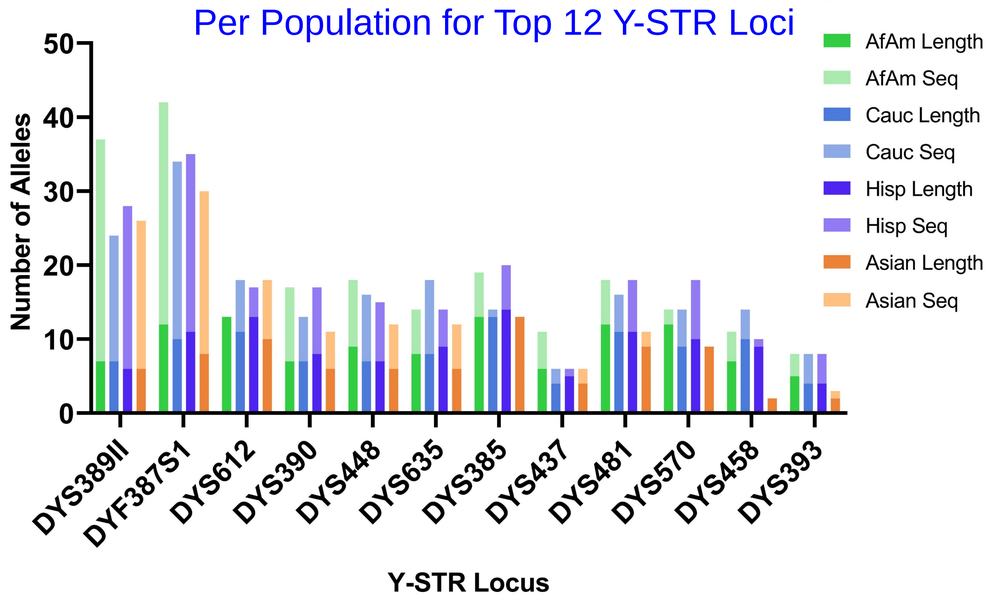
Additional Y-STR Alleles Observed by Sequencing



***12 Y-STR loci with > 2 sequence gains**

***DYS393, DYS456, and DYS458 are from PowerSeq 46GY and were only run with 656 male population samples**

Additional Alleles Observed by Sequencing Per Population for Top 12 Y-STR Loci



Y-STR Alleles Observed By Length & Sequence Per Population

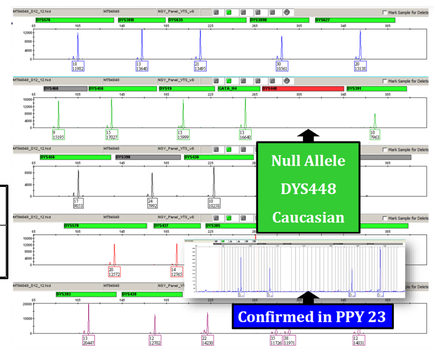
Y- STR Locus	Total		African American		U.S. Caucasian		U.S. Hispanic		U.S. Asian	
	Length	Sequence	Length	Sequence	Length	Sequence	Length	Sequence	Length	Sequence
DYS389II	9	49	7	30	7	17	6	22	6	20
DYF387S1	14	45	12	30	10	24	11	24	8	22
DYS612	15	18	13	0	11	7	13	4	10	8
DYS390	8	18	7	10	7	6	8	9	6	5
DYS448	13	17	9	9	7	9	7	8	6	6
DYS635	11	17	8	6	8	10	9	5	6	6
DYS385	17	13	13	6	13	1	14	6	13	0
DYS437	6	10	6	5	4	2	5	1	4	2
DYS481	14	9	12	6	11	5	11	7	9	2
DYS570	12	9	12	2	9	5	10	8	9	0
DYS548	12	6	7	4	10	4	9	1	2	0
DYS393	5	4	5	3	4	4	4	4	2	1
DYS456	9	2	7	0	7	0	7	2	6	0
DYS438	7	2	6	0	5	1	6	1	6	1
DYS549	7	2	6	1	7	1	6	0	5	0
DYS460	6	2	6	1	4	0	4	0	5	1
DYS533	6	2	6	1	6	0	6	0	4	1
DYS643	10	1	9	0	7	1	8	0	8	0
DYS576	9	1	9	1	9	0	8	0	7	0
DYS19	8	1	6	1	6	1	6	0	6	0
DYS391	7	1	5	0	5	1	5	0	4	0
DYS522	7	1	6	0	6	1	5	0	6	0
Y-GATA-H4	6	1	6	0	6	0	5	0	3	1
DYS392	10	0	7	0	7	0	7	0	5	0
DYS505	9	0	8	0	5	0	7	0	7	0
DYS439	7	0	6	0	6	0	5	0	5	0
DYS389I	6	0	4	0	5	0	4	0	5	0

