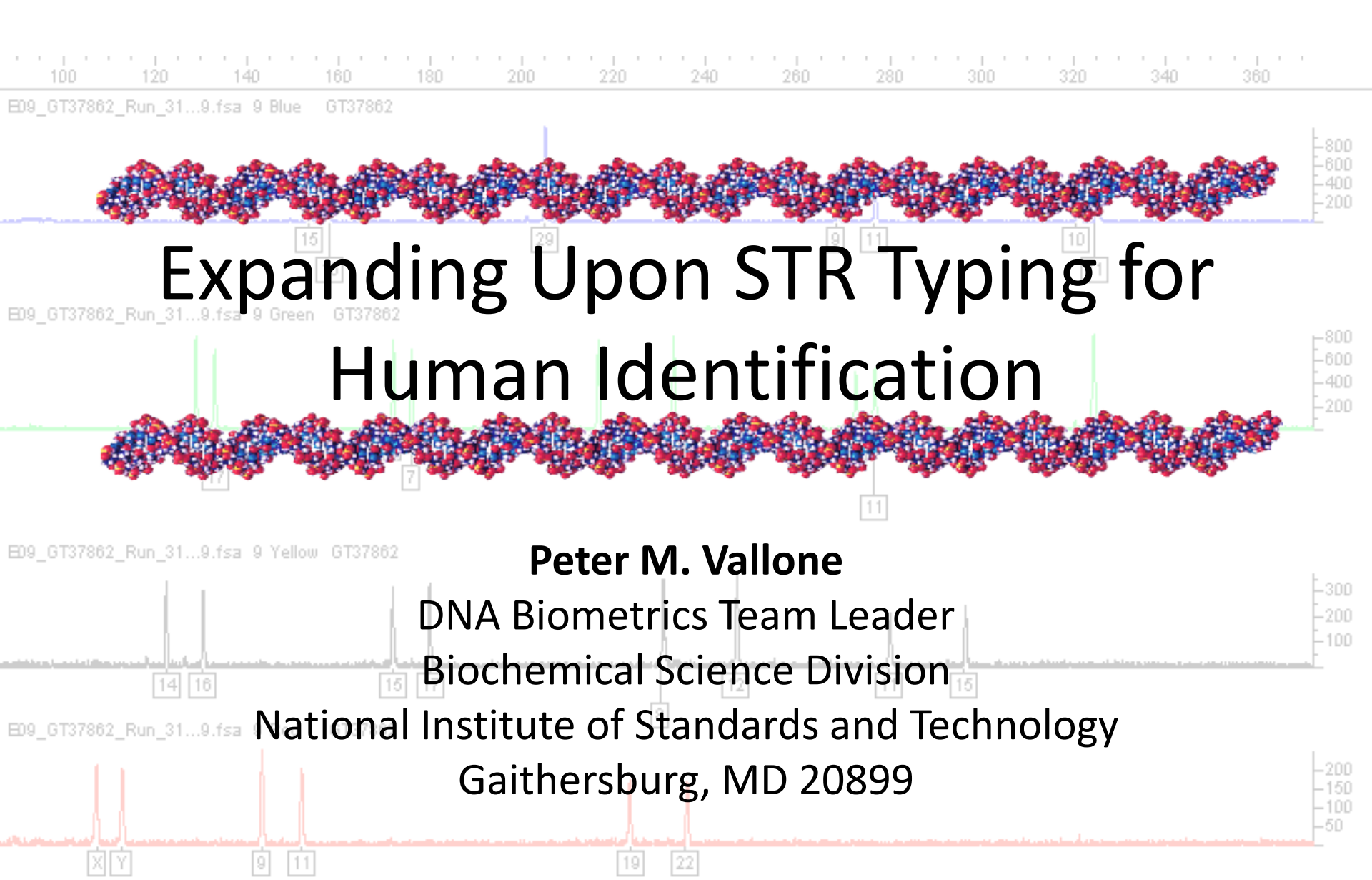


Expanding Upon STR Typing for Human Identification

Peter M. Vallone
DNA Biometrics Team Leader
Biochemical Science Division
National Institute of Standards and Technology
Gaithersburg, MD 20899

Types of Genetic Variation

- Length Variation

  short tandem repeats (STRs)

CTAGTCGT**[GATA][GATA][GATA]**GCGATCGT

- Sequence Variation

  single nucleotide polymorphisms (SNPs)

