Ongoing Projects in the Applied Genetics Group at NIST

Peter Vallone, Michael Coble, Becky Hill, Erica Butts, Kevin Kiesler, Margaret Kline
Applied Genetics Group
U.S. National Institute of Standards and Technology

Victoria Police Forensic Services Department
September 9, 2013
Macleod, Victoria, Australia

NIST Human Identity Project Team

within the Applied Genetics Group

Forensic DNA Team

Funding from the **National Institute of Justice (NIJ)** through NIST Office of Law Enforcement Standards


Pete Vallone
Group Leader


Mike Coble


Becky Hill


Margaret Kline

DNA Biometrics Team

Funding from the **FBI** through NIST Information Access Division


Erica Butts


Kevin Kiesler

Data Analysis Support


Dave Duewer

As of April 1, John Butler has moved into the Office of Special Programs and is working on Forensic Science efforts across NIST

Sources of external funding

NIJ National Institute of Justice

Rapid Instrumentation

- Several companies are currently developing prototype devices
- In collaboration with the FBI and DHS testing of these systems is being performed at NIST
  - Current focus on concordance, reproducibility, and reliability
  - Future focus is to conduct an interlaboratory study for each of the tested platforms

Rapid DNA Prototype Testing

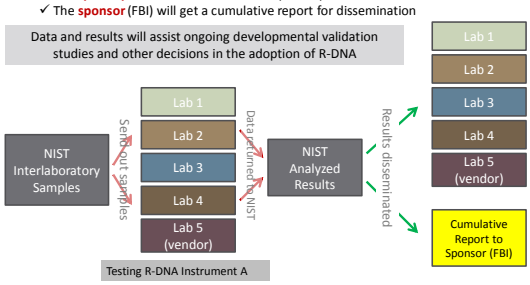
- Received first R-DNA prototypes in Sept 2012
- R-DNA platforms have been evolving quickly over the last 12 months
- Many developmental changes and upgrades within the past year
  - Software, hardware, data processing, etc

Purpose of Interlaboratory Testing

- As of July 2013 R-DNA success levels were high enough to carry out an interlaboratory assessment of the R-DNA prototypes
- Data was collected and analyzed in August 2013
- Results will only be shown from the NIST/DHS instrument
  - Results will be presented at the Biometrics Consortium Conference by Peter Vallone (September 18th, Tampa, FL)

Interlaboratory Testing

- Each participant will be sent a standard sample set for testing
- Provide participants and sponsor with data and feedback
  - ✓ Each **participant** and will receive their specific performance feedback
  - ✓ The **sponsor** (FBI) will get a cumulative report for dissemination



In coordination with the FBI Rapid DNA Program Office (Dr. Tom Callaghan)

Interlaboratory Testing Plan

- Anonymous buccal collection of 50 samples
  - (5 replicates of 10 unique individuals)
  - Swabs were collected 15 months prior to testing
- Schematic of runs (10 chips)

	Chip									
Lane	1	2	3	4	5	6	7	8	9	10
1	A	F	J	E	A	C	F	H	D	I
2	B	G	I	D	B	E	F	J	E	J
3	C	H	H	C	B	D	G	J	A	F
4	D	I	G	B	A	D	G	I	B	G
5	E	J	F	A	C	E	H	I	C	H
Ladder										

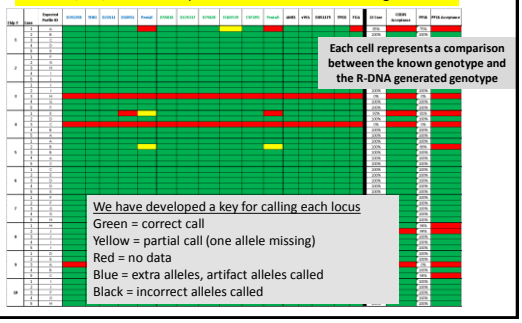
Defining Success

- A complete and correct CODIS core 13 STR profile (as called by the expert system software)
  - If any of the 13 loci allele calls were incorrect or absent this was deemed a lane failure
  - Comparing correct genotypes (lab generated) to the types exported to cmf
- Note: we are not including chips that failed due to hardware issues in success calculations

Run logs and heat maps

Information is recorded, logged, and reported

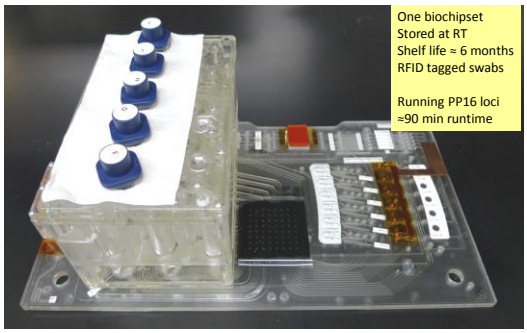
NIST, FBI, DFSC developed the worksheets and scoring rules



ANDE

Electrophoresis takes place on chip
Exports genotypes to .xml files
View egram as bitmap
.fsa files are exported (can be opened in GeneMapper/Genemarker)

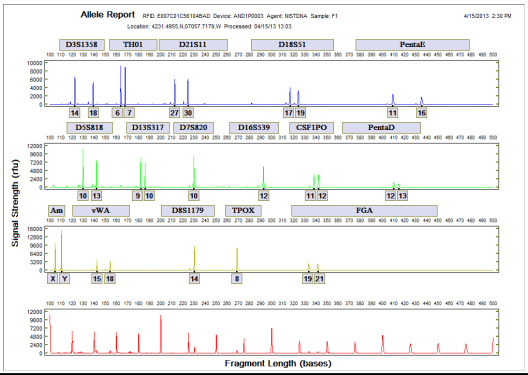
ANDE



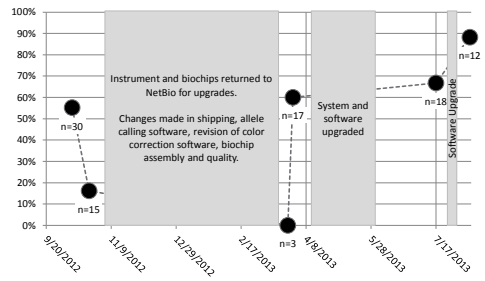
One biochipset
Stored at RT
Shelf life ≈ 6 months
RFID tagged swabs

