Population Sample Sequencing at NIST

FORENSICS@NIST – Forensic Genetics Session

9 November 2016

Katherine Butler Gettings, PhD
Research Biologist, Applied Genetics Group

Dr. Peter Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen

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.

Forensic DNA Current Technology:

Capillary Electrophoresis (CE)

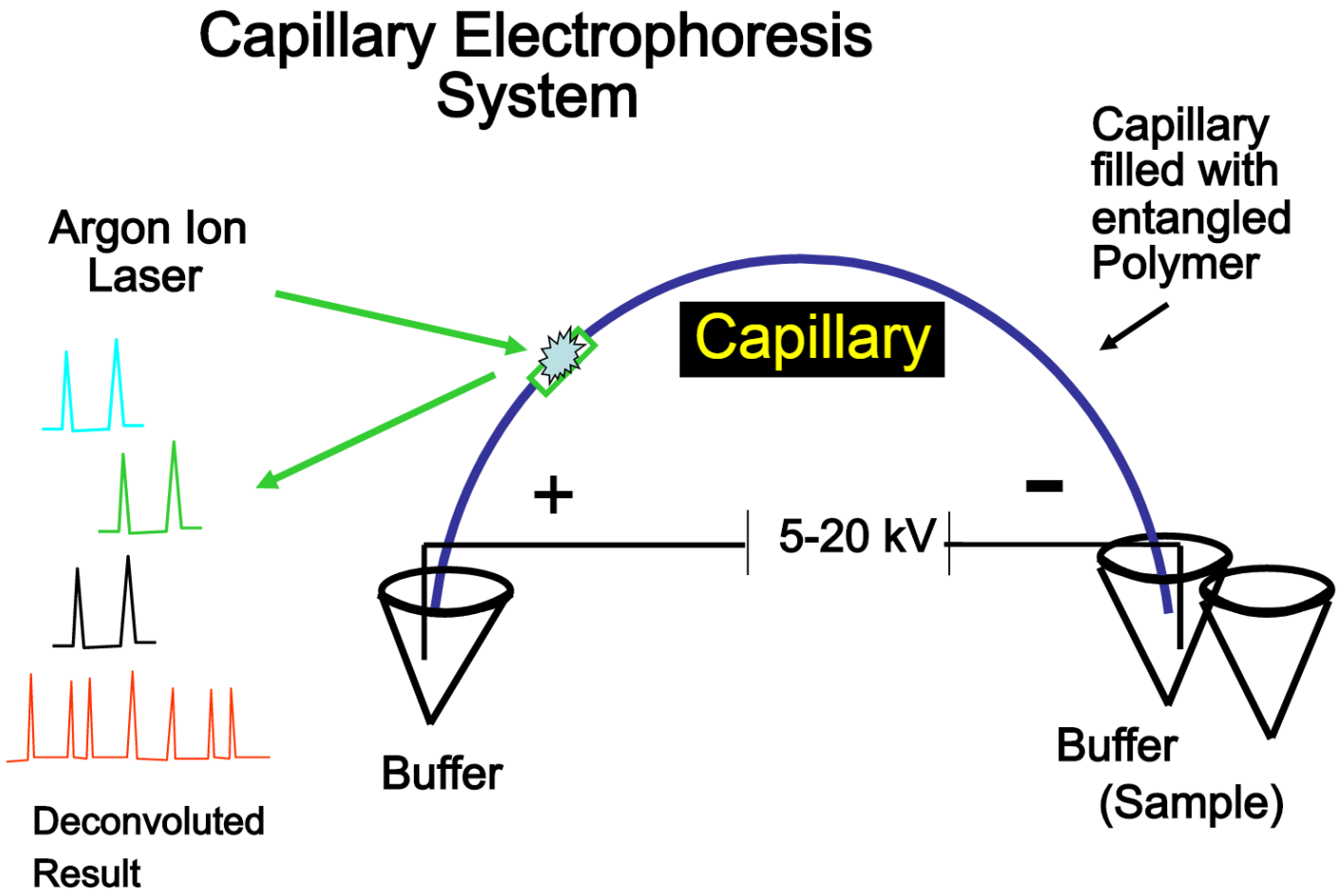
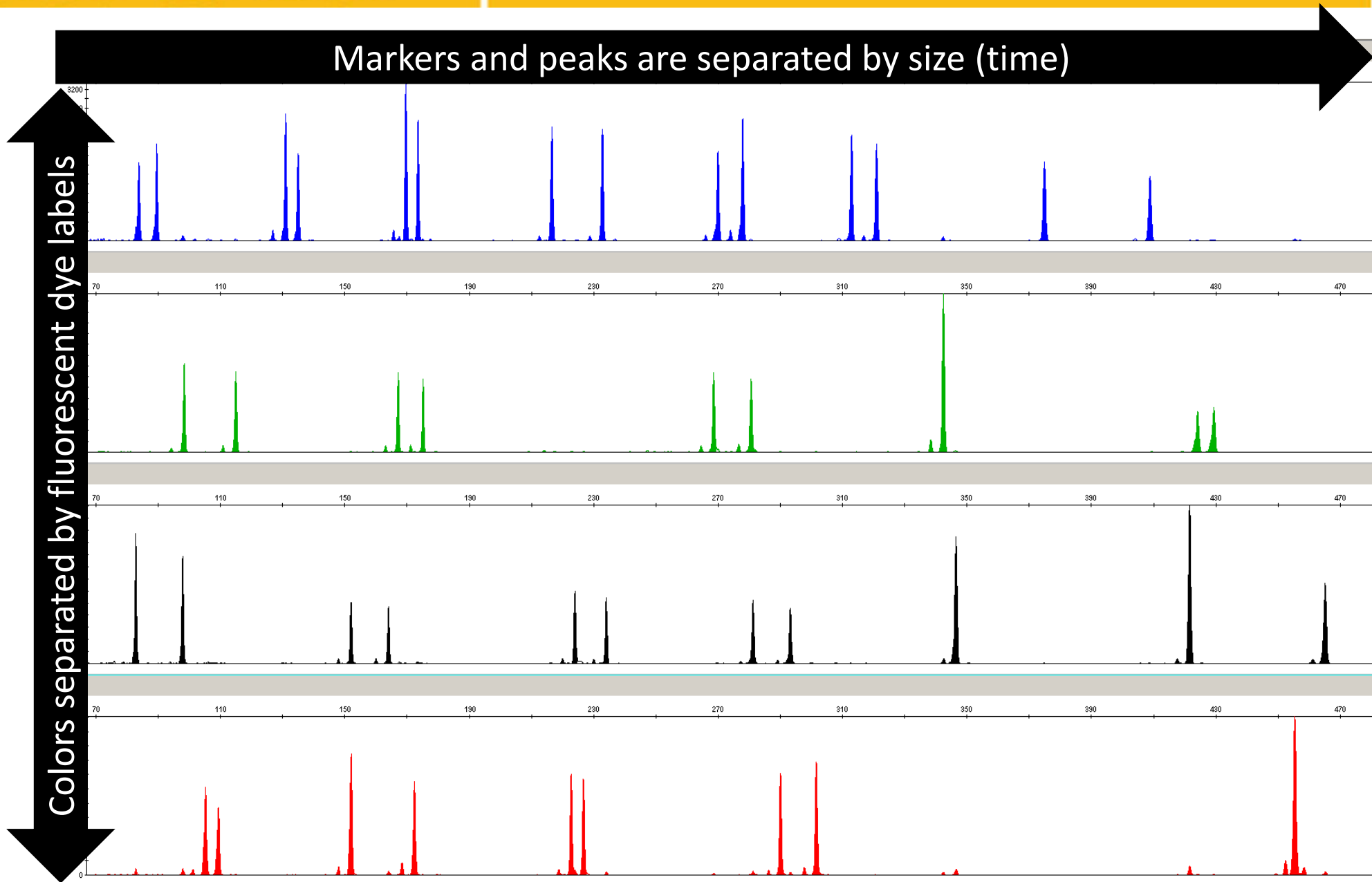


