

U.S. Population Sequence Data for 27 Autosomal STR Loci, 24 YSTR Loci and 7 XSTR Loci

Katherine Gettings, Ph.D., Kevin Kiesler, M.S., Becky Steffen, M.S., Lisa Borsuk, M.S., and Peter Vallone, Ph.D.

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

MATERIAL MEASUREMENT LABORATORY

Official Disclaimer

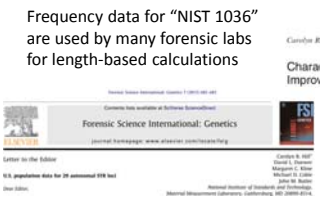
The opinions and assertions contained herein are solely those of the author and are not to be construed as official or as views of the U.S. Department of Commerce.

Commercial equipment, instruments, software, or materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the U.S. Department of Commerce, nor does it imply that any of the materials, instruments, software or equipment identified are necessarily the best available for the purpose.

NIST

Sequencing Forensic STRs in Population Samples

Frequency data for "NIST 1036" are used by many forensic labs for length-based calculations


Forensic Science International: Genetics

Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples

D4S2408

Allele	Total	Cauc.	Afr. Am.	Hisp.
7	0.0015		0.0039	
8	0.1904	0.2222	0.1417	0.2194
9	0.2791	**0.3161**	0.1870	**0.3777**
10	0.2301	0.2375	0.2441	0.1906
11	0.2378	0.1973	**0.3189**	0.1655
12	0.0596	0.0249	0.1024	0.0468
13	0.0015	0.0019	0.0020	

This new data set will serve the same purpose for **sequenced-based** calculations

NIST

genomeweb

Illumina Launches MiSeq FGx for Forensic Applications

Jan 21, 2015 | Monica Heger

Premium

NEW YORK (GenomeWeb) – Illumina has launched the MiSeq FGx Forensic Genomics System, a next-generation sequencing system validated specifically for forensic applications, the company said today.

The system includes the MiSeq FGx sequencing instrument, the ForenSeq DNA Signature Prep Kit, and ForenSeq Universal Analysis software. It evaluates both short tandem repeats (STRs) and SNPs, and is compatible with existing DNA databases like the Combined DNA Index System (CODIS).

27 autosomal STR loci
24 YSTR loci
7 XSTR loci
Identity, Ancestry and Phenotype SNPs

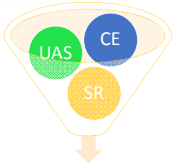
NIST

Data analysis

1. Length based genotypes from CE for **all samples at all loci**
2. Sequences and length genotypes from Illumina's **U**niversal **A**nalysis **S**oftware
3. Full string analysis with in-house pipeline based on:

STRait Razor v2.0: The improved STR Allele Identification Tool – Razor

David H. Warshauer, Jonathan L. King, Bruce Budowle



Final allele sequences

NIST

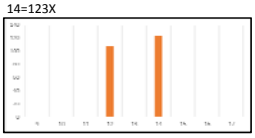
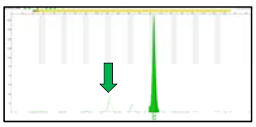
Discordance with CE Data

D9S1122

UAS: 12,14 NIST CE Assay: 14,14

12=107X

14=123X

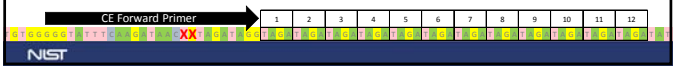
Flanking region analysis revealed a 2bp del in 12 allele (rs754976988)

NIST

Discordance with CE Data

D9S1122

1. NIST CE Assay called 14,14.
2. ForenSeq UAS called 12,14.
3. A larger CE amplicon would genotype as 11.2,14.
4. The full sequence reveals 12 (rs754976988, --/TG), 14.