Running PP16 loci
≈90 min runtime

ANDE

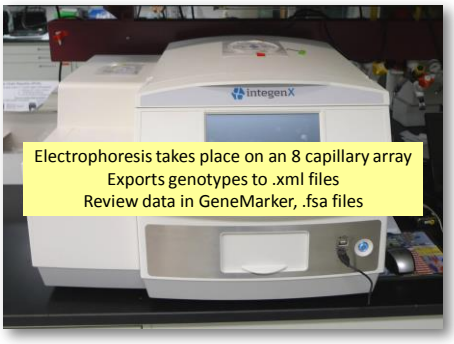


ANDE Timeline of Testing



n=95 chips tested from 9/20/2012 to 8/12/2013

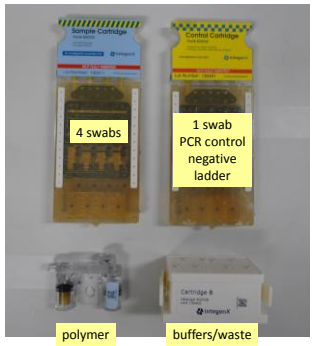
RapidHIT 200



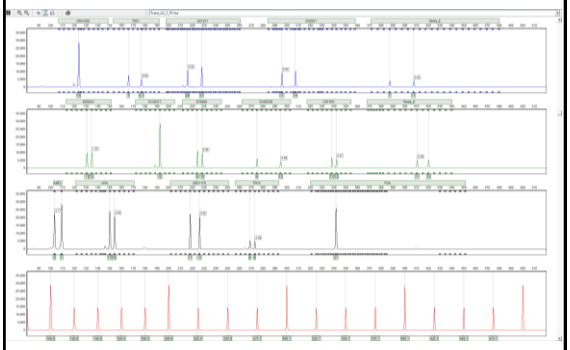
RapidHIT 200

Kit = 4 components
Stored at 4ºC
Shelf life ≈ 3 weeks
Cotton swabs

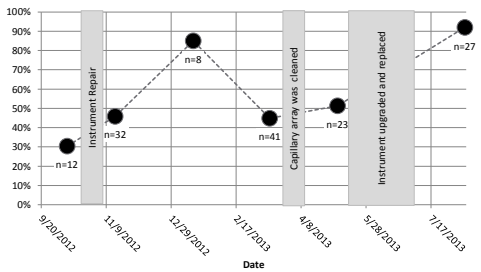
Running PP16 loci
≈90 min runtime



RapidHIT 200



RapidHIT 200 Timeline of Testing



n=143 chips tested from 9/20/2012 to 8/12/2013

When can developmental validation begin?

- When a final version of the software, chemistry, and hardware (the box) has been locked down
- Right now it would be similar to testing an STR kit that is still being optimized
  - Primer concentrations, PCR primer sequences, mastermix, annealing temperatures, etc
- Or CE instrument that is undergoing optimization
  - Capillary array material changes, formulation of spectral matrices, collection software, etc

Topics

- Rapid DNA (Pete Vallone)
- **Y STRs (Becky Hill)**
- Next Gen Sequencing (Kevin Kiesler)

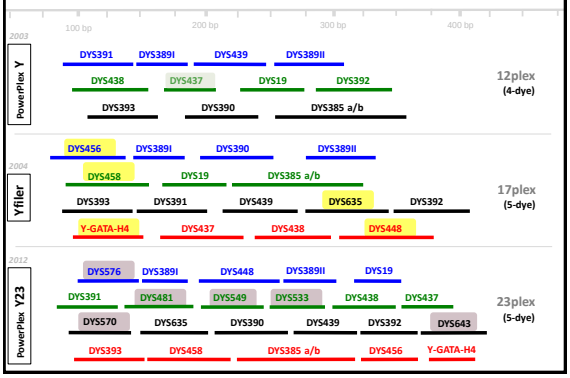
Outline

- Rapidly Mutating (RM) Y-STRs – Overview
- Population Genetic Parameters
  - current Y-STR kits
- Utility for common Y-STR haplotypes
- Utility for close relatives
- Conclusions

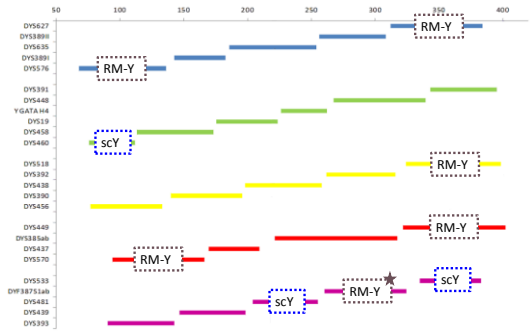
What has happened in the past decade…

- **Selection of core Y-STR loci** (SWGDAM Jan 2003)
- "Full" Y-chromosome sequence became available in June 2003; over 700 Y-STR loci identified (only ~20 in 2000)
- **Commercial Y-STR kits released**
  - ~~Y-PLEX 6,5,12 (2001-03);~~ **PowerPlex Y** (9/03), **Yfiler** (12/04), **PPY23** (6/12)
    Yfiler Plus (coming soon)
- Many population studies performed and online databases generated with thousands of Y-STR haplotypes
- Forensic casework demonstrations showing value of Y-STR testing along with court acceptance
- Some renewed interest in Y-STRs to aid familial searching

STR Marker Layouts for Y-STR Kits



Proposed Yfiler Plus Kit



RM = Rapidly Mutating, sc = single copy

Slide courtesy of Dr. Julio Mulero (Life Technologies)

NIST U.S. Samples (>1450)

- **NIST U.S. population samples**
  - 260 African American, 260 Caucasian, 140 Hispanic, 3 Asian
- **U.S. father/son paired samples**
  - **~100 fathers/100 sons for each group**: 200 African American, 200 Caucasian, 200 Hispanic, 200 Asian
- **NIST SRM 2391b**, PCR-based DNA Profiling Standard (highly characterized)
  - 10 genomic DNA samples, 2 cell line samples
  - Includes 9947A and 9948
- **NIST SRM 2391c**, PCR-based DNA Profiling Standard
  - 4 genomic DNA (one mixture)
  - 2 cell lines (903 and FTA paper)

Hill, C.R., et al. (2011) Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 Systems. *Forensic Sci. Int. Genet.* 5(4): 269-275.