  insertions/deletions

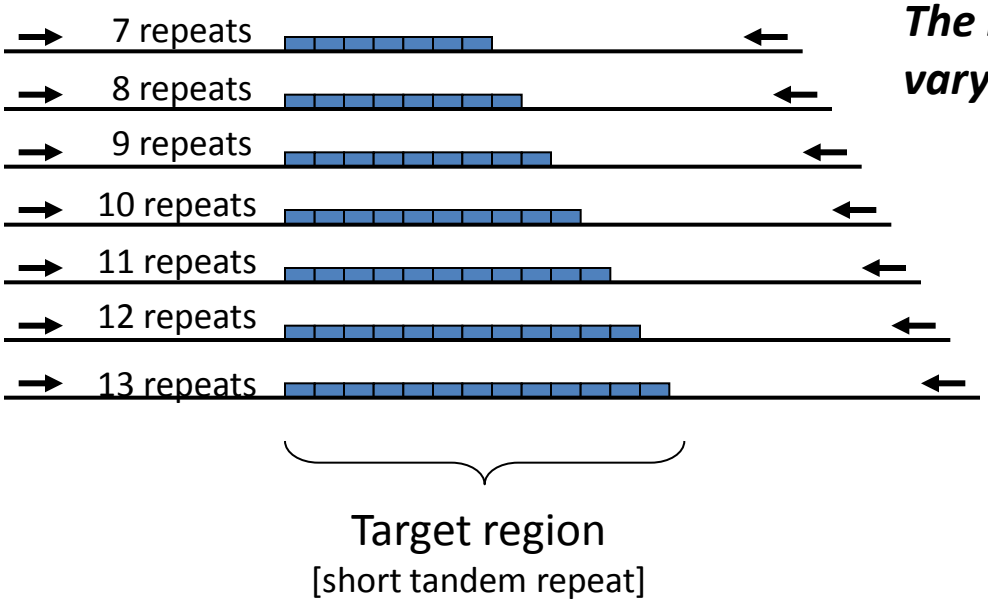
GCTAGTCGATGCTC**[G/A]**GCGTATGCTGTAGC

Also copy number variation, methylation, inversions…

Short Tandem Repeat (STR) Markers

An accordion-like DNA sequence that occurs between genes

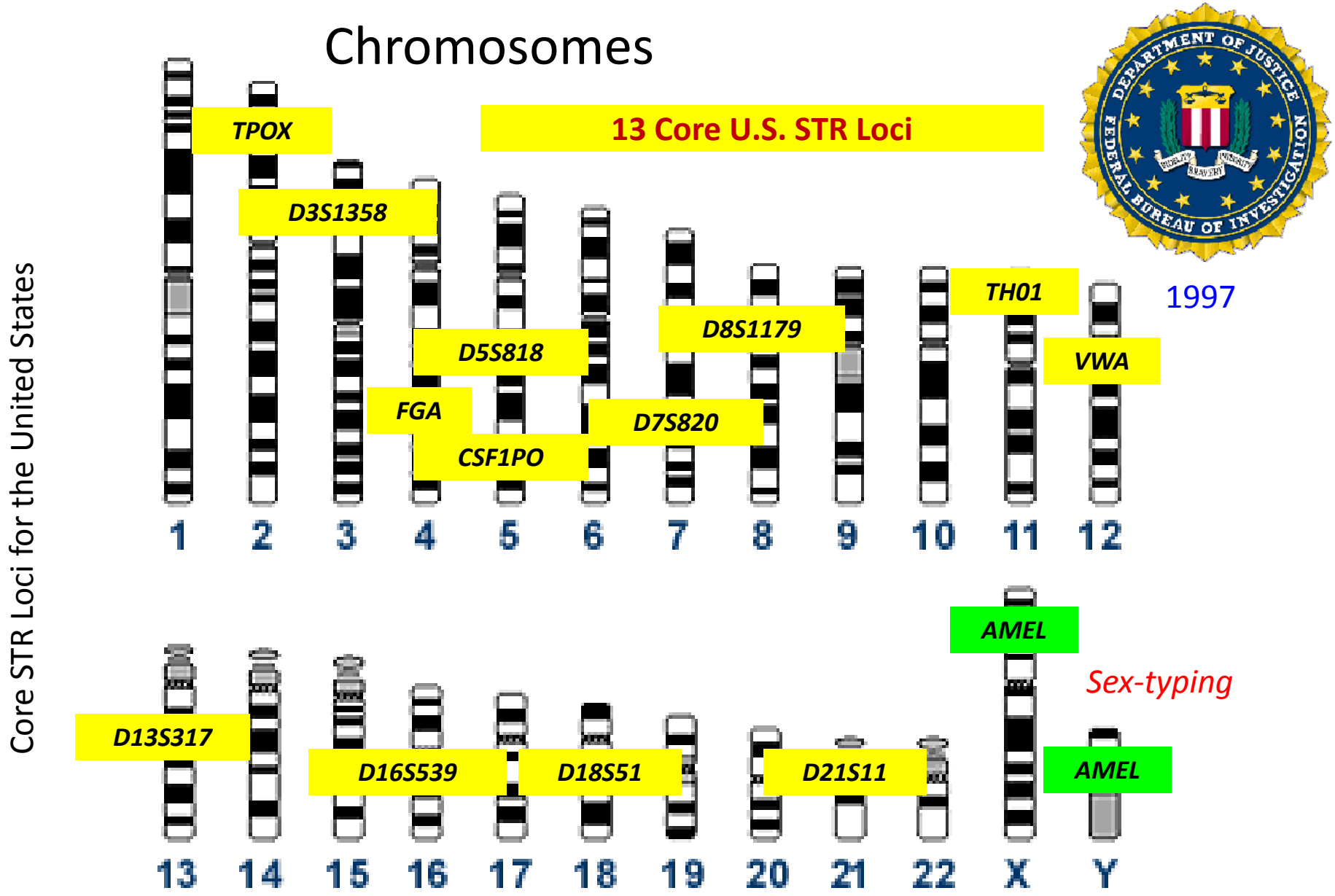
TCCCAAGCTCTTCCTCTTCCCTAGATCAATACAGACAGAAGACAGG
TGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA
GATATCATTGAAAGACAAAACAGAGATGGATGATAGATACATGCTT
ACAGATGCACAC **= 12 GATA repeats ("12" is reported)**



The number of consecutive repeat units can vary between individuals

The frequency of these repeats observed in the general population have been sampled and are used for the statistical representation of a DNA profile

Position of Forensic STR Markers on Human Chromosomes



STR Typing – Fragment Analysis

- Extract DNA from sample
- Quantitiate DNA
- PCR amplify DNA (multiplex PCR)
- Separate PCR products (electrophoresis)
- Assign alleles to peaks based on size
- Generally insensitive to sequence variations within the repeat or entire PCR product

Steps in Forensic DNA Analysis

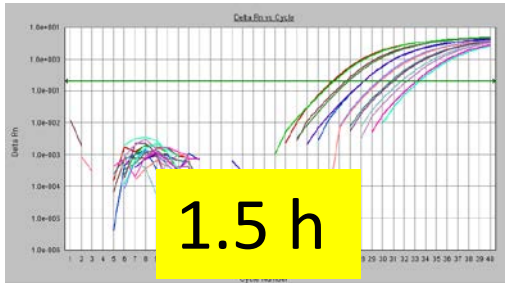
Usually 1-2 day process (a minimum of ~8 hours)

