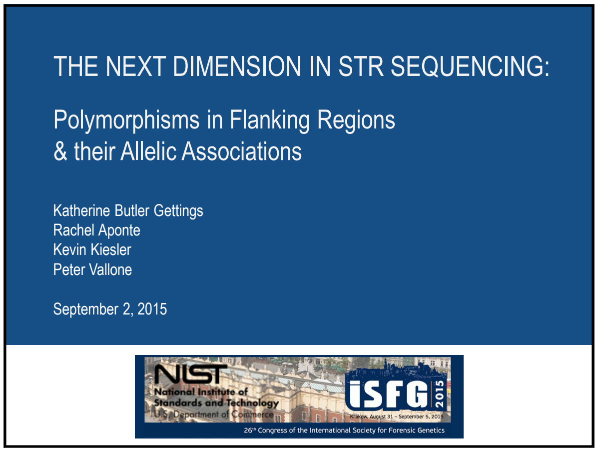


THE NEXT DIMENSION IN STR SEQUENCING:

Polymorphisms in Flanking Regions & their Allelic Associations

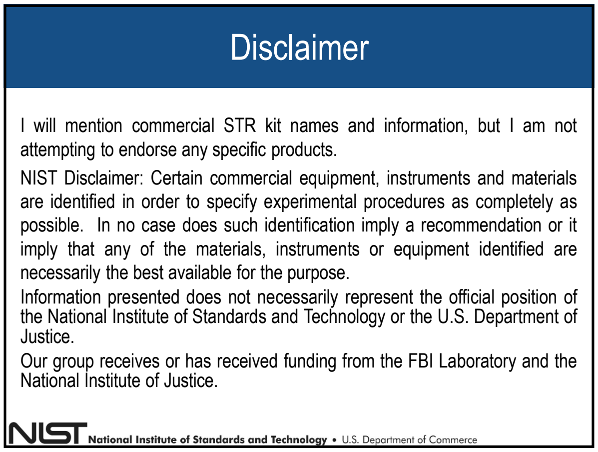
Katherine Butler Gettings
Rachel Aponte
Kevin Kiesler
Peter Vallone

September 2, 2015

NIST National Institute of Standards and Technology U.S. Department of Commerce

iSFG 2015 Krakow, August 31 – September 5, 2015

26th Congress of the International Society for Forensic Genetics



Disclaimer

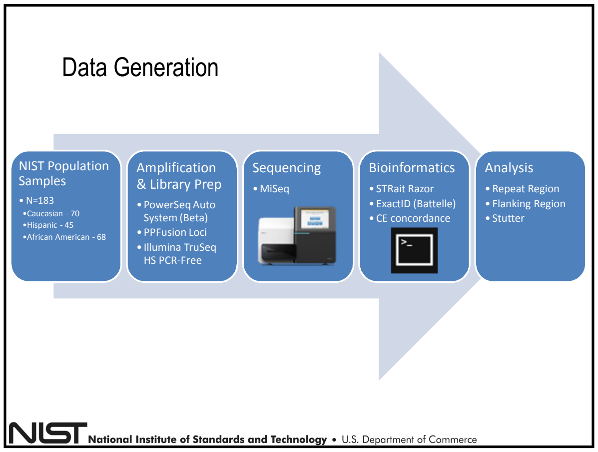
I will mention commercial STR kit names and information, but I am not attempting to endorse any specific products.

NIST Disclaimer: 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.

Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice.

Our group receives or has received funding from the FBI Laboratory and the National Institute of Justice.