Discrimination Capacity

- is a measure of the number of unique haplotypes in a given population

$$DC = \frac{\#H}{N}$$

# of Haplotypes

Population size

N = 948 males	PowerPlex Y	Yfiler	PowerPlex Y23	Yfiler Plus*
# haplotypes	816	930	945	946
discrimination capacity	0.8608	0.9810	0.9968	0.9979
# times haplotype observed	PPY (12 loci)	Yfiler (17 loci)	PPY23 (23 loci)	Yfiler Plus* (26 loci)
1	751	916	942	944
2	42	11	3	2
3	12	2	.	.
4	4	1	.	.
5	2	.	.	.
6	2	.	.	.
7	.	.	.	.
8	1	.	.	.
9	.	.	.	.
10	.	.	.	.
11	1	.	.	.
12	.	.	.	.
13	.	.	.	.
14	.	.	.	.
15	.	.	.	.
16	.	.	.	.
17	.	.	.	.
18	.	.	.	.
19	.	.	.	.
20	1	.	.	.

Number of unique and shared haplotypes observed with various combinations of Y-STR loci across 948 U.S. population samples

944 haplotypes occur once; and
2 sets of sample pairs cannot be resolved from one another

*Note: Analysis does not include information from DYS460 – only 26 of the 27 markers in Yfiler Plus were examined in this study.

N = 948 males	Yfiler	New Loci*	Yfiler Plus*
# haplotypes	930	945	946
discrimination capacity	0.9810	0.9842	0.9979
# times haplotype observed	Yfiler (17 loci)	New Loci* (9 loci)	Yfiler Plus* (26 loci)
1	916	918	944
2	11	15	2
3	2	.	.
4	1	.	.
5	.	.	.
6	.	.	.
7	.	.	.
8	.	.	.
9	.	.	.
10	.	.	.
11	.	.	.
12	.	.	.
13	.	.	.
14	.	.	.
15	.	.	.
16	.	.	.
17	.	.	.
18	.	.	.
19	.	.	.
20	.	.	.

9 of the 10 new loci alone perform slightly better than Yfiler

*Note: Analysis does not include information from DYS460 in this study

Disadvantages of the Y-Chromosome

- Loci are not independent of one another and therefore rare random match probabilities cannot be generated with the product rule; must use haplotypes (combination of alleles observed at all tested loci)
- **Paternal lineages possess the same Y-STR haplotype** (barring mutation) and thus fathers, sons, brothers, uncles, and paternal cousins cannot be distinguished from one another
- Not as informative as autosomal STR results
  - **More like addition (10 + 10 + 10 = 30) than multiplication (10 x 10 x 10 = 1,000)**

Rapidly Mutating (RM) Y-STRs

Trying to separate
close male relatives

Rapidly Mutating Y-STRs

The American Journal of Human Genetics 87, 341–353, September 10, 2010 ARTICLE

Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications

Kaye N. Ballantyne,[1] Miriam Goedbloed,[1] Rixun Fang,[2] Onno Schaap,[1] Oscar Lao,[1] Andreas Wollstein,[1,3] Ying Choi,[1] Kate van Duijn,[1] Mark Vermeulen,[1] Silke Brauer,[1,4] Ronny Decorte,[5] Micaela Poetsch,[6] Nicole von Wurmb-Schwark,[7] Peter de Knijff,[8] Damian Labuda,[9] Hélène Vézina,[10] Hans Knoblauch,[11] Rüdiger Lessig,[12] Lutz Roewer,[13] Rafal Ploski,[14] Tadeusz Dobosz,[15] Lotte Henke,[16] Jürgen Henke,[16] Manohar R. Furtado,[2] and Manfred Kayser[1,*]

Manfred Kayser

13 markers evaluated

Forensic Science International: Genetics 6 (2012) 208–218
Contents lists available at ScienceDirect
Forensic Science International: Genetics
journal homepage: www.elsevier.com/locate/fsig

A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages

Kaye N. Ballantyne[a,1,2], Victoria Keerl[a,1,3], Andreas Wollstein[a,b], Ying Choi[a], Sofia B. Zuniga[c], Arwin Ralf[a], Mark Vermeulen[a], Peter de Knijff[d], Manfred Kayser[a,*]

[a] Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, 3000 CA Rotterdam, The Netherlands
[b] Cologne Center for Genomics, University of Cologne, D-50674 Cologne, Germany
[c] Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

Using Y-STRs with a higher mutation rate, father-son and brother pairs can sometimes be distinguished

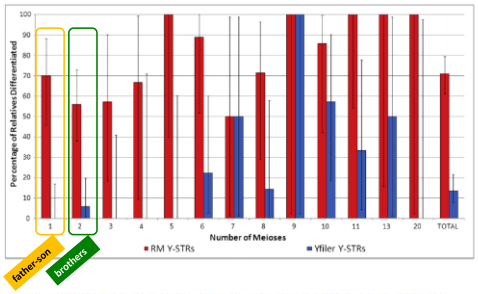


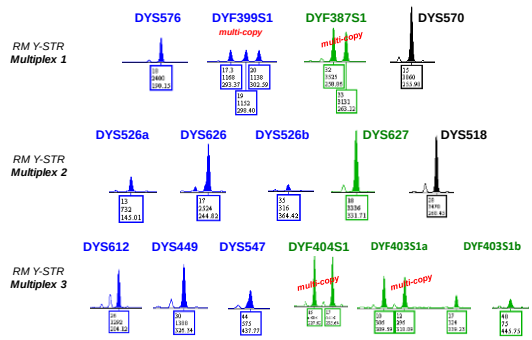
Figure 4. Male Relative Differentiation with Newly Identified 13 RM Y-STRs and Commonly Used 17 Yfiler Y-STRs
Results from differentiating between male relatives from analyzing 103 pairs from 80 male pedigrees, sorted according to the number of generations separating pedigree members, based on 13 RM Y-STRs (in red) and 17 Yfiler Y-STRs (in blue). Error bars represent 95% binomial confidence intervals. Note that these samples are independent from the father-son pairs initially used to establish the Y-STR mutation rates.