Observations and Examples from CE- or Sequence-Based Typing

Category	CE Observations (Obs)	Sequence Observations (Obs)
Null Alleles	1	1
Peak Height Imbalance	1	1
Flank Indel	3	0
Copy Number Variation	5	12
High Stutter	5	12
Artifacts	0	5

Null Alleles: DYS448

Null Allele observed in:
ForenSeq
PP Y23
Yfiler Plus

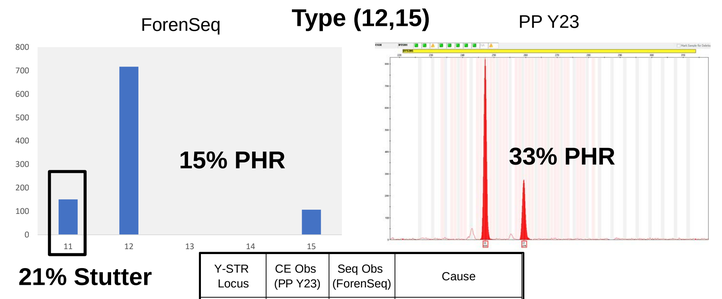


Y-STR Locus	CE Obs (PP Y23 & YF +)	Seq Obs (ForenSeq)	Cause
DYS448	1	1	possible deletion or primer binding site mutation, Sanger Seq needed

Peak Height Imbalance: DYS385a/b

Multi-Copy Y-STR Locus

Type (12,15)

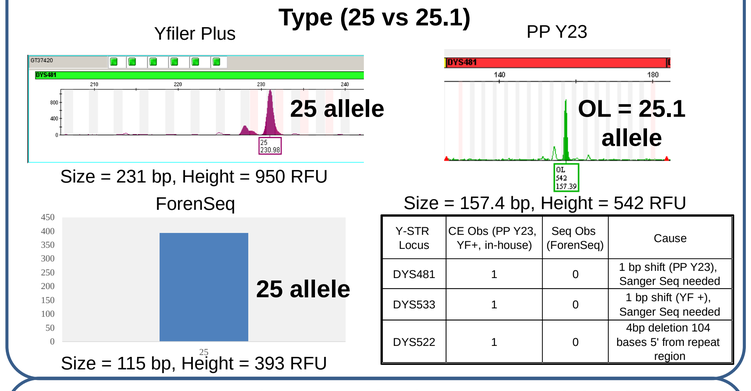


21% Stutter

Y-STR Locus	CE Obs (PP Y23)	Seq Obs (ForenSeq)	Cause
DYS385a/b	1	1	possible primer binding site mutation, Sanger Seq needed

Sequence Flank Indel: DYS481

Type (25 vs 25.1)

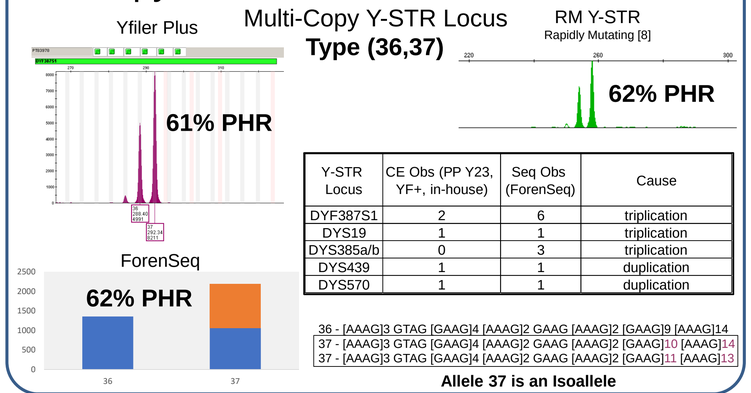


Y-STR Locus	CE Obs (PP Y23, YF+, in-house)	Seq Obs (ForenSeq)	Cause
DYS481	1	0	1 bp shift (PP Y23), Sanger Seq needed
DYS533	1	0	1 bp shift (YF +), Sanger Seq needed
DYS522	1	0	4bp deletion 104 bases 5' from repeat region

Copy Number Variation: DYF387S1

Multi-Copy Y-STR Locus

Type (36,37)



Y-STR Locus	CE Obs (PP Y23, YF+, in-house)	Seq Obs (ForenSeq)	Cause
DYF387S1	2	6	triplication
DYS19	1	1	triplication
DYS385a/b	0	3	triplication
DYS439	1	1	duplication
DYS570	1	1	duplication