NIST

Summary of Discord

Method	Null	Imbalance
ForenSeq	1	3*
CE‡	2	0
Both	2	0
Total	5	3

‡across CE kits *excluding D22S1045

Analysis Method	Flanking Indel
UAS vs CE‡	10
Custom vs CE‡	3*

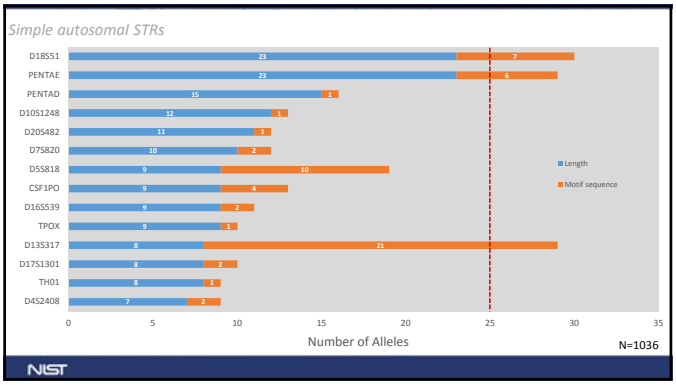
‡across CE kits *these also vary across CE assays

27 STR, 24 YSTR and 7 XSTR loci across 1036 samples
>100,000 alleles sequenced

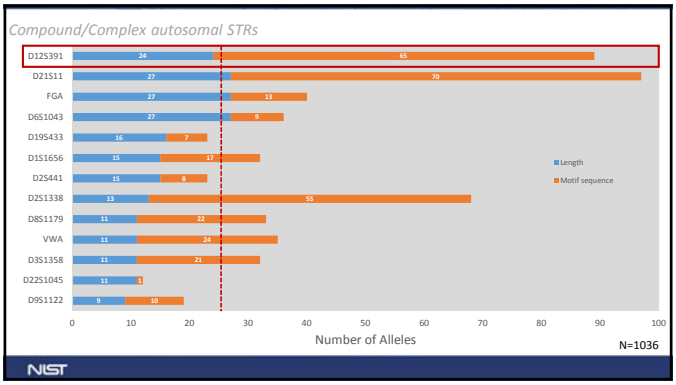
18 differences
(4 involve core loci)
>99.98% concordant