Blood Stain Buccal swab

Sample Collection & Storage

Biology

DNA Extraction



DNA Quantitation

Multiplex PCR Amplification

DNA separation and sizing

STR Typing

Technology

Interpretation of Results

Genetics

Statistics Calculated

DNA Database search
Paternity test
Reference sample

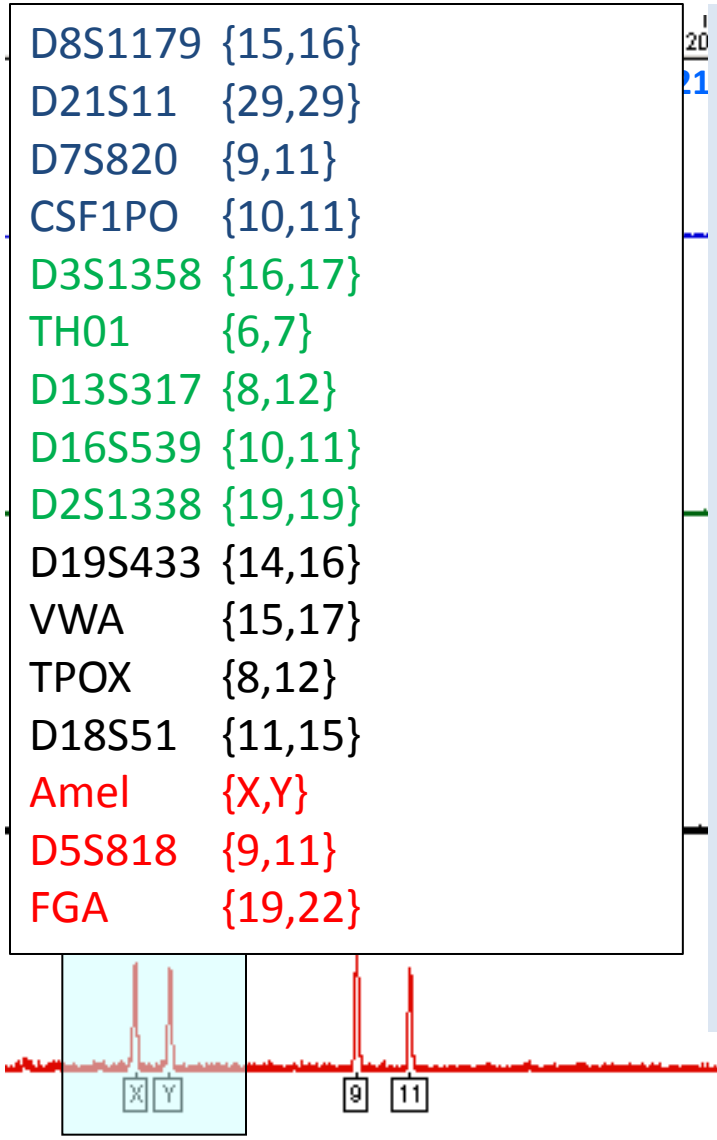
Applied Use of Information

Post extraction, STR typing requires ~1 ng of DNA template (100-200 copies)

Identifiler [Applied Biosystems] 15 STR Loci Kit

Information is tied together with multiplex PCR and data analysis

D8S1179	{15,16}
D21S11	{29,29}
D7S820	{9,11}
CSF1PO	{10,11}
D3S1358	{16,17}
TH01	{6,7}
D13S317	{8,12}
D16S539	{10,11}
D2S1338	{19,19}
D19S433	{14,16}
VWA	{15,17}
TPOX	{8,12}
D18S51	{11,15}
Amel	{X,Y}
D5S818	{9,11}
FGA	{19,22}



Multiplying the frequency of each genotype at each locus gives us the Random Match Probability (RMP) of 1.25×10^{-15} for **unrelated individuals**

*The chance of an **unrelated individual** having this exact same profile is **1 in 800 trillion***

This test contains the 13 FBI core loci

Electrophoretic Analysis of STRs

Fragment Analysis

- Applications
  - Human Identity Testing
  - Missing persons, mass fatalities
  - Kinship/paternity testing (limited)
- Profiles can be developed in a day
- ~1 ng of DNA required (100s of copies)
- Established typing technology, kits, core markers
- Simple data analysis (single source sample)
- Cost ~$30 per sample
- **Limited information about ancestry, phenotype (eye color – hair color), complex kinship scenarios**

Subject to common issues with degraded samples, mixtures, inhibitors

More information (sequence) is required to address these questions

STR sequence characterization

Sanger sequencing

Forensic Science International: Genetics 5 (2011) 329–332

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Short communication

STR sequence analysis for characterizing normal, variant, and null alleles

Margaret C. Kline *, Carolyn R. Hill, Amy E. Decker [1], John M. Butler

National Institute of Standards and Technology, 100 Bureau Drive, M/S 8312, Gaithersburg, MD 20899, USA