NIST National Institute of Standards and Technology • U.S. Department of Commerce



Data Generation

NIST Population Samples
- N=183
  - Caucasian - 70
  - Hispanic - 45
  - African American - 68

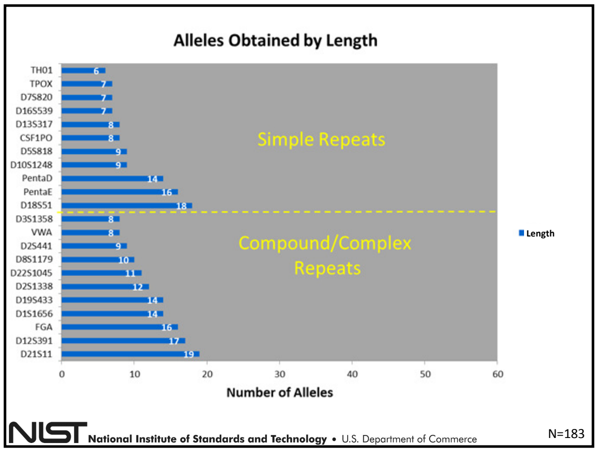
Amplification & Library Prep
- PowerSeq Auto System (Beta)
- PPFusion Loci
- Illumina TruSeq HS PCR-Free

Sequencing
- MiSeq

Bioinformatics
- STRait Razor
- ExactID (Battelle)
- CE concordance

Analysis
- Repeat Region
- Flanking Region
- Stutter

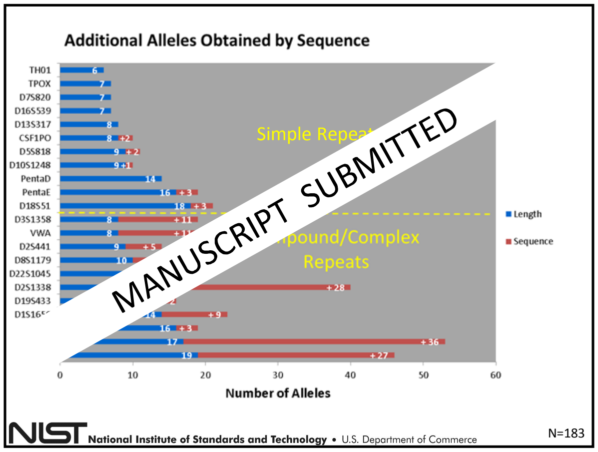
NIST National Institute of Standards and Technology • U.S. Department of Commerce