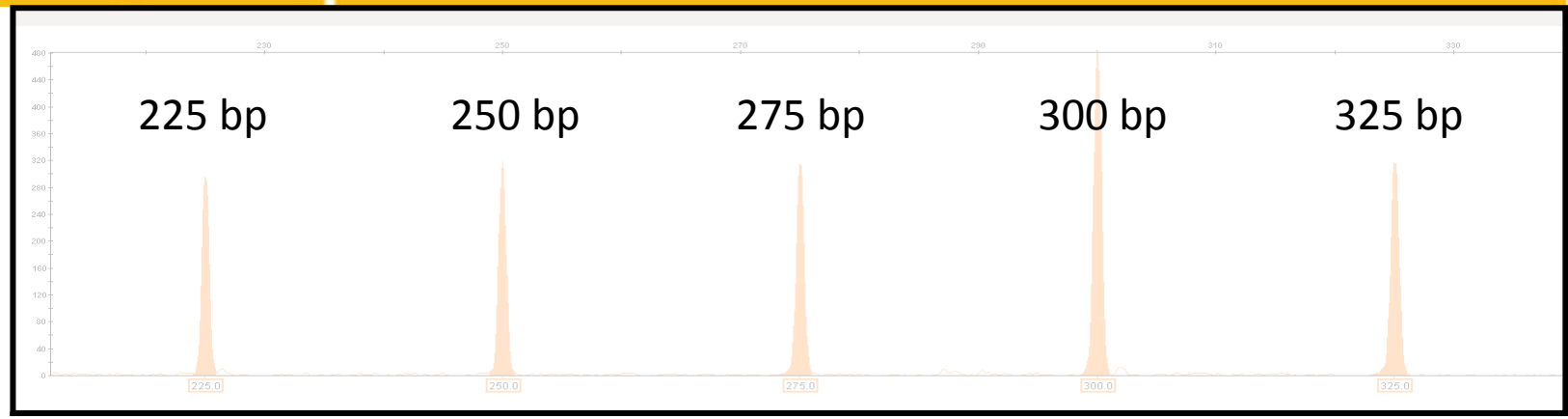
Image courtesy Bruce McCord, Florida International University