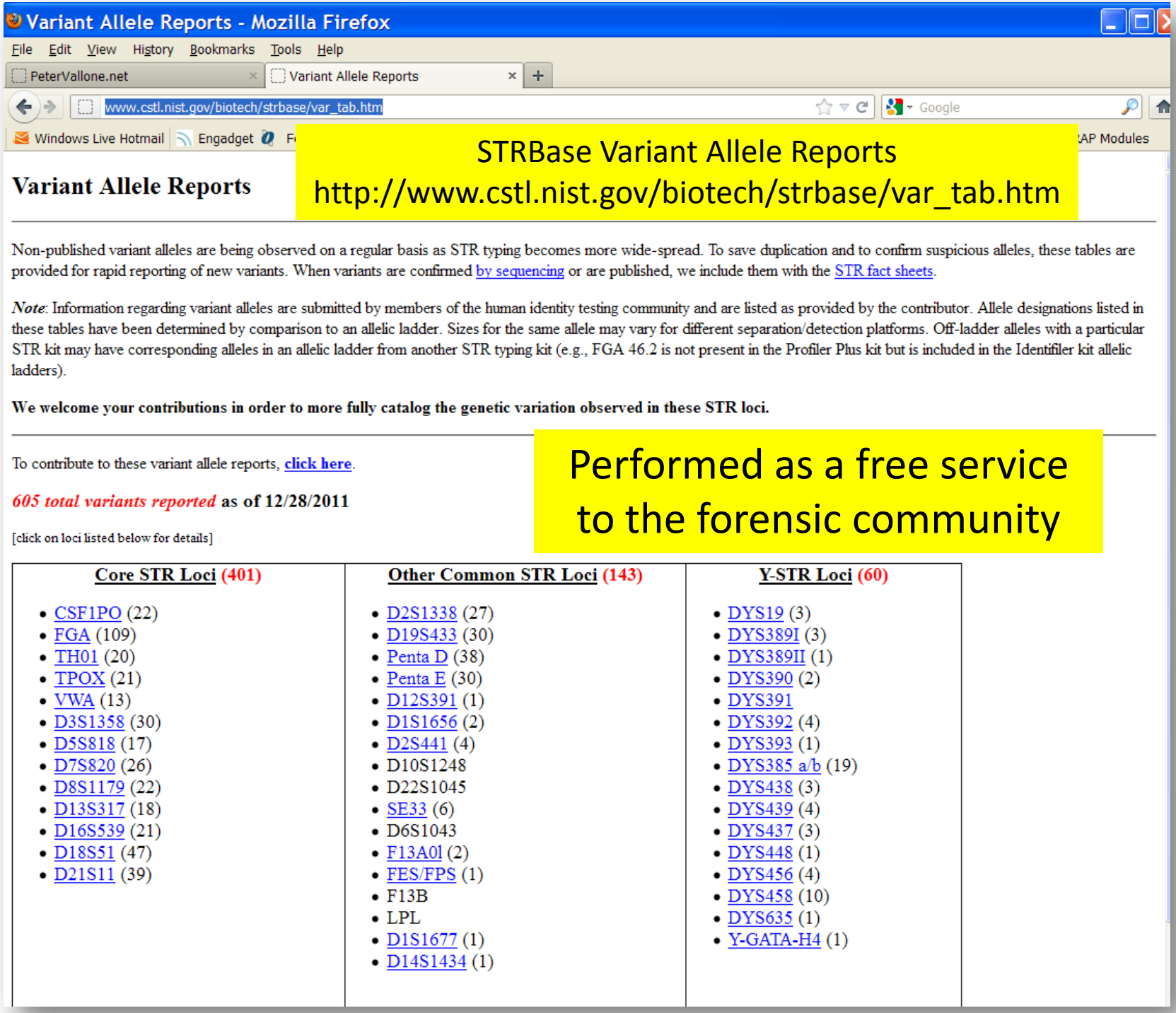
ARTICLE INFO

Article history:
Received 5 April 2010
Received in revised form 23 July 2010
Accepted 8 September 2010

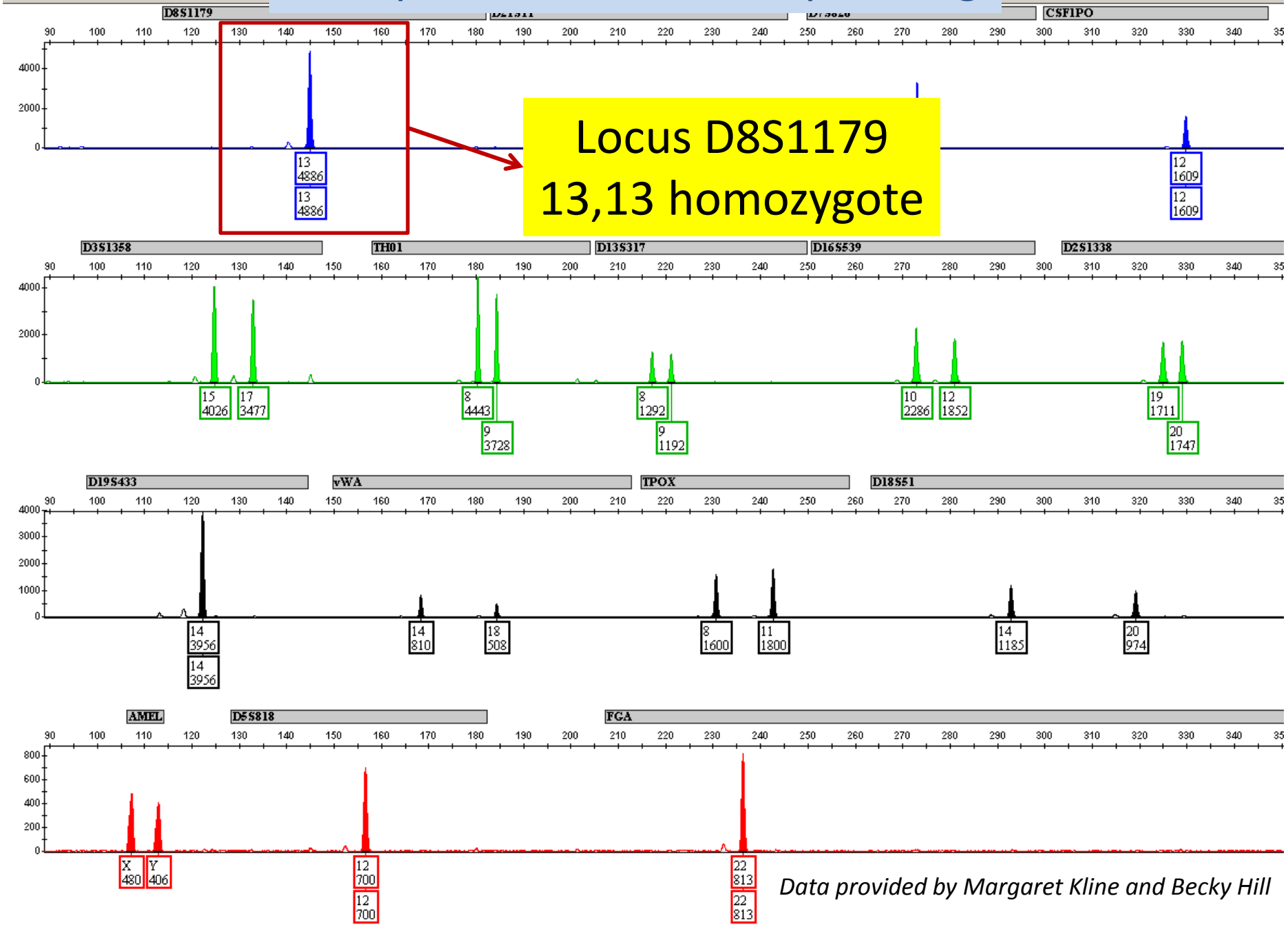
Keywords:
Short tandem repeat
STR typing
DNA sequencing
Allele dropout
Null allele