Concordance is improved by including NGS flanking region

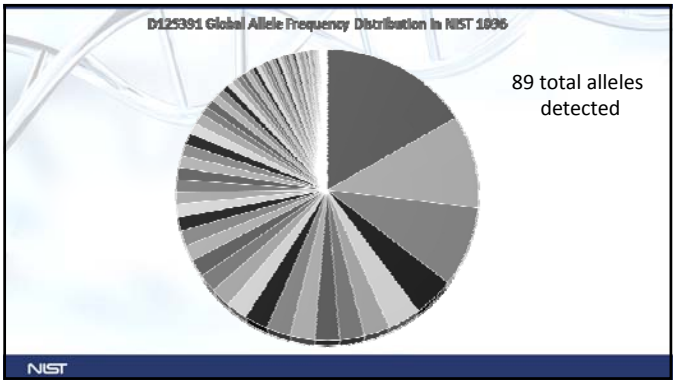
NIST



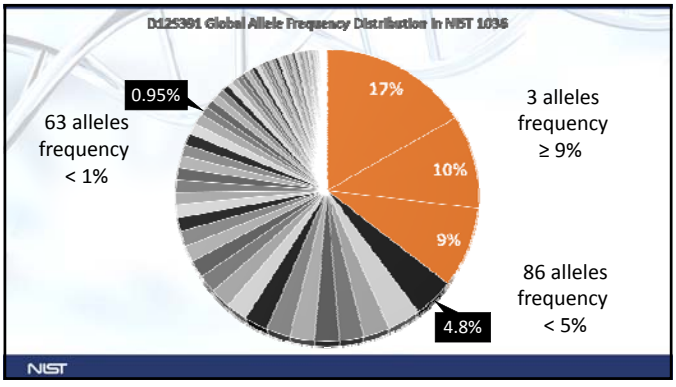
NIST



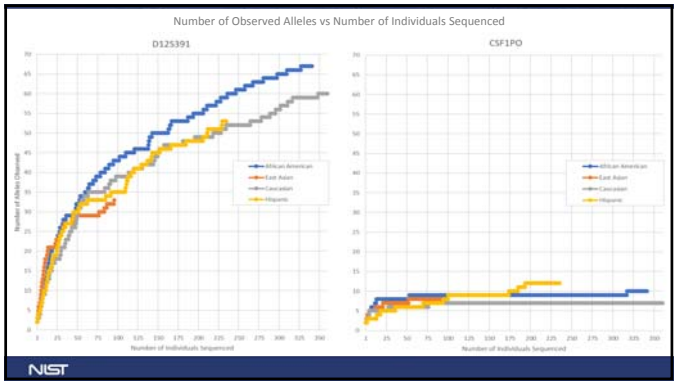
NIST

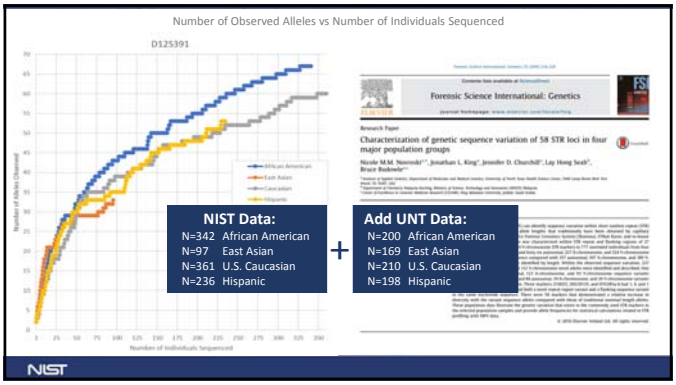


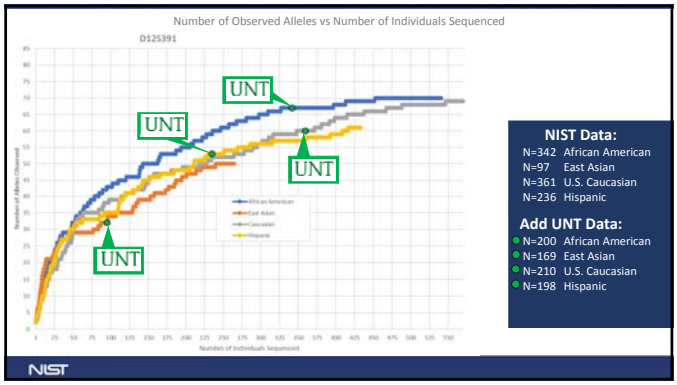
NIST

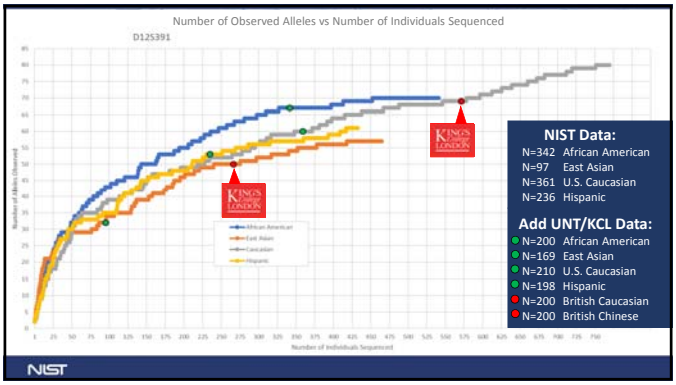


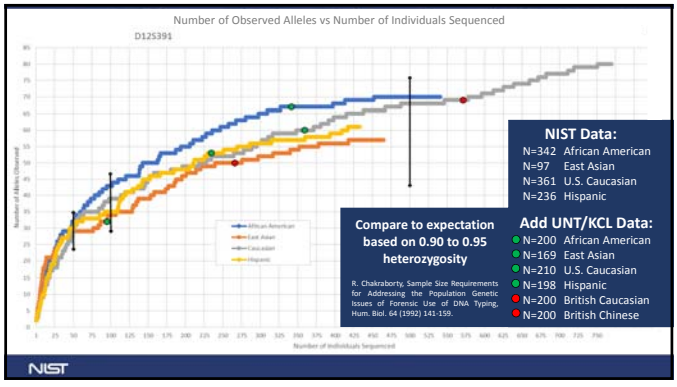
NIST

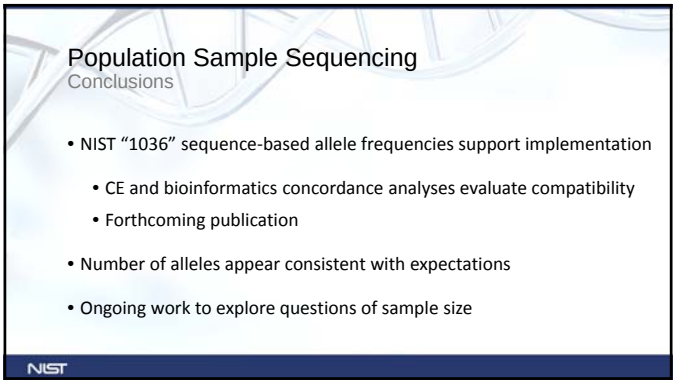

Number of Observed Alleles vs Number of Individuals Sequenced
D12S391
CSF1PO
NIST