The American Journal of Human Genetics 87, 341–353, September 10, 2010

K. Ballantyne et al. 2010; K. Ballantyne et al. 2012

Rapidly Mutating (RM) Y-STRs

NIST supplied data from 1,296 U.S. samples (634 population + 331 father/son pairs) to RM Y-STR Study Group led by Manfred Kayser



Why do these markers mutate "rapidly"?

Locus	(average mutation rate)	"Large" number of repeats
Markers in Yfiler Plus: DYS449	(1.2%)	
DYS518	(1.8%)	
DYS547	(2.4%)	→ DYS547
DYS570	(1.2%)	
DYS576	(1.4%)	
DYS612	(1.4%)	
DYS626	(1.2%)	
DYS627	(1.2%)	
DYF387S1	(1.6%)	Multi-copy Markers
DYF399S1	(7.7%)	→ DYF399S1
DYF403S1 a/b	(3.1/1.2%)	
DYF404S1	(1.3%)	
DYS526 a/b	(1.3%)	

DYS458 (0.64%) is highest in Yfiler loci where average is ~0.2%

Gene Diversity

- is a measure of the uniqueness of a particular marker in a given population

$$GD = (1 - \sum_i x_i^2)$$

Relative frequency of each allele

N = 100

$$GD = (1 - \sum_i x_i^2)$$

0

DC = 1/100 = 0.01

N = 100

$$GD = (1 - \sum_i x_i^2)$$

0.75

DC = 4/100 = 0.04

N = 100

GD = $(1 - \sum_i x_i^2)$

0.99

DC = 100/100 = 1.0

Gene Diversity of the Markers

Marker	GD	DC
DYS576	0.766	0.035
DYF399S1	0.993	0.587
DYF387S1	0.870	0.098
DYS570	0.743	0.035
RM-01 (all)	**0.9998**	**0.9764**

Marker	GD	DC
DYS526a/b	0.923	0.138
DYS626	0.794	0.043
DYS627	0.848	0.043
DYS518	0.791	0.039
RM-02 (all)	**0.9985**	**0.8661**

DYS385a/b GD = 0.929

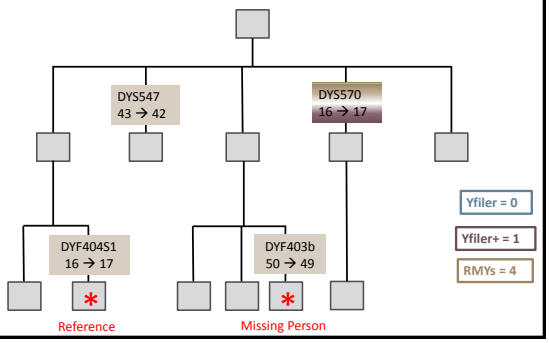
Marker	GD	DC
DYF403S1a/b	0.923	0.791
DYF404S1	0.902	0.110
DYS612	0.832	0.043
DYS449	0.796	0.043
DYS547	0.798	0.039
RM-03 (all)	**1.000**	**0.9984**

All 13 RM Y-STRs resolved 948 males

Paternal Relatives in the Database

	PPY-23	mtDNA	Kinship Index	RM Mutations
Y27	match	n/a	Father-Son	0
Y28	match	n/a	254,325,532	
Y16	match	match	Full Sib	0
Y17	match	match	155,463	
ZT79994	match	match	Full Sib	1
ZT79995	match	match	56,327	
GT37828	match	C1 (Native)	Cousin	4
C87H	match	n/a	0.228	
PT84348	match	L1b (African)	Cousin	3
ZT80369	match	C1 (Native)	0	
ZT79304	match	L2a (African)	Cousin	3
PT84253	match	L1b (African)	0.568	

Y-STR mutations in a Paternal Lineage



Mutation Rate Information

Meioses	Mutations	Group
63	15	AfAm
89	25	Asian
91	11	Caucasian
88	20	Hispanic
331	71	total
		(21.4%)

+1 Repeat (Son)	-1 Repeat (Son)	
8	6	AfAm
11	13	Asian
5	6	Caucasian
8	12	Hispanic
+2 Repeat (Son)	-2 Repeat (Son)	
0	1	AfAm
1	0	Asian
0	0	Caucasian
0	0	Hispanic

Marker	# of Mutations
DYF399S1	15
DYF403S1a/b	11
DYS627	7
DYS612	7
DYS518	6
DYS570	5
DYS626	5
DYS547	4
DYS526a/b	3
DYS576	3
DYS449	3
DYF404S1	1
DYF387S1	1

Interpretational Issues

- We will need to move away from simply "excluding" based upon a set number of discordant markers.
- A Likelihood Ratio can provide weight to the evidence based upon competing propositions.
- This will require information on the ***haplotype frequency*** and ***mutation rate data***.

Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping

Manfred Kayser[a,*], Mark Vermeulen[a,b], Hans Knoblauch[c], Herbert Schuster[d], Michael Krawczak[e], Lutz Roewer[f]

Forensic Science International: Genetics 1 (2007) 125–128.

NIST Y-STR Data

- All PPY23 Y-STR haplotypes have been submitted to the Y-HRD and U.S. Y-STR databases
- Much of this data presented has been recently published in Profiles in DNA and FSI: Genetics

Short communication

Haplotype data for 23 Y-chromosome markers in four U.S. population groups

Michael D. Coble*, Carolyn R. Hill, John M. Butler

Variability of New STR Loci and Kits in US Population Groups

Summary

- Rapidly Mutating Y-STRs are highly diverse markers that can discriminate common haplotypes and close relatives.
- These markers may create interpretational issues for paternity/missing persons cases, but LRs can be useful for evaluating these situations.
- An international consortium is gathering frequency and mutation rate data.
- We plan on testing Yfiler Plus with our population samples as soon as the Material Transfer Agreement (MTA) gets signed by both parties (NIST and Life Tech)

Topics

- Rapid DNA (Pete Vallone)
- Y STRs (Becky Hill)
- **Next Gen Sequencing (Kevin Kiesler)**

Use of NGS for forensic applications

Highly-parallel/high-throughput next-generation sequencing technologies provide the ability to directly sequence forensically relevant targets
Issues: sample input amounts, back compatibility, new workflows, cost, throughput, etc