Current Technology: CE Electro-pherogram

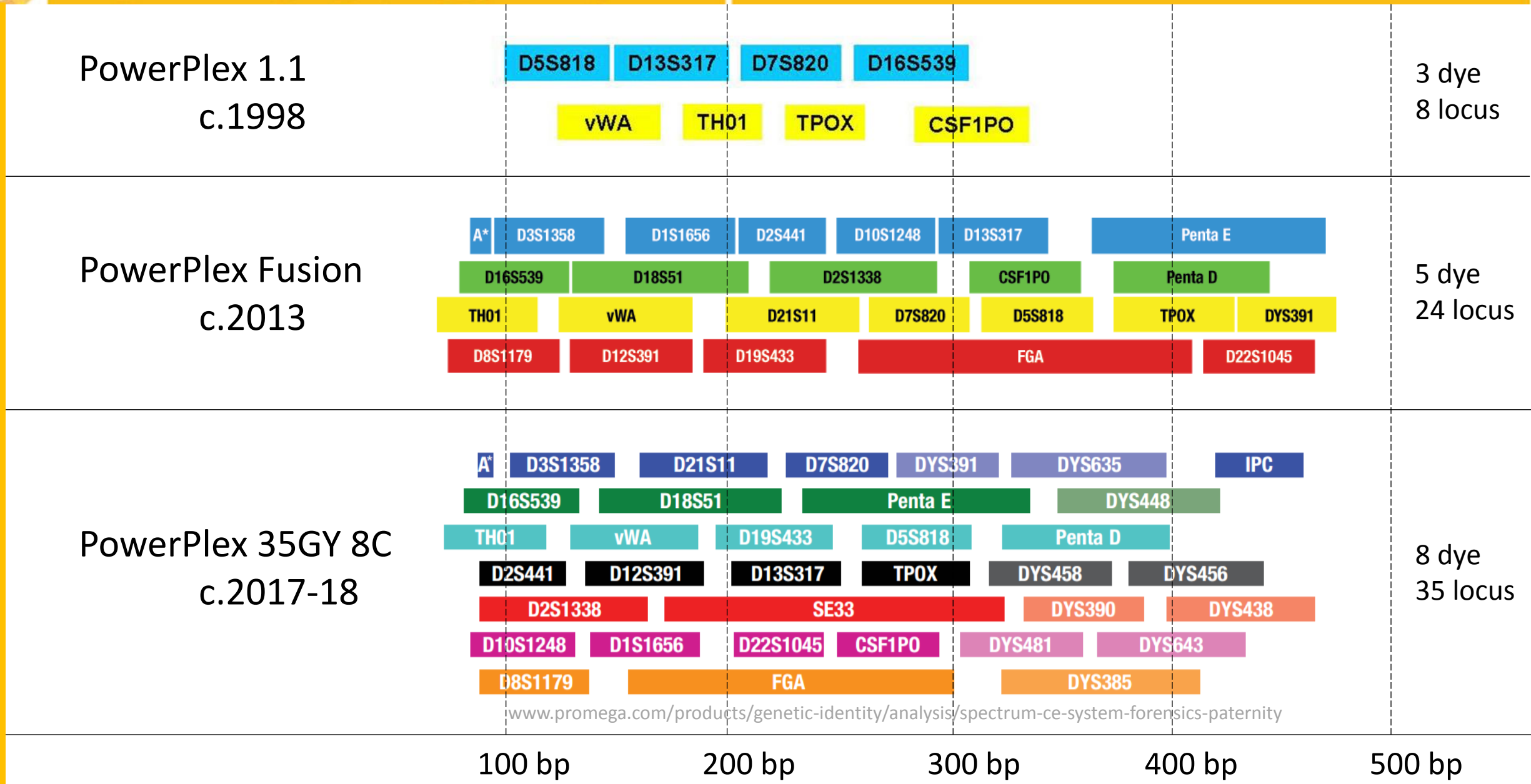


CE Analysis

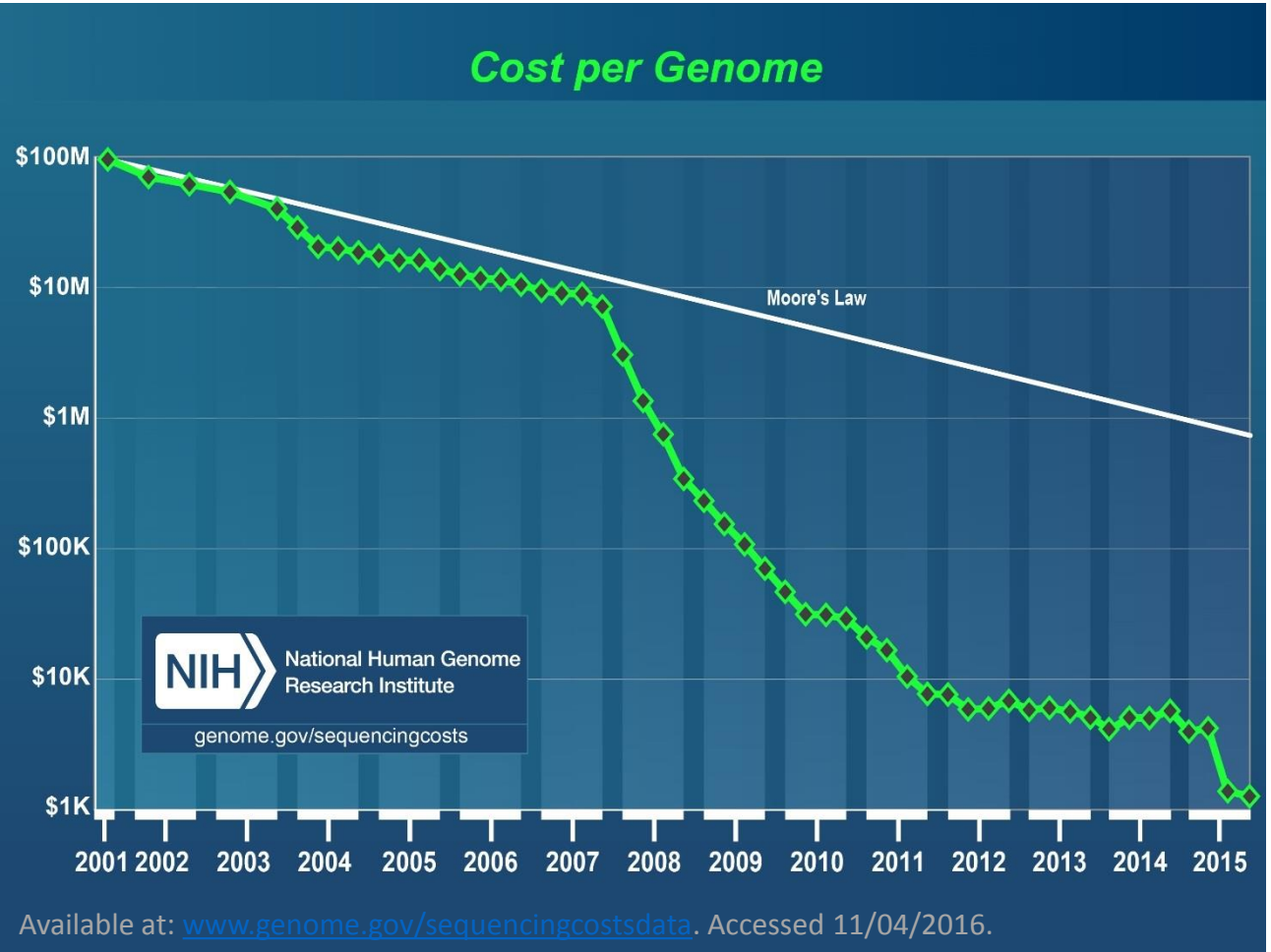
Internal Lane Standard



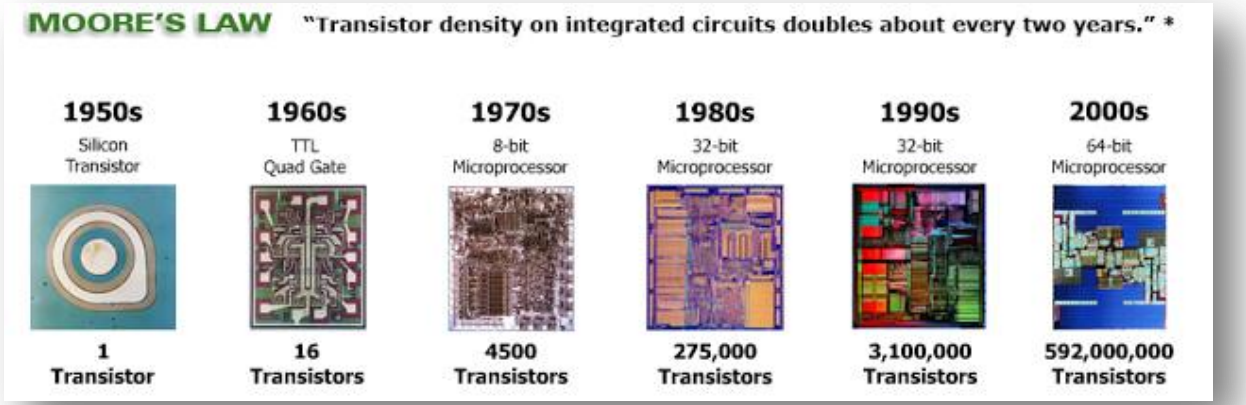
CE Multiplex Evolution Examples



Advances in DNA Sequencing



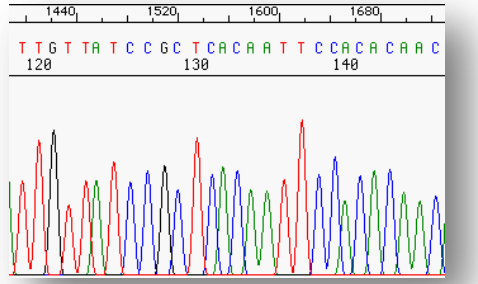
Available at: www.genome.gov/sequencingcostsdata. Accessed 11/04/2016.