ABSTRACT

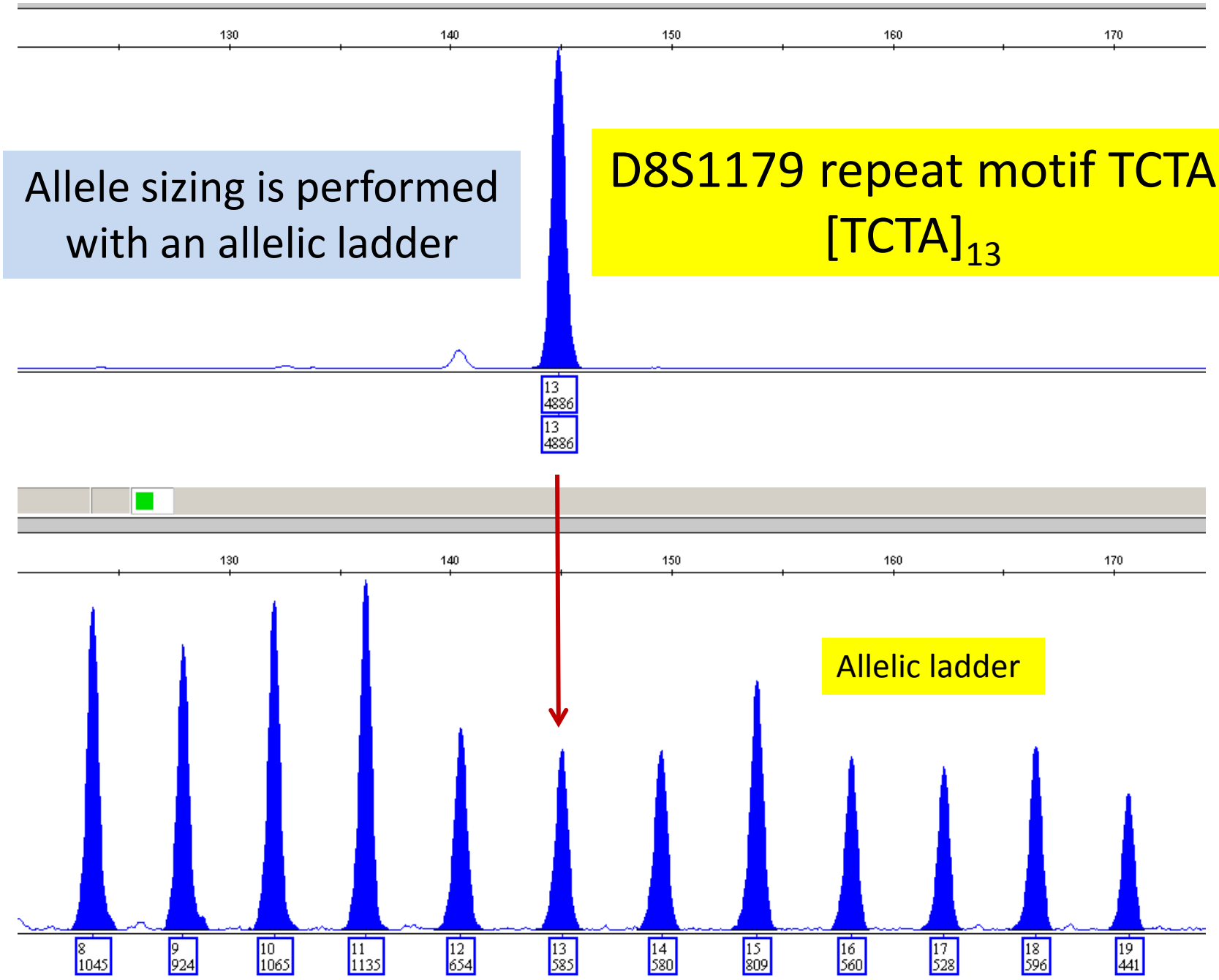
DNA sequence variation is known to exist in and around the repeat region of short tandem repeat (STR) loci used in human identity testing. While the vast majority of STR alleles measured in forensic DNA laboratories worldwide type as "normal" alleles compared with STR kit allelic ladders, a number of variant alleles have been reported. In addition, a sequence difference at a polymerase chain reaction (PCR) primer binding site in the DNA template can cause allele drop-out (i.e., a "null" or "silent" allele) with one set of primers and not with another. Our group at the National Institute of Standards and Technology (NIST) has been sequencing variant and null alleles supplied by forensic labs and cataloging this information on the NIST STRBase website for the past decade. The PCR primer sequences and strategy used for our STR allele sequencing work involving 23 autosomal STRs and 17 Y-chromosome STRs are described along with the results from 111 variant and 17 null alleles.


Variant Allele Reports - Mozilla Firefox
File Edit View History Bookmarks Tools Help
PeterVallone.net
Variant Allele Reports
www.cstl.nist.gov/biotech/strbase/var_tab.htm
Google
Windows Live Hotmail
Engadget
AP Modules
STRBase Variant Allele Reports
http://www.cstl.nist.gov/biotech/strbase/var_tab.htm
Variant Allele Reports
Non-published variant alleles are being observed on a regular basis as STR typing becomes more wide-spread. To save duplication and to confirm suspicious alleles, these tables are provided for rapid reporting of new variants. When variants are confirmed by sequencing or are published, we include them with the STR fact sheets.
Note: Information regarding variant alleles are submitted by members of the human identity testing community and are listed as provided by the contributor. Allele designations listed in these tables have been determined by comparison to an allelic ladder. Sizes for the same allele may vary for different separation/detection platforms. Off-ladder alleles with a particular STR kit may have corresponding alleles in an allelic ladder from another STR typing kit (e.g., FGA 46.2 is not present in the Profiler Plus kit but is included in the Identifiler kit allelic ladders).
We welcome your contributions in order to more fully catalog the genetic variation observed in these STR loci.
To contribute to these variant allele reports, click here.
605 total variants reported as of 12/28/2011
[click on loci listed below for details]
Performed as a free service
to the forensic community
Core STR Loci (401)
CSF1PO (22)
FGA (109)
TH01 (20)
TPOX (21)
VWA (13)
D3S1358 (30)
D5S818 (17)
D7S820 (26)
D8S1179 (22)
D13S317 (18)
D16S539 (21)
D18S51 (47)
D21S11 (39)
Other Common STR Loci (143)
D2S1338 (27)
D19S433 (30)
Penta D (38)
Penta E (30)
D12S391 (1)
D1S1656 (2)
D2S441 (4)
D10S1248
D22S1045
SE33 (6)
D6S1043
F13A01 (2)
FES/FPS (1)
F13B
LPL
D1S1677 (1)
D14S1434 (1)
Y-STR Loci (60)
DYS19 (3)
DYS389I (3)
DYS389II (1)
DYS390 (2)
DYS391
DYS392 (4)
DYS393 (1)
DYS385 a/b (19)
DYS438 (3)
DYS439 (4)
DYS437 (3)
DYS448 (1)
DYS456 (4)
DYS458 (10)
DYS635 (1)
Y-GATA-H4 (1)