- Mitochondrial whole genome analysis
  - Potential for improved sensitivity, mixture detection, multiplex sequencing of full mitochondrial genomes
  - Detection of minor SNP variants – heteroplasmy
- Going in depth into STR loci
  - STRs are useful for legacy (databases)
  - SNPs within STRs identify 'sub-alleles'
- Forensically relevant SNPs: newer human identity applications: biogeographical ancestry, externally visible traits, complex kinship, degraded samples, low template

Initial Goals

- To characterize NIST forensic SRMs with NGS
  - Further characterizes the materials with a new technique
  - Supports adoption of NGS in forensic community
  - SRM 2391c: PCR Based DNA Profiling Standard
    - Not all STR loci have full sequence information
  - SRMs 2392 and 2392-I: Mitochondrial DNA Sequencing
    - Confirm Sanger data with a high coverage sequencing technology
  - Understand bias between NGS platforms: chemistry and bioinformatics
- Is there a need for a new material?
  - Forensic validation

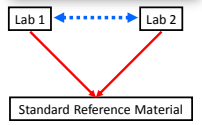
NIST Standard Reference Materials

http://www.nist.gov/srm/

Traceable standards to ensure accurate measurements in our nation's crime laboratories

Human Identity SRMs
SRM 2391c – PCR-Based DNA Profiling
SRM 2392 & 2392-I – mitochondrial DNA
SRM 2395 – Y-STR DNA Profiling
SRM 2372 – Human DNA quantitation

Genomic DNAs characterized for the expanded CODIS core loci and Y-STRs



Calibration with SRMs enables confidence in comparisons of results between laboratories and technologies

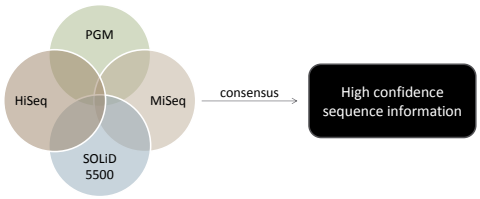
Characterization of the existing SRMs

Current Status

- 2391c **PCR Based DNA profiling standard**
  - 68 STR markers (51 autosomal + 17 Y chromosome)
  - STR repeat lengths (alleles) were certified using multiple (unique) PCR primer sets
  - Sanger sequencing was only performed for loci without multiple PCR primer sets (only 10% of markers)
- 2392 & 2392-I **Mitochondrial DNA sequencing standard**
  - Entire mtGenome (≈16,569 bp) was certified by Sanger sequencing
  - Multiple F/R strand coverage across the mtGenome

Multiple NGS Platforms

- Use of multiple platforms to obtain a consensus sequence for the SRMs
  - Identify and reduce the false positives and negatives
  - Identify and account for bias in a specific chemistry and/or informatics pipeline

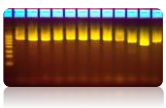


Sequencing Studies

Performed on four NGS platforms for SRM 2392 and 2392-I

- **Ion Torrent PGM**
  - Experiments performed at NIST
  - Edge Biosystems (outsourced)
- **Illumina HiSeq 2000**
  - Beckman-Coulter Genomics (outsourced)
- **Illumina MiSeq**
  - Edge Biosystems (outsourced)
- **SOLiD 5500**
  - Experiments performed at NIST

12 Amplicon PCR
0.8 kb to 1.9 kb



Illumina MiSeq and HiSeq platforms will be online at NIST by the end of this year

Data Processing Pipeline:
Signal Processing, Alignment, and Variant Calling

	Ion Torrent PGM	Illumina MiSeq	Illumina HiSeq	SOLiD 5500
Signal Processing Output: FASTQ	Torrent Server	MiSeq Reporter	HiSeq Control	LifeScope
Read Mapping Output: BAM	Torrent Server	Novoalign	BWA	LifeScope
Variant Calling Output: VCF	Torrent Server	GATK	GATK	GATK

Abbreviations:
FASTQ – Unaligned reads in text format with quality scores
BAM – Binary Alignment Map (Aligned reads)
VCF – Variant Call File
BWA – Burrows Wheeler Aligner
GATK – Genome Analysis Tool Kit

Sequence Coverage Summary

Experiment	Average Read Depth (AQ20*)	Experiment Design
EdgeBio PGM	280 x	Seven mtGenomes + spike-in controls**
NIST PGM Run 1	6,500 x	Three mtGenomes
NIST PGM Run 2	9,000 x	Three mtGenomes
Illumina MiSeq	49,000 x	Seven mtGenomes
Illumina HiSeq	41,000 x	Seven mtGenomes + spike-in controls**
NIST SOLiD	29,000 x	Seven mtGenomes + spike-in controls**

* AQ20 = reads with alignment quality score of 20 or above
= less than 1 error per 100 bases

**Spike-in control was NIST SRM 2374: DNA Sequence Library for External RNA Controls

False Positives and False Negatives
Using platform specific informatics pipeline

		PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
9947A	FP	1	5	3	21	9	11
	FN	3	4	3	3	3	3
CHR	FP	2	6	10	21	9	10
	FN	3	5	4	3	3	4
HL-60	FP	1	8	8	20	9	8
	FN	1	2	1	1	1	1
Avg Coverage		280	6,500	9,000	49,000	41,000	29,000

Calls made to the rCRS
On average 0.04 % error rate

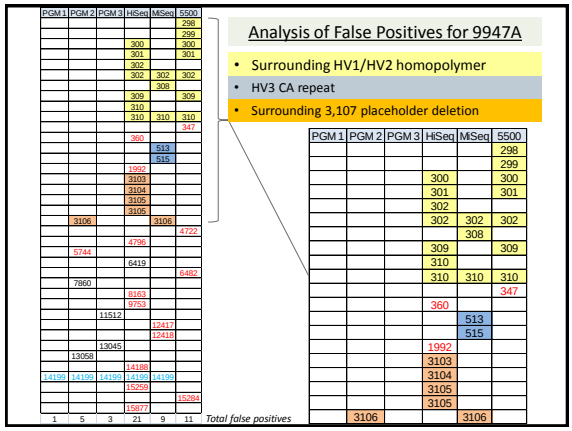
False Positives and False Negatives

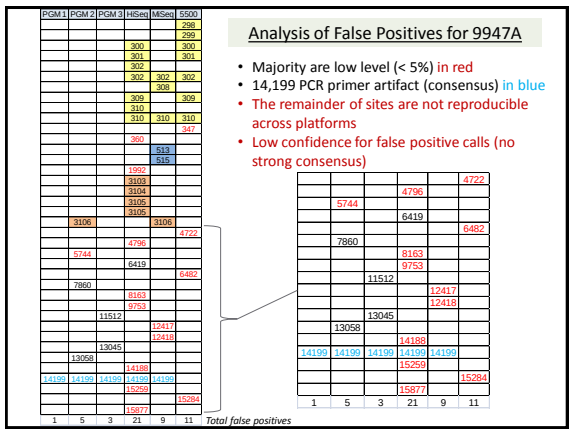
Using platform specific informatics pipeline

		PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
9947A	FP	1	5	3	21	9	11
	FN	3	(4)	3	3	3	3
CHR	FP	2	6	10	21	9	10
	FN	3	5	4	3	3	4
HL-60	FP	1	8	8	20	9	8
	FN	1	(2)	1	1	1	1
Avg Coverage		280	6,500	9,000	49,000	41,000	29,000

False negatives were concentrated in C stretch regions of the genome
The FN sites 13,759 and 5,228 were due to low coverage

9947A (FN) = 309.1, 309.2, 315.1, 13,759
CHR (FN) = 309.1, 315.1, 16193.1, 16183, 16189
HL-60 (FN) = 315.1, 5,228





CHR

PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
					298
			299		
					299
					300
			301		301
			302	302	302
			309		309
310			310	310	310
			310	310	
			360		
				515	
			639		
			1992		
			3103		
			3104		
		3105			
			3105		
			3105		
	3106	3106		3106	
					4547
					4722
			4796		
	5744	5744			
	6220	6220			
			6419		
			9163		
	8230	8230			
		9546	9753		
		11512			
					11826
		12417		12417	
				12418	
	12704				
		13045			
			14188		
14199	14199	14199	14199	14199	
			15259		
			15877		
			16182		
2	6	10	21	9	10

HL-60

PGM 1	PGM 2	PGM 3	HiSeq	MiSeq	5500
				152	
			302		
			309		
					309
			310	310	310
			310	310	
			360		
				515	
			1992		
	2445	2445			
			3103		
			3104		
			3105		
			3105		
	3106			3106	
					3675
					4547
					4722
			4796		
	5149	5149			
	5297				
			6419		
					7508
			8163		
	8230	8230			
		8695			
		9541			
			9753		
		11512			
					11787
					11826
		12417			
				12417	
				12418	
	13058				
			14188		
14199	14199	14199	14199	14199	
			15259		
			15877		
	16361				
			16564		
			16965		
				16568	
1	8	8	20	9	8

Analysis of False Positives for CHR and HL-60

Similar issues with homopolymers, low abundance variants, PCR primer artifacts

Again, the remainder of the false positive sites are not reproducible across platforms

Variant Calls – Concordance by Consensus

SRM 2392 Component B (9947A)

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G	G
4135	T	C	C	C	C	C	C	C
4769	A	G	G	G	G	G	G	G
7645	T	C	C	C	C	C	C	C
7861	T	C	T/C	T/C	T/C	T/C	T/C	T/C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

Variant Calls – Concordance by Consensus

SRM 2392 Component B (9947A)

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393								G/A
1438								G
4135								C
4769								G
7645								C
7861								T/C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

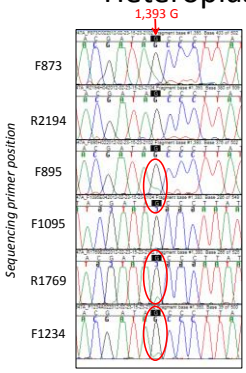
All polymorphisms from Sanger sequencing confirmed
Exception: C-stretch insertions/deletions
Future variant caller algorithms may improve in/del performance

Heteroplasmy at Positions 1,393 and 7,861

SRM 2392 Component B (9947A)

Nucleotide Position	rCRS Reference Sequence	SRM 2392 Component B Sanger Call	EdgeBio PGM	NIST PGM run 1	NIST PGM run 2	EdgeBio Illumina MiSeq	Beckman Genomics Illumina HiSeq	NIST SOLiD
93	A	G	G	G	G	G	G	G
195	T	C	C	C	C	C	C	C
214	A	G	G	G	G	G	G	G
263	A	G	G	G	G	G	G	G
309.1	:	C						
309.2	:	C						
315.1	:	C						
750	A	G	G	G	G	G	G	G
1393	G	G	G/A	G/A	G/A	G/A	G/A	G/A
1438	A	G	G	G	G	G	G	G
4135	T	C	C	C	C	C	C	C
4769	A	G	G	G	G	G	G	G
7645	T	C	C	C	C	C	C	C
7861	T	C	T/C	T/C	T/C	T/C	T/C	T/C
8448	T	C	C	C	C	C	C	C
8860	A	G	G	G	G	G	G	G
9315	T	C	C	C	C	C	C	C
13572	T	C	C	C	C	C	C	C
13759	G	A	A		A	A	A	A
15326	A	G	G	G	G	G	G	G
16311	T	C	C	C	C	C	C	C
16519	T	C	C	C	C	C	C	C

Heteroplasmy at 1,393?



- 6x coverage by Sanger
- 3/6 of reads indicate low-level heteroplasmy
  - Red circles
- Not reproducible in all reads
  - Not always detected by Sanger sequencing

Heteroplasmy detected by NGS at Site 1,393 and 7,861

- Agreement across platforms (high confidence)
  1,393 ≈ 18.0% (± 2.2%) minor component "A"
  7,861 ≈ 14.2% (± 2.9%) minor component "T"

Experiment	1,393 "G" (rCRS)	1,393 "A" (Var.)	Coverage	7,861 "T" (rCRS)	7,861 "C" (Var.)	Coverage
PGM at EdgeBio	77.3 %	22.7 %	97 x	14.1 %	85.9 %	71 x
PGM NIST Run 1	83.9 %	16.1 %	1385 x	20.0 %	80.0 %	191 x
PGM NIST Run 2	83.3 %	16.7 %	1571 x	15.0 %	85.0 %	571 x
NIST SOLiD	82.9 %	17.1 %	22,719 x	12.5 %	87.5 %	17,499 x
Illumina MiSeq	82.1 %	17.8 %	6,517 x	11.6 %	88.4 %	3,715 x
Illumina HiSeq	82.3 %	17.7 %	48,071 x	11.8 %	88.2 %	42,101 x
Average	**82.0 %**	**18.0 %**		**14.2 %**	**85.8 %**	