http://electronicsbyexamples.blogspot.com/2013/03/milestones-in-digital-electronics.html

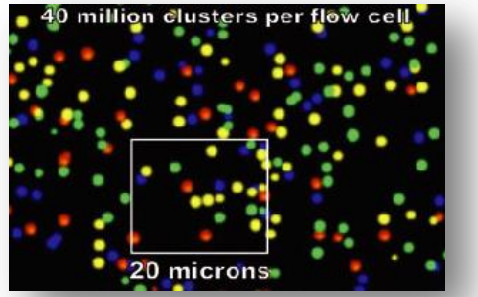
2001:

First human genome published,
requiring 15 years of effort
at a cost of 3 billion dollars



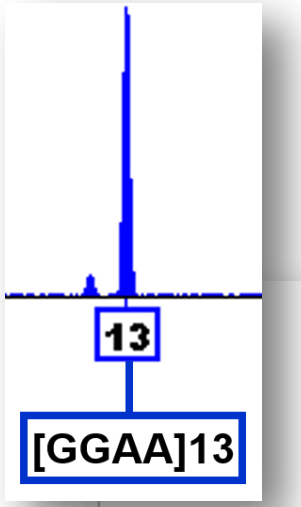
2014:

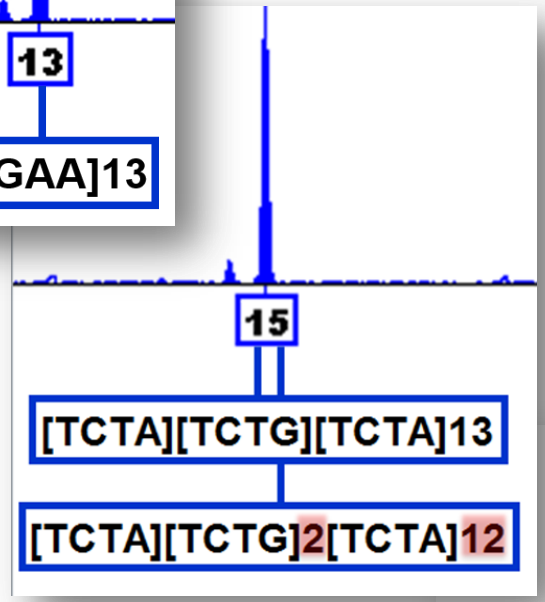
One instrument can sequence
45 human genomes in one day
for $1000 each