Alleles Obtained by Length

National Institute of Standards and Technology • U.S. Department of Commerce

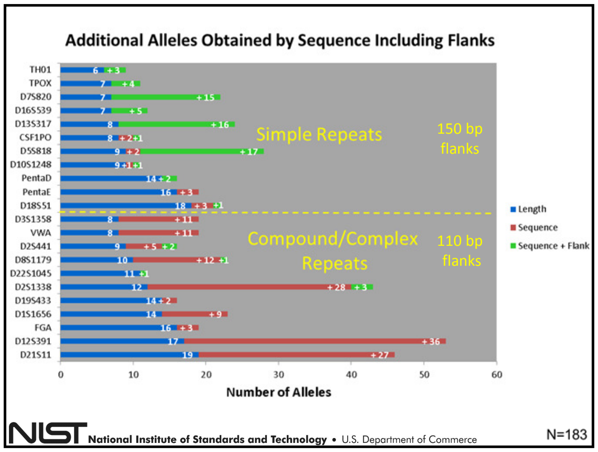
N=183



Additional Alleles Obtained by Sequence

National Institute of Standards and Technology • U.S. Department of Commerce

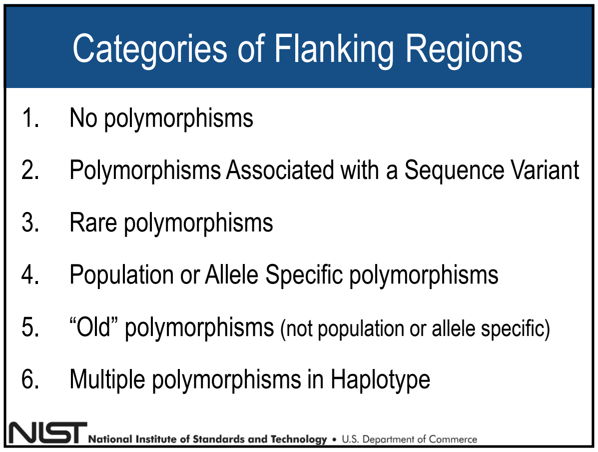
N=183



Additional Alleles Obtained by Sequence Including Flanks

National Institute of Standards and Technology • U.S. Department of Commerce

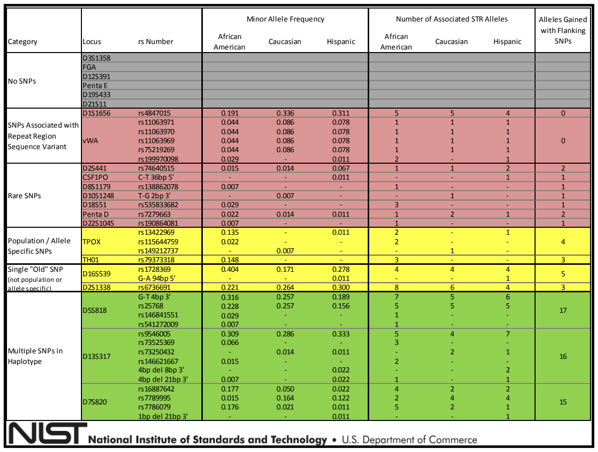
N=183



Categories of Flanking Regions

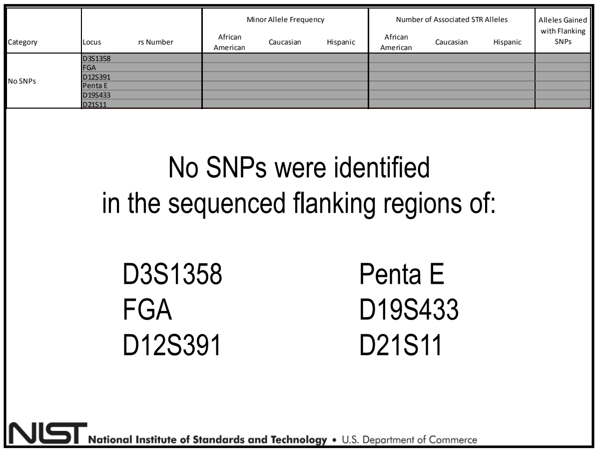
1. No polymorphisms
2. Polymorphisms Associated with a Sequence Variant
3. Rare polymorphisms
4. Population or Allele Specific polymorphisms
5. "Old" polymorphisms (not population or allele specific)
6. Multiple polymorphisms in Haplotype

NIST National Institute of Standards and Technology • U.S. Department of Commerce



Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
No SNPs	D3S1358								
	FGA								
	D12S391								
	Penta E								
	D19S433								
	D21S11								
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0
	vWA	rs11063971	0.044	0.086	0.078	1	1	1	0
		rs11063970	0.044	0.086	0.078	1	1	1	
		rs11063969	0.044	0.086	0.078	1	1	1	
		rs75219269	0.044	0.086	0.078	1	1	1	
		rs199970098	0.029	-	0.011	2	-	1	
Rare SNPs	D2S441	rs74640515	0.015	0.014	0.067	1	1	2	2
	CSF1PO	C-T 36bp 5'	-	-	0.011	-	-	1	1
	D8S1179	rs138862078	0.007	-	-	1	-	-	1
	D10S1248	T-G 2bp 3'	-	0.007	-	-	1	-	1
	D18S51	rs535833682	0.029	-	-	3	-	-	1
	Penta D	rs7279663	0.022	0.014	0.011	1	2	1	2
	D22S1045	rs190864081	0.007	-	-	1	-	-	1
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1	4
		rs115644759	0.022	-	-	2	-	-	
		rs149212737	-	0.007	-	-	1	-	
	TH01	rs79373318	0.148	-	-	3	-	-	3
Single "Old" SNP (not population or allele specific)	D16S539	rs1728369	0.404	0.171	0.278	4	4	4	5
		G-A 94bp 5'	-	-	0.011	-	-	1	
	D2S1338	rs6736691	0.221	0.264	0.300	8	6	4	3
Multiple SNPs in Haplotype	D5S818	G-T 4bp 3'	0.316	0.257	0.189	7	5	6	17
		rs25768	0.228	0.257	0.156	5	5	5	
		rs146841551	0.029	-	-	1	-	-	
		rs541272009	0.007	-	-	1	-	-	
	D13S317	rs9546005	0.309	0.286	0.333	5	4	7	16
		rs73525369	0.066	-	-	3	-	-	
		rs73250432	-	0.014	0.011	-	2	1	
		rs146621667	0.015	-	-	2	-	-	
		4bp del 8bp 3'	-	-	0.022	-	-	2	
		4bp del 21bp 3'	0.007	-	0.022	1	-	1	
	D7S820	rs16887642	0.177	0.050	0.022	4	2	2	15
		rs7789995	0.015	0.164	0.122	2	4	4	
		rs7786079	0.176	0.021	0.011	5	2	1	
		1bp del 21bp 3'	-	-	0.011	-	-	1	

NIST National Institute of Standards and Technology • U.S. Department of Commerce

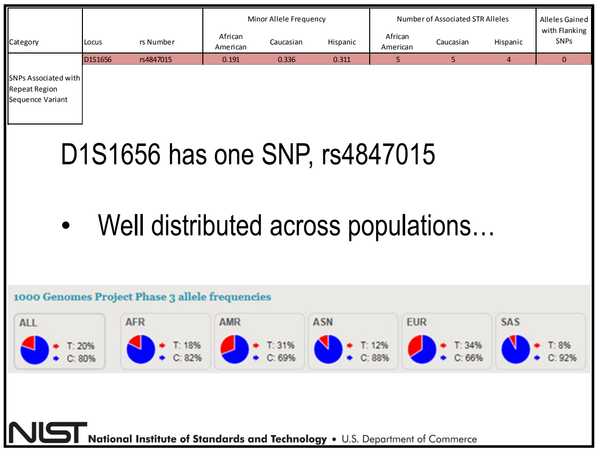


Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
No SNPs	D3S1358								
	FGA								
	D12S391								
	Penta E								
	D19S433								
	D21S11								

No SNPs were identified
in the sequenced flanking regions of:

D3S1358	Penta E
FGA	D19S433
D12S391	D21S11

NIST National Institute of Standards and Technology • U.S. Department of Commerce



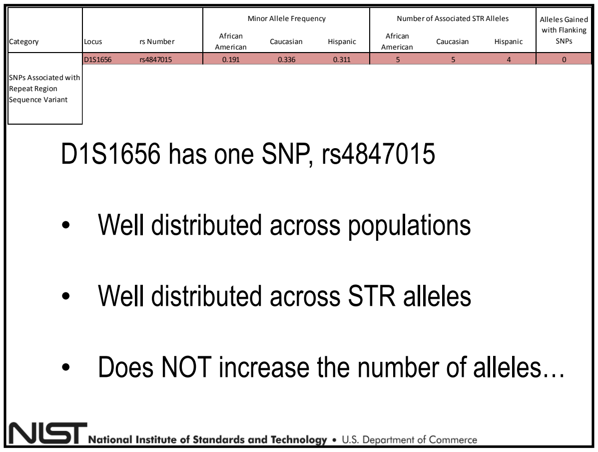
Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015

- Well distributed across populations…

1000 Genomes Project Phase 3 allele frequencies

NIST National Institute of Standards and Technology • U.S. Department of Commerce