Summary of mtDNA Work

- The consensus data from the four NGS platforms for the mitochondrial SRMs agree with Sanger sequencing data
  - G/A heteroplasmy at 1,393 confirmed
  - T/C heteroplasmy at 7,861 confirmed
  - C insertions and deletions are issues (assemblers/variant callers)
  - The majority of false positives are of low abundance and not reproducible across platforms
- Continuing work
  - Experiments for setting a variant calling threshold
  - Evaluate a three amplicon approach for mitochondrial DNA enrichment
  - Sequence the mitoSRMs on the PacificBiosciences platform (Collaboration with Children's National Medical Center)
  - Benefit from a standardized (forensic) informatics pipeline (CLC bio software, NextGENe)
  - Evaluate improved variant callers from Life Technologies and Illumina

NIST SRM 2391c

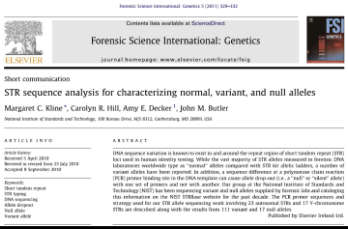
PCR Based Profiling Standard

- Certified Reference Material for STR typing
  - Five components
    - A – Single-source female
    - B – Single-source male
    - C – Single-source male
    - D – Mixed-source (Components A and C)
    - E – Single-source female cells on 903 paper
    - F – Single-source male cells on FTA paper

- Components A, B, and C have been sequenced at NIST on the PGM

Sample Preparation

- PCR primers and conditions were from Kline *et al.* 2011
- All loci amplified in single-plex, purified, then pooled
- Libraries were barcoded to run all samples in one run



Forensic Science International: Genetics

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

Short communication

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

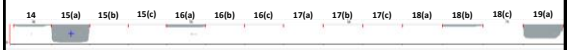
Margaret C. Kline*, Carolyn R. Hill, Amy E. Decker [1], John M. Butler

Analysis Methods

- We are using NextGENe for mapping reads to alleles
  - from Softgenetics
  - Forensic setting filters out reads < 80 % of reference
  - Reads must span all the way across repeat
- NextGENe needs a reference file for mapping reads
  - Create virtual allelic ladders for each locus using known sequence variants
  - Contains repeat structures for STR alleles
    - …CAGGTG GATA GATA GATA GATA GATA TCATTG…
    - …CAGGTG GATA GATA GATA GATA GATA TCATTG…
    - …CAGGTG GATA GATA GATA GATA GATA GATA TCATTG…
  - CODIS core loci plus Amelogenin (also D2, D19, Penta D & E)
    - 702 alleles (in Butler - Advanced Topics in Forensic DNA Typing: Methodology)
  - How much flanking sequence is needed?
    - This has an effect on the analysis
  - What about unknown alleles?
    - This is a limitation

Results from NextGENe

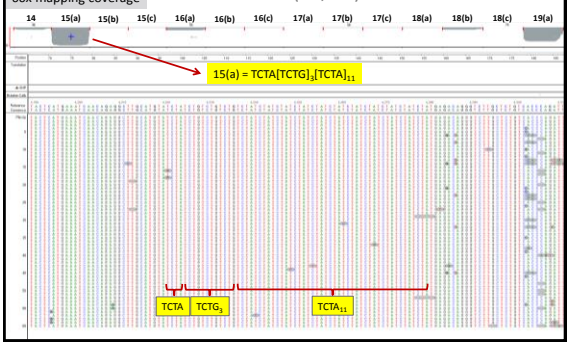
D3S1358 (15, 19) for SRM Component B



Results from NextGENe

D3S1358 (15, 19)

60x mapping coverage



Results for SRM 2391c Component A

- Alleles sequenced for all CODIS loci
  - All alleles confirmed and agree with Sanger sequence
  - D2 and D19 dropped out at the PCR stage

2391c Component A				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	18,23	TBD	To Be Determined	To Be Determined
D3S1358	15,16	15, 16	TCTA[TCTG]$_3$[TCTA]$_{12}$	TCTA[TCTG]$_3$[TCTA]$_{13}$
D5S818	11,12	11,12	[AGAT]$_{11}$	[AGAT]$_{12}$
D7S820	11,11	11,11	[GATA]$_{11}$	[GATA]$_{11}$
D8S1179	13,14	13,14	[TCTA]$_{13}$	[TCTA]$_3$TCTG[TCTA]$_{11}$
D13S317	8,8	8,8	[TATC]$_8$	[TATC]$_8$
D16S539	10,11	10,11	[GATA]$_{10}$	[GATA]$_{11}$
D18S51	12,15	12,15	[AGAA]$_{12}$	[AGAA]$_{15}$
D19S1443	13,14	TBD	To Be Determined	To Be Determined
D21S11	28,32.2	28,32.2	[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCATA)[TCTA]$_{10}$	[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCA TA) [TCTA]$_{11}$ TA TCTA
CSF1PO	10,10	10,10	[AGAT]$_{10}$	[AGAT]$_{10}$
FGA	21,23	21,23	[TTTC]$_3$TTTT TTCT[CTTT]$_{11}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{10}$CTCC[TTCC]$_2$
Penta D	9,13	9,13	[AAAGA]$_9$	[AAAGA]$_{13}$
Penta E	5,10	5,10	[AAAGA]$_5$	[AAAGA]$_{10}$
TH01	8,9.3	8,9.3	[AATG]$_8$	[AATG]$_4$ATG[AATG]$_5$
TPOX	8,8	8,8	[AATG]$_8$	[AATG]$_8$
VWA	18,19	18,19	TCTA[TCTG]$_4$[TCTA]$_{13}$	TCTA[TCTG]$_4$[TCTA]$_{14}$
AMEL	X, X	X, X	No Polymorphisms Observed	No Polymorphisms Observed