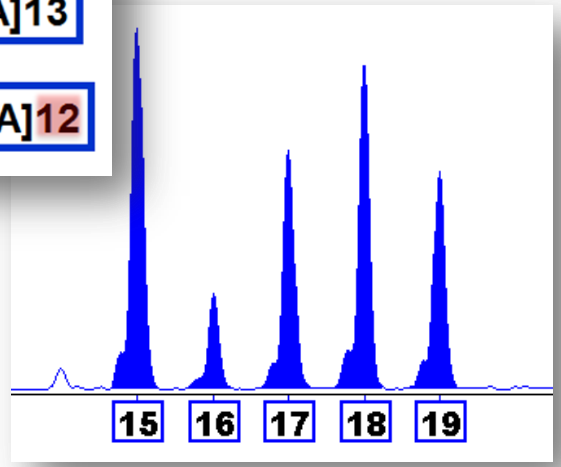


Sequencing Forensic STRs

- STR = Short Tandem Repeat
  - Example [GGAA]13
- Targeted sequencing reveals sequence variation within STR amplicons
- Greater degree of multiplexing
  - Not confined by dye colors; smaller PCR amplicons
  - Other loci (SNPs)







Forensic NGS Kits for STR and SNP Typing

Illumina ForenSeq and FGx

Applied Biosystems Precision ID STR, SNP and Mixture Panels

Promega PowerSeq Auto/YSTR/Mito System

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

ForenSeq

- 27 autosomal STR loci

When a match is made in a forensic case, allele frequencies are used to calculate how common or rare the DNA profile is in a given population

J Forensic Sci, January 2008, Vol. 53, No. 1
doi: 10.1111/j.1556-4029.2008.00595.x
Available online at: www.blackwell-synergy.com

Carolyn R. Hill, M.S.; Margaret C. Kline, M.S.; Michael D. Coble,† Ph.D.; and John M. Butler, Ph.D.

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	

Forensic Science International: Genetics 7 (2013) e82–e83

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Letter to the Editor

U.S. population data for 29 autosomal STR loci

Carolyn R. Hill*
David L. Duewer
Margaret C. Kline
Michael D. Coble
John M. Butler
National Institute of Standards and Technology, Material Measurement Laboratory, Gaithersburg, MD 20899-8314,

Dear Editor,

ForenSeq

- 27 autosomal STR loci

When a match is made in a forensic case, allele frequencies are used to calculate how common or rare the DNA profile is in a given population

D4S2408

Allele	N	Freq	Sequence Allele	N	Freq
7	1	0.6%	[ATCT]7	1	0.6%
8	23	14.4%	[ATCT]8	23	14.4%
9	60	37.5%	[ATCT]9	18	11.3%
			[ATCT] GTCT [ATCT]7	42	26.3%
10	53	33.1%	[ATCT]10	53	33.1%
11	21	13.1%	[ATCT]11	21	13.1%
12	2	1.3%	[ATCT]12	2	1.3%

Example data for illustration purposes only

Example: Profile from crime scene is single source, and matches POI.

D4S2408

Allele	N	Freq	Sequence Allele	N	Freq
7	1	0.6%	[ATCT]7	1	0.6%
8	23	14.4%	[ATCT]8	23	14.4%
9	60	37.5%	[ATCT]9	18	11.3%
			[ATCT] **G**TCT [ATCT]7	42	26.3%
10	53	33.1%	[ATCT]10	53	33.1%
11	21	13.1%	[ATCT]11	21	13.1%
12	2	1.3%	[ATCT]12	2	1.3%

Example data for illustration purposes only

At D4S2408, the length genotype is 9,10.

The Random Match Probability (RMP) is 2*p*q. p and q are the frequencies of 9 and 10.

The length-based RMP is 2pq = 2(0.375)(0.331) = 0.248, meaning approximately **25%** of individuals in this population would not be excluded as possible contributors.

At D4S2408, the sequence genotype is [ATCT]9, [ATCT]10.

The associated statistic is 2*p*q. p and q are the frequencies of [ATCT]9 and [ATCT]10.

The sequence-based RMP is 2pq = 2(0.113)(0.331) = 0.074, meaning approximately **7%** of individuals in this population would not be excluded as possible contributors.