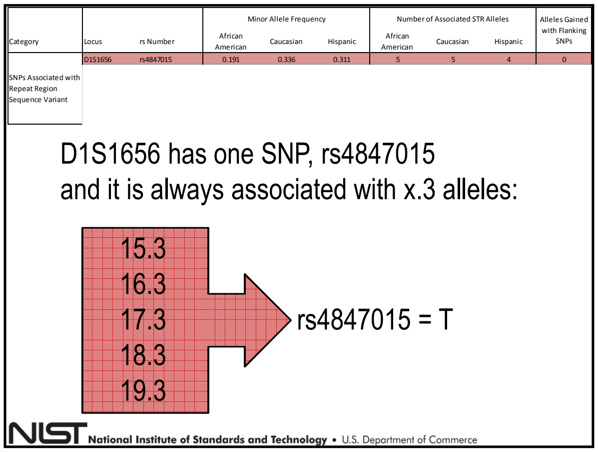


Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015

- Well distributed across populations
- Well distributed across STR alleles
- Does NOT increase the number of alleles…

NIST National Institute of Standards and Technology • U.S. Department of Commerce



Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015
and it is always associated with x.3 alleles:

15.3
16.3
17.3
18.3
19.3

rs4847015 = T

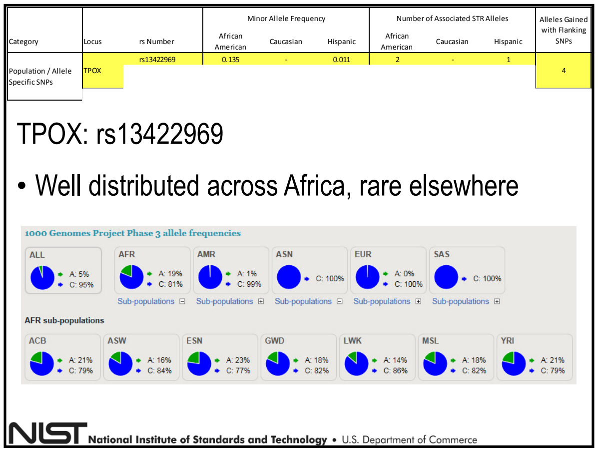
NIST National Institute of Standards and Technology • U.S. Department of Commerce

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Rare SNPs	D2S441	rs74640515	0.015	0.014	0.067	1	1	2	2
	CSF1PO	C-T 36bp 5'	-	-	0.011	-	-	1	1
	D8S1179	rs138862078	0.007	-	-	1	-	-	1
	D10S1248	T-G 2bp 3'	-	0.007	-	-	1	-	1
	D18S51	rs535833682	0.029	-	-	3	-	-	1
	Penta D	rs7279663	0.022	0.014	0.011	1	2	1	2
	D22S1045	rs190864081	0.007	-	-	1	-	-	1

Rare SNPs were identified
in the sequenced flanking regions of:

D2S441
CSF1PO
D8S1179
D10S1248
D18S51
Penta D
D22S1045

NIST National Institute of Standards and Technology • U.S. Department of Commerce

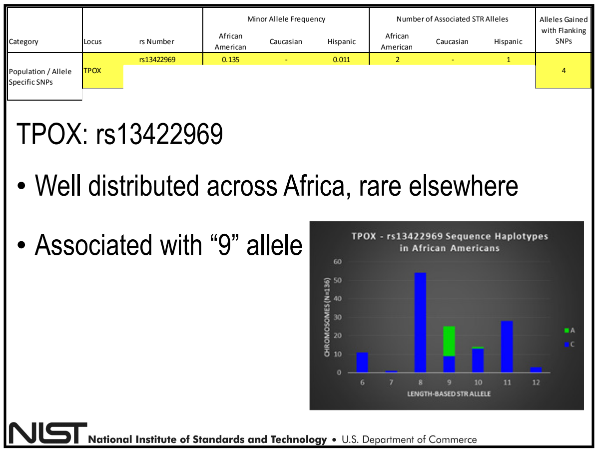


Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1	4

TPOX: rs13422969

- Well distributed across Africa, rare elsewhere

NIST National Institute of Standards and Technology • U.S. Department of Commerce