Number of Observed Alleles vs Number of Individuals Sequenced
D12S391
Forensic Science International: Genetics
Characterization of genetic sequence variation of 58 STR loci in four major population groups
NIST Data:
N=342 African American
N=97 East Asian
N=361 U.S. Caucasian
N=236 Hispanic
+
Add UNT Data:
N=200 African American
N=169 East Asian
N=210 U.S. Caucasian
N=198 Hispanic
NIST


Number of Observed Alleles vs Number of Individuals Sequenced
D12S391
UNT
UNT
UNT
UNT
NIST Data:
N=342 African American
N=97 East Asian
N=361 U.S. Caucasian
N=236 Hispanic
Add UNT Data:
N=200 African American
N=169 East Asian
N=210 U.S. Caucasian
N=198 Hispanic
NIST


Number of Observed Alleles vs Number of Individuals Sequenced
D12S391
KCL
KCL
NIST Data:
N=342 African American
N=97 East Asian
N=361 U.S. Caucasian
N=236 Hispanic
Add UNT/KCL Data:
N=200 African American
N=169 East Asian
N=210 U.S. Caucasian
N=198 Hispanic
N=200 British Caucasian
N=200 British Chinese
NIST


Number of Observed Alleles vs Number of Individuals Sequenced
D12S391
NIST Data:
N=342 African American
N=97 East Asian
N=361 U.S. Caucasian
N=236 Hispanic
Compare to expectation based on 0.90 to 0.95 heterozygosity
R. Chakraborty, Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing, Hum. Biol. 64 (1992) 141-159.
Add UNT/KCL Data:
N=200 African American
N=169 East Asian
N=210 U.S. Caucasian
N=198 Hispanic
N=200 British Caucasian
N=200 British Chinese
NIST


Population Sample Sequencing
Conclusions
• NIST "1036" sequence-based allele frequencies support implementation
• CE and bioinformatics concordance analyses evaluate compatibility
• Forthcoming publication
• Number of alleles appear consistent with expectations
• Ongoing work to explore questions of sample size
NIST

NCBI BioProject—STRseq

Mission: To provide high-confidence STR allele sequence records with uniform annotation, facilitating exchange of information across forensic laboratories.

- Collaborators with large datasets "seed" the BioProject
- NIST evaluates raw sequence data with agnostic bioinformatic pipeline
- GenBank record for **all unique sequences**
- BioProject searchable by string (BLAST), locus, allele…

NIST

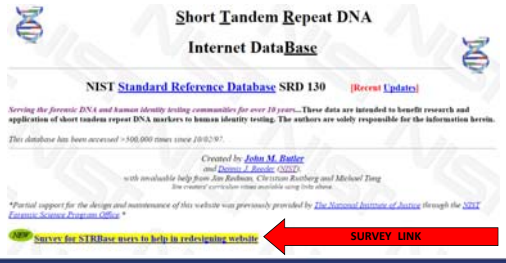
NCBI BioProject STRseq

Publication containing sequence	
Description of STRseq	
Metadata	Kit / Instrument / Quality Length-based allele Location on reference genome
Annotation	SNPs or InDels Bracketed Sequence
Sequence String	

NIST

STRBase Update

Please go to http://www.cstl.nist.gov/strbase *for survey:*



NIST

Acknowledgments

NIST Applied Genetics
Dr. Peter Vallone
Lisa Borsuk
Kevin Kiesler
Becky Steffen

NIST Special Programs
Dr. John Butler

Collaborators
Dr. Kim Pruitt
Dr. Ilene Mizrachi
Lori Black
Dr. George Riley
Dr. Christopher Phillips
Jonathan King
Dr. Walther Parson
Dr. David Ballard
Laurence Devesse

Funding

NIST

katherine.gettings@nist.gov