ForenSeq™ admin@nist.gov illumina

PROJECT: 2015_12_23_NEQUALS_32 ▸ ANALYSIS: 1.0 - 32 SAMPLE RUN ▸ SAMPLE:S226 ▸ POPULATION STATISTICS

Project Name

1.0 - 32 sample run
Completed 25 Dec 2015
Control Review:
1 Samples
P N Q

S226

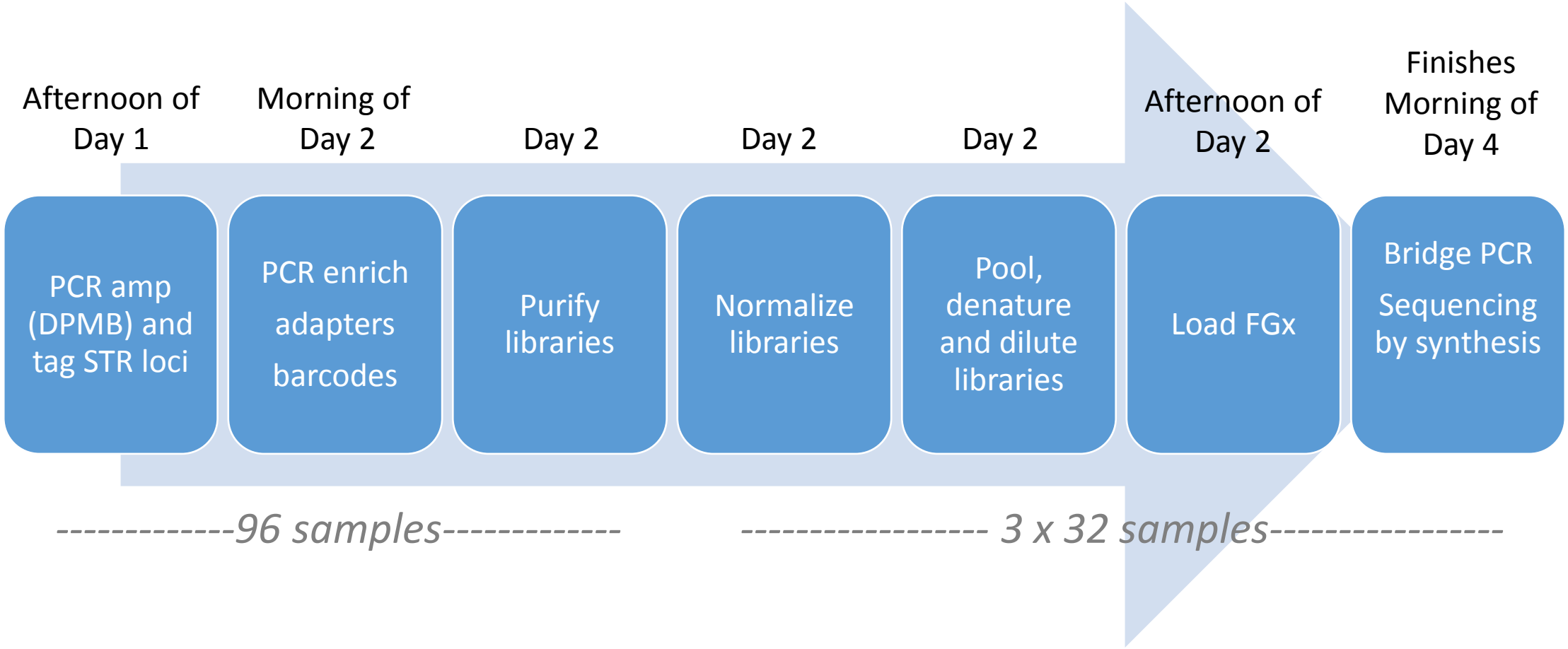
Population Statistics

Length-based allele frequencies:

(Sample)
S226
1.0 - 32 sample run

African American: NIST 1036 U.S. Population Dataset	1 in 2.712E+28
Asian: NIST 1036 U.S. Population Dataset	1 in 4.639E+33
Caucasian: NIST 1036 U.S. Population Dataset	1 in 3.093E+32
Hispanic: NIST 1036 U.S. Population Dataset	1 in 7.139E+31