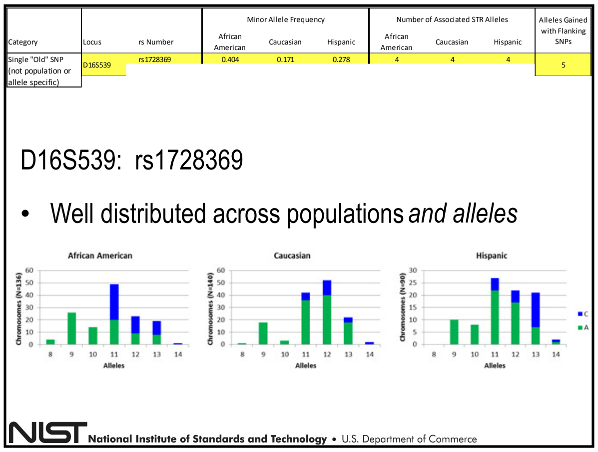


Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1	4

TPOX: rs13422969

- Well distributed across Africa, rare elsewhere
- Associated with “9” allele

NIST National Institute of Standards and Technology • U.S. Department of Commerce

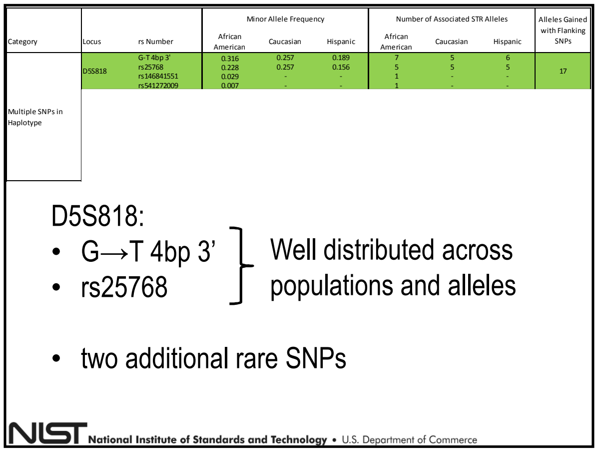


Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Single "Old" SNP (not population or allele specific)	D16S539	rs1728369	0.404	0.171	0.278	4	4	4	5

D16S539: rs1728369

- Well distributed across populations *and alleles*

National Institute of Standards and Technology • U.S. Department of Commerce



Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Multiple SNPs in Haplotype	D5S818	G-T 4bp 3'	0.316	0.257	0.189	7	5	6	17
		rs25768	0.228	0.257	0.156	5	5	5	
		rs146841551	0.029	-	-	1	-	-	
		rs541272009	0.007	-	-	1	-	-	

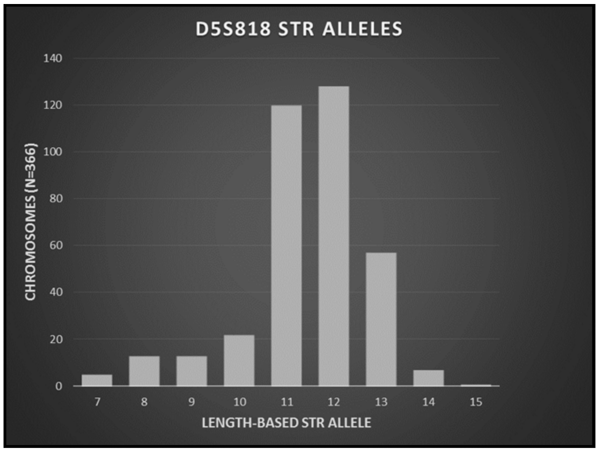
D5S818:

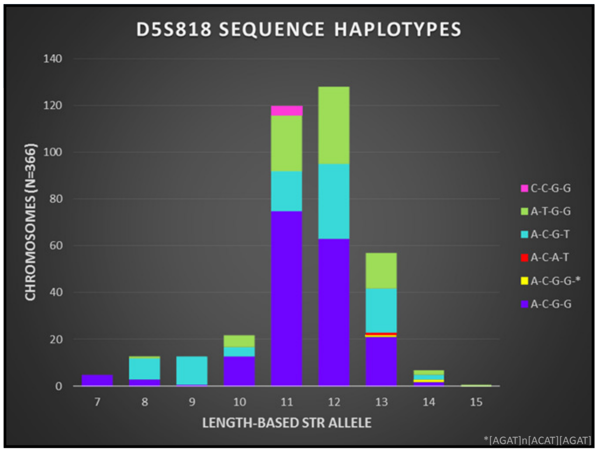
- G→T 4bp 3'
- rs25768

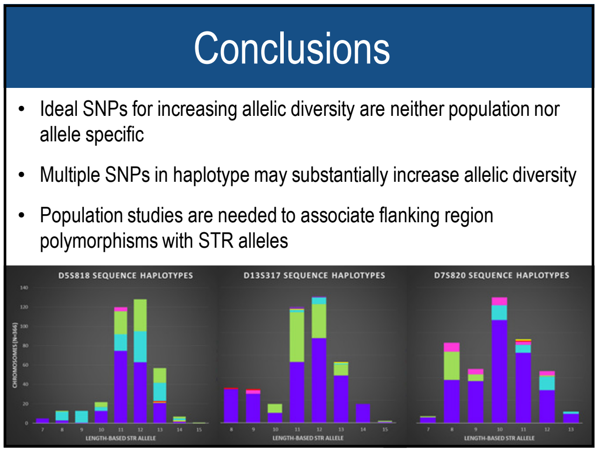
} Well distributed across populations and alleles

- two additional rare SNPs

National Institute of Standards and Technology • U.S. Department of Commerce

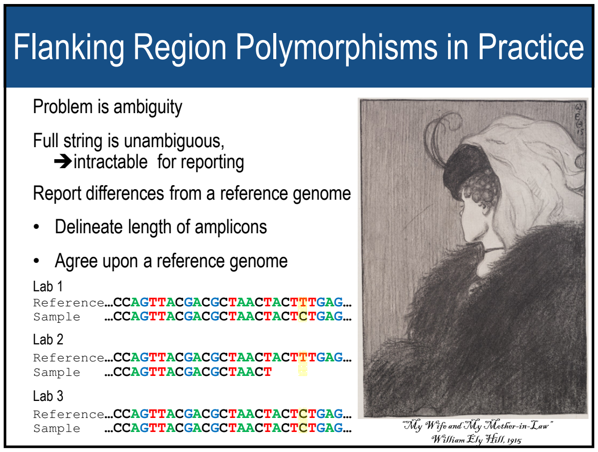






Conclusions

- Ideal SNPs for increasing allelic diversity are neither population nor allele specific
- Multiple SNPs in haplotype may substantially increase allelic diversity
- Population studies are needed to associate flanking region polymorphisms with STR alleles



Flanking Region Polymorphisms in Practice

Problem is ambiguity

Full string is unambiguous,
→intractable for reporting

Report differences from a reference genome

- Delineate length of amplicons
- Agree upon a reference genome

Lab 1

```
Reference…CCAGTTACGACGCTAACTACTTTGAG…
Sample   …CCAGTTACGACGCTAACTACTCTGAG…
```