Example of STR Allele Sequencing



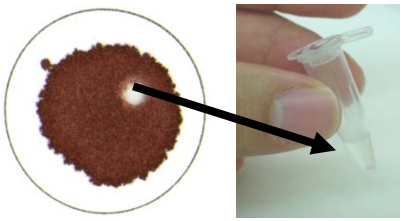
Data provided by Margaret Kline and Becky Hill



Data provided by Margaret Kline and Becky Hill

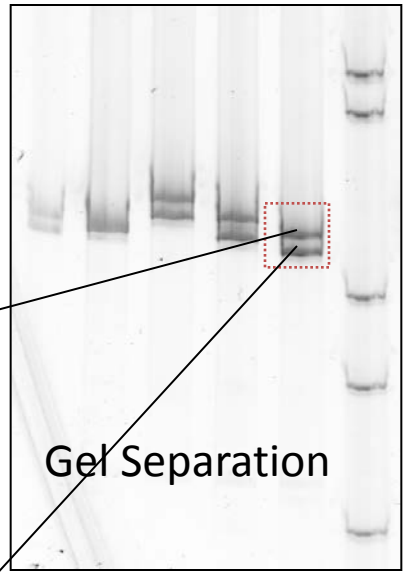
Steps in STR Allele Sequencing

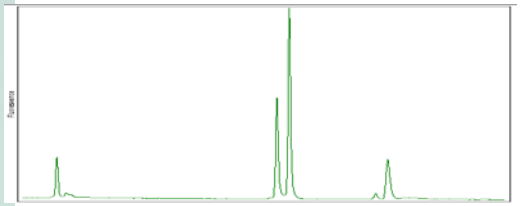
- DNA Extraction
- Amplification with primers external to kit primers
- Gel Cutouts with Heterozygotes
- Re-Amplification
- Amplicon Quantitation
- ExoSAP
- Cycle Sequencing
- Dye Terminator Removal
- F/R Sequence Alignment to Reference Sequence