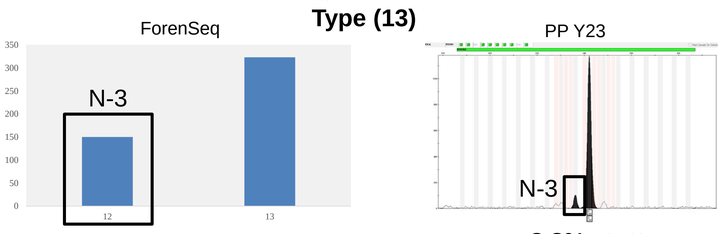
36 - [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]14
37 - [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]14
37 - [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]11 [AAAG]13

Allele 37 is an Isoallele

High Stutter: DYS392

$[ATA]_n$ trinucleotide

Type (13)



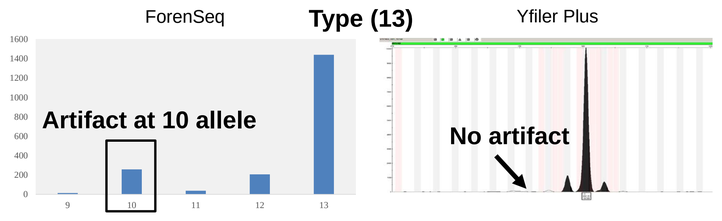
46.4% stutter

8.8% stutter

Y-STR Locus	CE Obs (PP Y23, YF+, in-house)	Seq Obs (ForenSeq)	Cause
DYS392	0	7	n-3 stutter
DYS439	1	1	n+4 stutter
DYS612	2	2	n-3 and n+3 stutter
DYS635	1	1	N+4 stutter
DYS643	1	1	n+4 stutter

Artifacts: DYS392

Type (13)



Allele	ForenSeq Coverage	Yfiler Plus RFU
9	14	ND
10	258	ND
11	37	ND
12	206	1140
13	1440	10089

Y-STR Locus	CE Obs (PP Y23, YF+, in-house)	Seq Obs (ForenSeq)	Cause
DYS392	0	4	sequence artifact
DYS505	0	1	sequence artifact

Conclusions

- Sequencing adds additional alleles for 24 of 27 Y-STR markers ranging from one allele (six Y-STR markers) to 49 alleles (DYS389II).
- Additional alleles from sequencing lowered the probability of identity (P_I) for nine Y-STR markers as compared to CE typing (DYS393, DYS389 II, DYS390, DYS391, DYS437, DYS448, DYS481, DYS570, and DYS635).
- DYF387S1, DYS612, DYS570, and DYS576 are rapidly mutating (RM) Y-STR loci [8] that have added 73 additional alleles collectively with sequencing.
- This information will be published in a forthcoming manuscript (in preparation).

References:
1. Verogen, ForenSeq DNA Signature Prep Reference Guide, Document #VD2018005, Rev. A (June 2018).
2. Warshauer, D.H., King, J.L., Budowle, B. (2015) STRait Razor v2.0: The improved STR Allele Identification Tool – Razor, Forensic Sci Int Genet. 14:182-186.
3. Steffen C.R., Coble M.D., Gettings K.B., Vallone P.M. (2017) Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci' [Forensic Sci. Int. Genet. 7 (2013) e82-e83]. Forensic Sci Int Genet. 31:e36-e40.
4. Coble, M.D., Hill, C.R., Butler J.M. (2013) Haplotype data for 23 Y-chromosome markers in four U.S. population groups. Forensic Sci. Int. Genet. 7: e66-e68.
5. Gettings K.B., Borsuk L.A., Steffen C.R., Kiesler K.M., Vallone P.M. (2018) Sequence-based US population data for 27 autosomal STR loci. Forensic Sci Int Genet 37:106-115.
6. Kivisild T. (2017) The study of human Y chromosome variation through ancient DNA. Human Genetics 136:529-546.
7. Lao O., Vallone P.M., Coble M.D., et al. (2010) Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. Human Mutation 31:E1875-E1893.
8. Ballantyne K.N., Keerl V., Wollstein, A., et al. (2012) A new future of forensic Y-chromosome analysis: Rapidly mutating &-STRs for differentiating male relatives and paternal lineages. Forensic Sci. Int. Genet. 6: 208-218.

Funding: NIST Special Programs Office: Forensic DNA

Acknowledgements: We would like to thank Doug Storts and the Promega Genetic Identity Team for providing us the prototype PowerSeq 46GY System for this study.

Poster available for download from STRBase: https://strbase.nist.gov/NISTpub.htm#Presentations