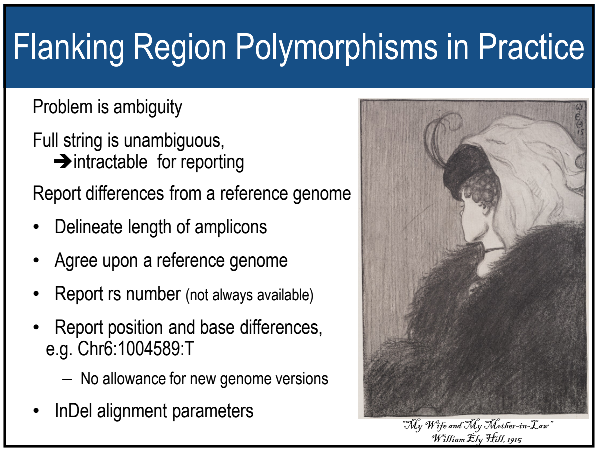
Lab 2

```
Reference…CCAGTTACGACGCTAACTACTTTGAG…
Sample   …CCAGTTACGACGCTAACT
```

Lab 3

```
Reference…CCAGTTACGACGCTAACTACTCTGAG…
Sample   …CCAGTTACGACGCTAACTACTCTGAG…
```

"My Wife and My Mother-in-Law"
William Ely Hill, 1915



Flanking Region Polymorphisms in Practice

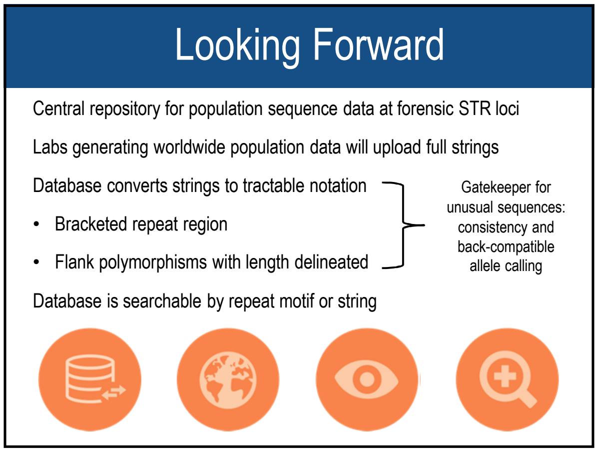
Problem is ambiguity

Full string is unambiguous,
➔intractable for reporting

Report differences from a reference genome

- Delineate length of amplicons
- Agree upon a reference genome
- Report rs number (not always available)
- Report position and base differences, e.g. Chr6:1004589:T
  - No allowance for new genome versions
- InDel alignment parameters

"My Wife and My Mother-in-Law"
William Ely Hill, 1915



Looking Forward

Central repository for population sequence data at forensic STR loci

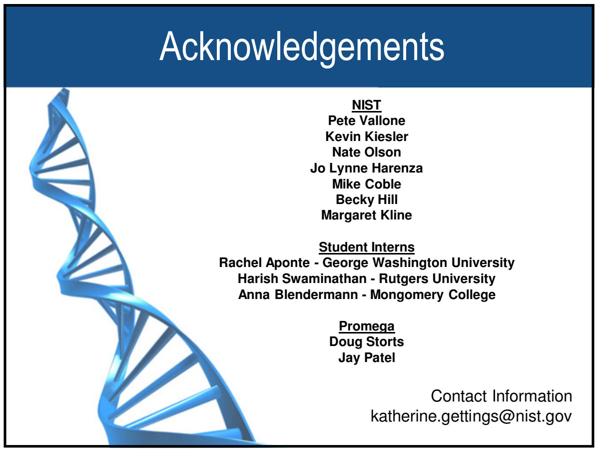
Labs generating worldwide population data will upload full strings

Database converts strings to tractable notation

- Bracketed repeat region
- Flank polymorphisms with length delineated

Gatekeeper for unusual sequences: consistency and back-compatible allele calling

Database is searchable by repeat motif or string



Acknowledgements

NIST
Pete Vallone
Kevin Kiesler
Nate Olson
Jo Lynne Harenza
Mike Coble
Becky Hill
Margaret Kline

Student Interns
Rachel Aponte - George Washington University
Harish Swaminathan - Rutgers University
Anna Blendermann - Mongomery College

Promega
Doug Storts
Jay Patel

Contact Information
katherine.gettings@nist.gov