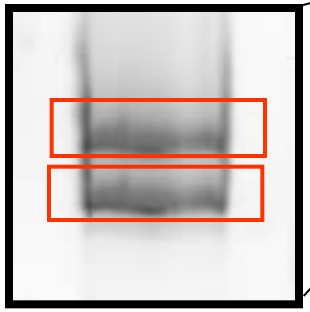
DNA Extraction

PCR Amplification





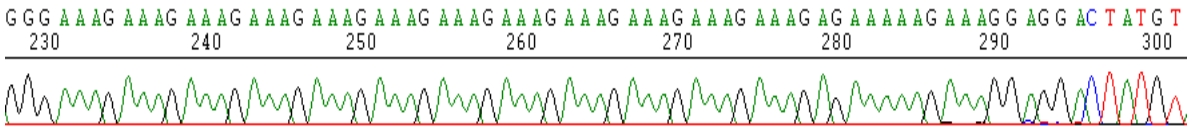
Amplicon Quantitation



Allele Isolation with gel cutouts

Re-Amplification

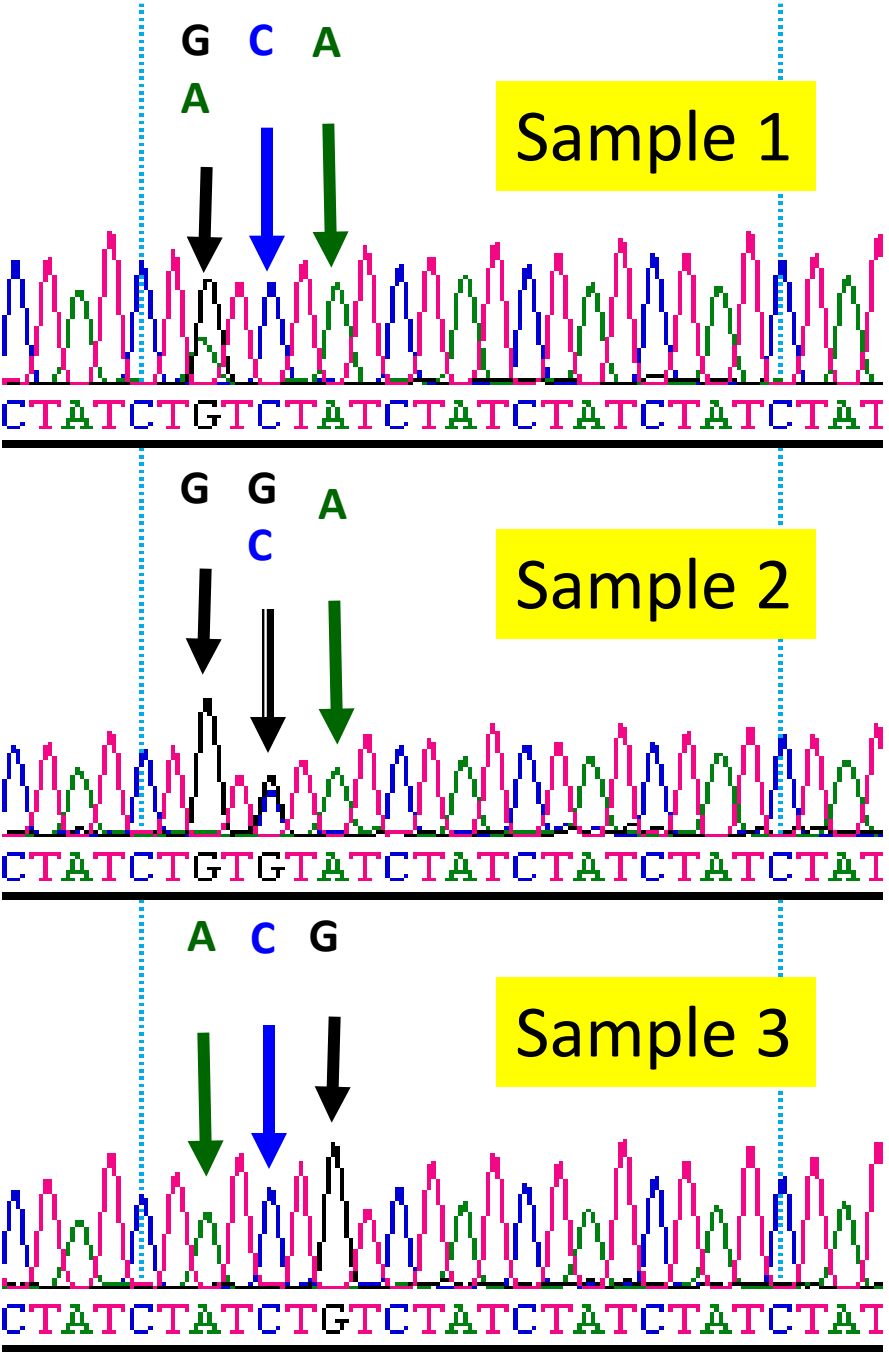
12 GAAA repeats



DNA sequence analysis

SNPs within the D8S1179 repeat

All 3 samples '13,13' $[TCTA]_{13}$



Allele A - $[TCTA]_{13}$

Allele B - TCTA TCT**G** $[TCTA]_{11}$

Allele B - TCTA TCT**G** $[TCTA]_{11}$

Allele C - TCTA TCT**G** T**G**TA $[TCTA]_{10}$

There are **4** different '13' alleles in these 3 samples.

Allele D - $[TCTA]_2$ TCT**G** $[TCTA]_{10}$

Allele D - $[TCTA]_2$ TCT**G** $[TCTA]_{10}$

Data provided by Margaret Kline and Becky Hill

Sequencing of STRs

- Sanger sequencing can provide more information than fragment analysis
  - Increased resolution (one-to-one matching)
  - Can assist with kinship applications

- Detect
  - SNPs within the STR region or PCR products
  - Off ladder alleles, null alleles
  - Microvariants

- **Cannot multiplex, manual workflow, data analysis is more involved than STR typing**

Mass Spectrometry

Determine the base composition
of a PCR product containing STRs

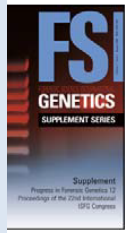
Not sequencing, but SNPs can be detected

$A_{10}G_{20}C_{12}T_4$ -> $A_{10}G_{20}C_{11}T_5$

One less C, one more T

T to C SNP

'Provides Content not Context'



Planz et al. 2009

597 samples were analyzed by ESI-TOF
7/13 core loci contain a significant number of SNPs within STRs

Table 1
Descriptive statistics for seven most polymorphic STR loci containing SNPs

Locus	Population	STR only analysis on IBIS T5000		
		n	Alleles detected	DP
D13S317	Caucasian	182	7	0.9213
	African Am.	214	7	0.8607
	Hispanic	193	7	0.9445
D21S11	Caucasian	182	14	0.9540
	African Am.	214	20	0.9589
	Hispanic	193	14	0.9521
D3S1358	Caucasian	182	8	0.9226
	African Am.	214	8	0.8923
	Hispanic	193	8	0.8939
D5S818	Caucasian	182	9	0.8432
	African Am.	214	9	0.8932
	Hispanic	193	9	0.8679
D7S820	Caucasian	182	8	0.9349
	African Am.	214	8	0.7
	Hispanic	193	9	0.7358
D8S1179	Caucasian	182	10	0.9324
	African Am.	214	10	0.9239
	Hispanic	193	9	0.9303
vWA	Caucasian	182	10	0.9388
	African Am.	214	11	0.9403
	Hispanic	193	7	0.9108

DP, discrimination power; H_e, expected heterozygosity; H_o, observed heter

Planz et al. 2009

Table 1
Descriptive statistics for seven most polymorphic STR loci containing SNPs

Locus	Population	STR only analysis on IBIS T5000			STR-SNP analysis on IBIS T5000		
		n	Alleles detected	DP	*n*	Alleles detected	DP
D13S317	Caucasian	182	7	0.9213	181	12	0.9705
	African Am.	214	7	0.8607	213	12	0.9528
	Hispanic	193	7	0.9445	193	13	0.9751
D21S11	Caucasian	182	14	0.9540	181	23	0.9780
	African Am.	214	20	0.9589	213	33	0.9708
	Hispanic	193	14	0.9521	193	25	0.9752
D3S1358	Caucasian	182	8	0.9226	181	18	0.9671
	African Am.	214	8	0.8923	213	18	0.9775
	Hispanic	193	8	0.8939	193	18	0.9455
D5S818	Caucasian	182	9	0.8432	181	15	0.9260
	African Am.	214	9	0.8932	213	17	0.9102
	Hispanic	193	9	0.8679	193	13	0.9554
D7S820	Caucasian	182	8	0.9349	181	15	0.9600
	African Am.	214	8	0.7	213	12	0.9376
	Hispanic	193	9	0.7358	193	14	0.9482
D8S1179	Caucasian	182	10	0.9324	181	14	0.9627
	African Am.	214	10	0.9239	213	19	0.9489
	Hispanic	193	9	0.9303	193	16	0.9639
vWA	Caucasian	182	10	0.9388	181	22	0.9580
	African Am.	214	11	0.9403	213	26	0.9766
	Hispanic	193	7	0.9108	193	16	0.9305