General workflow for 96 samples at NIST

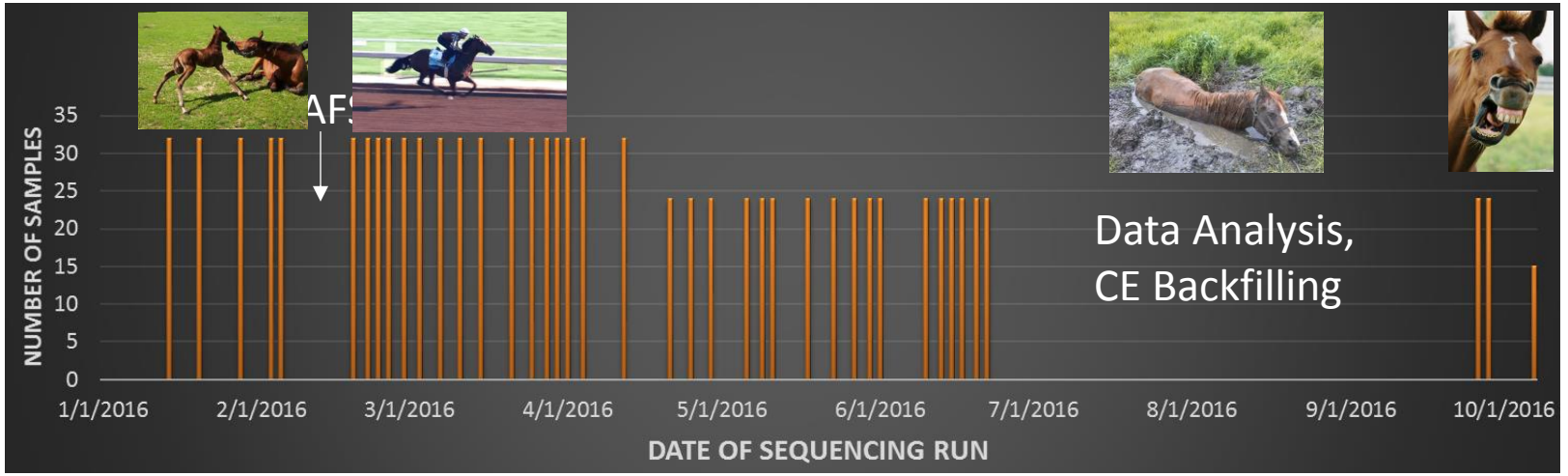


3rd plate is loaded afternoon of day 6, 96 samples complete on day 8

Population Sample Sequencing Metrics

1036 Samples from four population groups

- 342 African American
- 361 Caucasian
- 97 East Asian
- 236 Hispanic



Sequenced in batches of 24 or 32

41 total sequencing runs

Additional CE analysis to provide complete data for concordance check

Data analysis

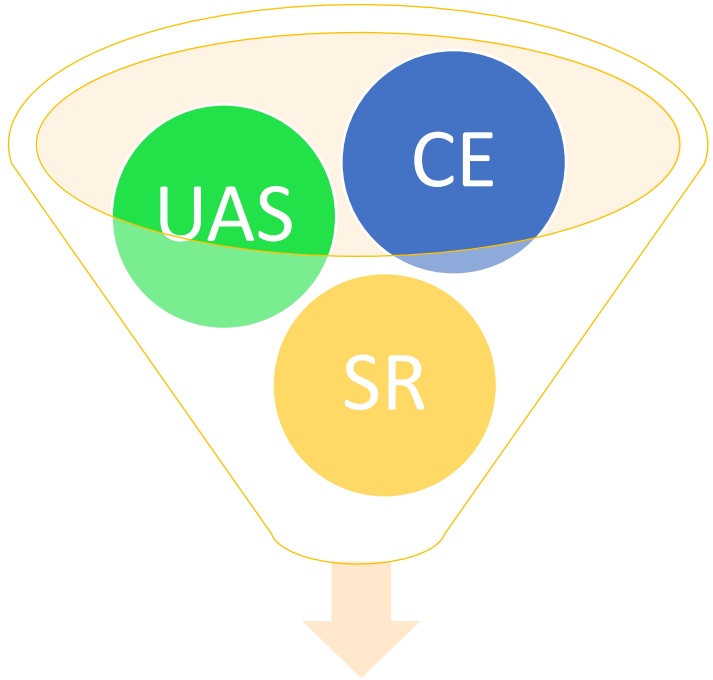
- Length based genotypes from CE
- Sequences and length genotypes from UAS

  Illumina's **U**niversal **A**nalysis **S**oftware
- Tandem analysis with in-house pipeline based on:

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

David H. Warshauer [a], Jonathan L. King [a], Bruce Budowle [a,b,*]

[a] *Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Boulevard, Fort Worth, TX 76107, USA*
[b] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia*



Final allele sequences

Slides have been removed from the original presentation, manuscript under preparation

Population Sample Sequencing

Conclusions

- Many forensic STR loci contain underlying sequence variation
- This will increase allelic diversity, thus increasing the ability to discriminate among individuals in a mixture
- NIST “1036” sequence-based allele frequencies support implementation
- CE concordance analysis ensures back-compatibility

Acknowledgments

Dr. Peter Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen

Funding
- NIST SPO - Forensic DNA
- FBI - DNA as a Biometric
- NIJ

Questions? katherine.gettings@nist.gov