Results for SRM 2391c Component B

- Alleles sequenced for all CODIS loci
  - D18 allele 16 needs to be confirmed

★ Novel variant

2391c Component B				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	17,17	TBD	To Be Determined	To Be Determined
D3S1358	15,19	15, 19	TCTA[TCTG]$_3$[TCTA]$_{11}$	TCTA[TCTG]$_3$[TCTA]$_{15}$
D5S818	12,13	12,13	[AGAT]$_{12}$	[AGAT]$_{13}$
D7S820	10,10	10,10	[GATA]$_{10}$	[GATA]$_{10}$
D8S1179	10,13	10,13	[TCTA]$_{10}$	[TCTA]$_{13}$
D13S317	9,12	9,12	[TATC]$_9$	[TATC]$_{12}$
D16S539	10,13	10,13	[GATA]$_{10}$	[GATA]$_{13}$
D18S51	13,16	13,TBD	[AGAA]$_{13}$	To Be Determined
D19S1443	16,16.2	TBD	To Be Determined	To Be Determined
D21S11	32,32.2	★	**[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCATA)[TCTA]$_{12}$**	[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCATA)[TCTA]$_{13}$ TA TCTA
CSF1PO	10,11	10,11	[AGAT]$_{10}$	[AGAT]$_{11}$
FGA	20,23	20,23	[TTTC]$_3$TTTT TTCT[CTTT]$_{11}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{13}$CTCC[TTCC]$_2$
Penta D	8,12	8,12	[AAAGA]$_8$	[AAAGA]$_{12}$
Penta E	7,15	7,15	[AAAGA]$_7$	[AAAGA]$_{15}$
TH01	6,9.3	6,9.3	[AATG]$_6$	[AATG]$_4$ATG[AATG]$_5$
TPOX	8,11	8,11	[AATG]$_8$	[AATG]$_{11}$
VWA	17,18	17,18	TCTA[TCTG]$_4$[TCTA]$_{12}$	TCTA[TCTG]$_4$[TCTA]$_{13}$
AMEL	X, Y	X, Y	No Polymorphisms Observed	No Polymorphisms Observed

Results for SRM 2391c Component C

- Alleles sequenced for all CODIS loci
  - D18 allele 16 needs to be confirmed
  - Novel alleles found at D8, D13

★ Novel variant

2391c Component C				
Locus	Certified Value	NGS Value	Repeat Structure - Allele 1	Repeat Structure - Allele 2
D2S1338	19,19	TBD	To Be Determined	To Be Determined
D3S1358	16, 18	16, 18	TCTA[TCTG]$_3$[TCTA]$_{12}$	TCTA[TCTG]$_3$[TCTA]$_{13}$
D5S818	10,11	10,11	[AGAT]$_{10}$	[AGAT]$_{11}$
D7S820	10,12	10,12	[GATA]$_{10}$	[GATA]$_{12}$
D8S1179	10,17	10,17	[TCTA]$_{10}$	★ **[TCTA]$_4$[TCTG]$_3$[TCTA]$_{14}$**
D13S317	11,11	12,12	★ **[TATC]$_{12}$ Del ATCA 6 bp ds**	★ **[TATC]$_{12}$ Del ATCA 6 bp ds**
D16S539	10,10	10,10	[GATA]$_{10}$	[GATA]$_{10}$
D18S51	16,19	TBD, 19	To Be Determined	[AGAA]$_{19}$
D19S1443	13.2, 15.2	TBD	To Be Determined	To Be Determined
D21S11	29,30	29,30	[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCATA)[TCTA]$_{11}$	[TCTA]$_4$[TCTG]$_6$ ([TCTA]$_3$TA[TCTA]$_2$TCA[TCTA]$_3$TCCATA)[TCTA]$_{11}$
CSF1PO	10,12	10,12	[AGAT]$_{10}$	[AGAT]$_{12}$
FGA	24,26	24,26	[TTTC]$_3$TTTT TTCT[CTTT]$_{10}$CTCC[TTCC]$_2$	[TTTC]$_3$TTTT TTCT[CTTT]$_{12}$CTCC[TTCC]$_2$
Penta D	10,11	10,11	[AAAGA]$_{10}$	[AAAGA]$_{11}$
Penta E	12,13	12,13	[AAAGA]$_{12}$	[AAAGA]$_{13}$
TH01	6,8	6,8	[AATG]$_6$	[AATG]$_8$
TPOX	11,11	11,11	[AATG]$_{11}$	[AATG]$_{11}$
VWA	16,18	16,18	TCTA[TCTG]$_4$[TCTA]$_{13}$	TCTA[TCTG]$_4$[TCTA]$_{15}$
AMEL	X, Y	X, Y	No Polymorphisms Observed	No Polymorphisms Observed

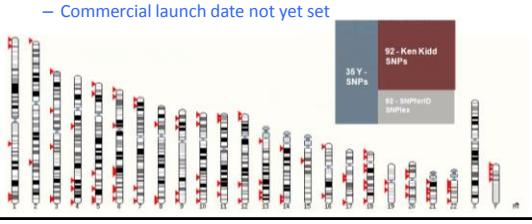
Summary of STR Work

- SRM 2391c Components A, B, and C sequenced
  - Used both 200 bp and 400 bp chemistry
  - Ion Torrent PGM – 314 (v2) and 316 (v2) chips
- Analysis parameters are being optimized
  - Some loci were more difficult to analyze (e.g. CSP1PO)
    - Changed flanking region sequence to 35 bp either side
    - Improved specificity to alleles, reduced # of reads mapping
  - Dropouts at D18 (16 allele) for SRM Components B and C
- Illumina MiSeq being installed at NIST next week

Life Technologies

Coming Soon for PGM

- HID SNP Panel v2.2
  - Autosomal loci chosen for high heterozygosity and low Fst
  - Genotype match probability 10^{-31} to 10^{-35}
  - 179 loci amplified in a single multiplex PCR
  - Short amplicons ≈ 150 bp
  - Commercial launch date not yet set



Life Technologies

Future Plans

- Ancestry informative and phenotypic SNP panel
- For generating investigative leads, subject exclusion
- 245 SNPs
  - 202 Ancestral SNPs
  - 45 Hair and eye color SNPs

Thanks for your attention!

Questions?

Peter.Vallone@nist.gov
Becky.Hill@nist.gov
Kevin.Kiesler@nist.gov

Outside funding agencies:
FBI - Evaluation of Forensic DNA Typing as a Biometric Tool
NIJ – Interagency Agreement with the Office of Law Enforcement Standards