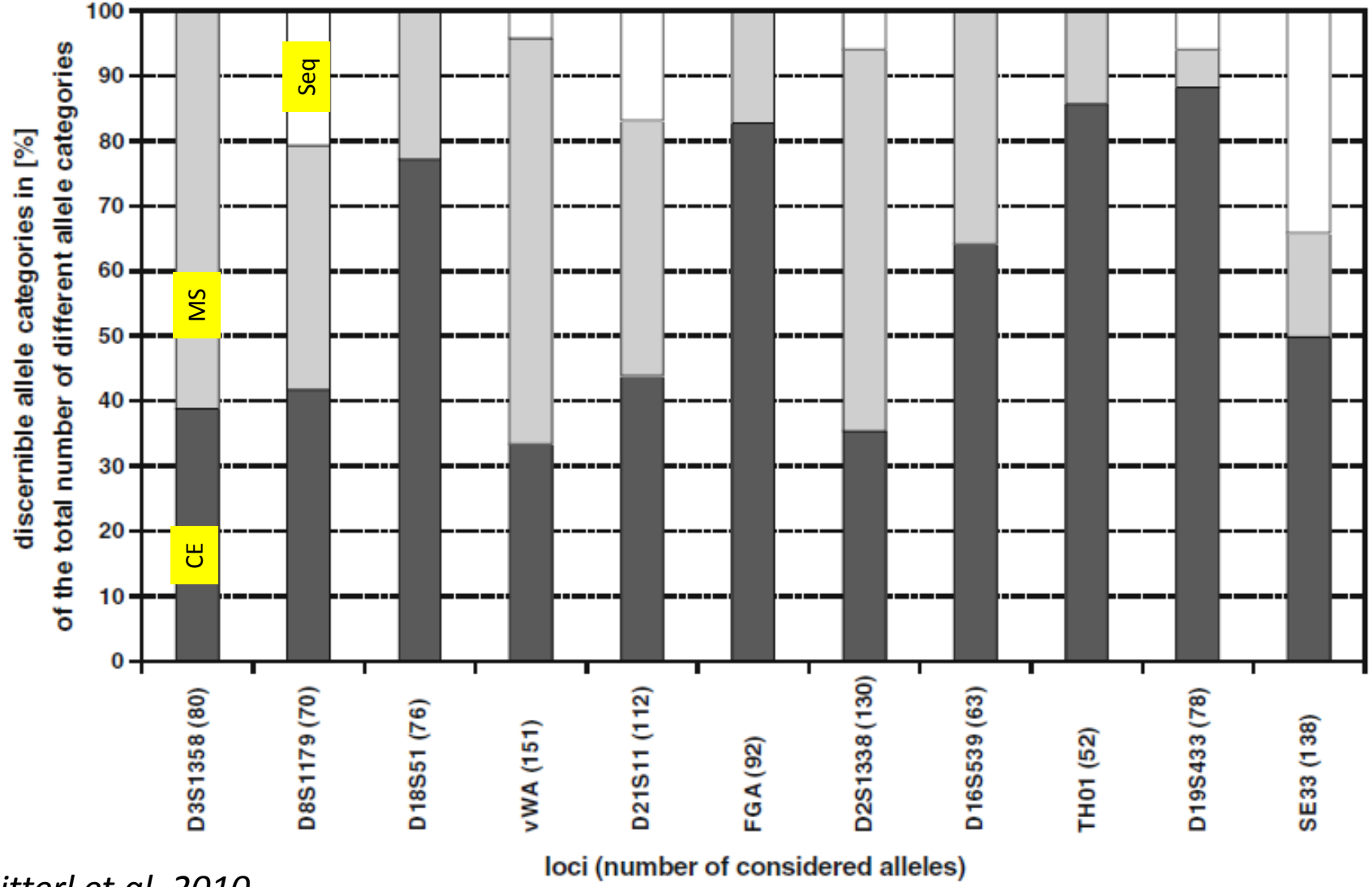
DP, discrimination power; H_e, expected heterozygosity; H_o, observed heter

- DP increased 3.5–5% per locus compared to nominal STR typing
- SNP-containing alleles could be traced though several generations in some pedigrees

Planz et al. 2009

Comparison of the number of differentiable allele categories by three different genotyping technologies

Black = Capillary electrophoresis
Gray = Mass spectrometry
White = Sanger sequencing



Pitterl et al. 2010

Next Generation Sequencing

Ultra High Throughput Sequencing

- Going in depth into STR loci and beyond
  - STRs are useful for legacy (databases)
  - Millions of bases of sequence variants (SNPs)

- Opens up new human identity applications: complex kinship, biogeographical ancestry, externally visible traits, degraded samples?, mixtures?, other applications

Applications are currently being addressed
by the forensic genetics community (K*ayser and deKnijff 2011)*

Next Generation Sequencing

Ultra High Throughput Sequencing

- Challenges
  - Repeating sequences (STRs) and read lengths
  - Sample requirements (10 ng to 5 µg)
  - Cost and time per unit of information
  - Data analysis (storage, assembly, interpretation)
  - Policy, privacy, disease related markers
  - Validation
  - Standards/reference materials
    - Accuracy of sequence information
    - Errors, platform and bioinformatics-based bias

Multiplexing samples and reduce data set while maintaining quality coverage

A single NGS experiment

Single sample – full genome coverage

Multiplexing samples and reduce data set while maintaining quality coverage

A single NGS experiment

1	9	17	25	33	41	49	57	65	73	81	89
2	10	18	26	34	42	50	58	66	74	82	90
3	11	19	27	35	43	51	59	67	75	83	91
4	12	20	28	36	44	52	60	68	76	84	92
5	13	21	29	37	45	53	61	69	77	85	93
6	14	22	30	38	46	54	62	70	78	86	94
7	15	23	31	39	47	55	63	71	79	87	95
8	16	24	32	40	48	56	64	72	80	88	96

96 samples, high depth coverage of the **forensically relevant markers**

100s, 1000s, 500k, 1M per sample

- STRs and SNPs for one-to-one matching
- Ancestry markers (X, Y, mito, autosomal)
- Phenotypic markers (eye color, hair color, etc)
- Kinship (linked and unlinked markers)
- Other
- If possible, avoid disease related markers

Mitigate costs by multiplexing samples and sequencing forensically relevant information

Points of discussion

- **The range of applications that are envisioned for human DNA sequencing within your organization**
- **Technical considerations/limitations for application of Next-Generations sequencing to your problems that may be unique to your organization**
- **Your programmatic plans for developing and implementing this technology including past and current investments as well as timelines for making future investments**
- **Policy implications that you anticipate from the expansion of human DNA analysis for your intended applications**
- **Plans and/or issues associated with human genomic data archiving, analysis, and curation.**
- **Your organization's position on the privacy and security issues related to your envisioned use of human genomic sequence information and your vision and approach for addressing these issues**

Questions?

peter.vallone@nist.gov
DNA Biometrics Team Leader
Biochemical Science Division
National Institute of Standards and Technology
Gaithersburg, MD 20899
301-975-4872

Acknowledgments

Margaret Kline and Becky Hill