FORENSICS @ NIST #NISTForensics

The Use of Statistical Analysis to Characterize Noise and Zygosity in Targeted Sequencing of Forensic STR Markers

Sarah Riman[1], PhD; Hari Iyer[2], PhD; Lisa Borsuk[1], MS; Peter M. Vallone[1], PhD

[1]Applied Genetics Group
[2]Statistical Design, Analysis, and Modeling Group

NIST APPLIED GENETICS

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

Disclaimer

Points of view in this presentation are mine and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Commerce.

NIST Disclaimer Certain commercial products and instruments are identified in order to specify experimental procedures as completely as possible. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of these products are necessarily the best available for the purpose.

Overview

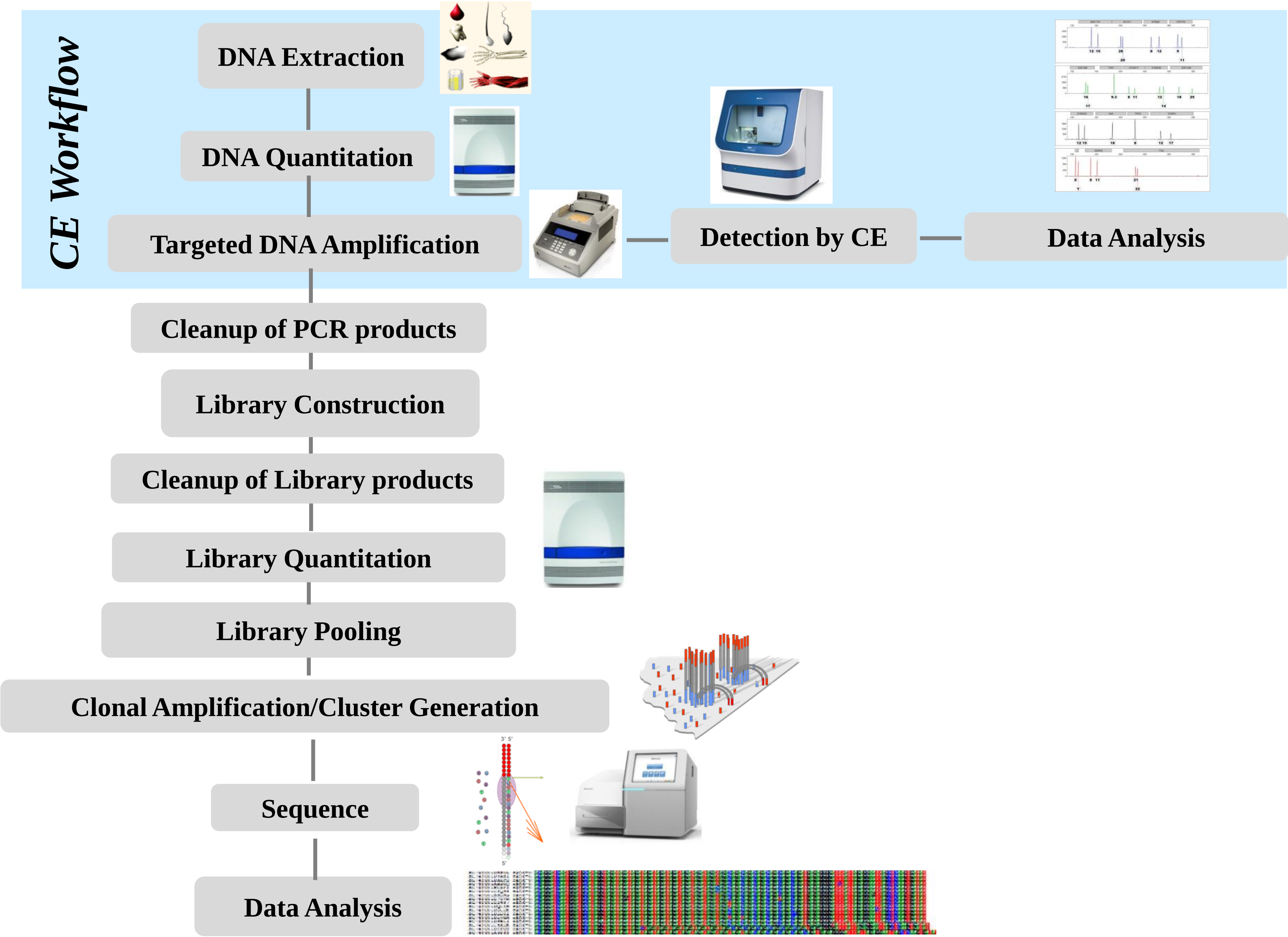
- Forensic DNA typing using the gold standard “Capillary Electrophoresis” (CE) technology vs. “Next Generation Sequencing” (NGS) technology

- Why implement NGS if you can accomplish DNA typing by CE?

- Characterization of single-source PowerSeq 46GY DNA profiles

Forensic DNA typing using the gold standard “Capillary Electrophoresis” (CE) technology vs. “Next Generation Sequencing” (NGS) technology

General Workflow for Next Generation Sequencing (NGS)

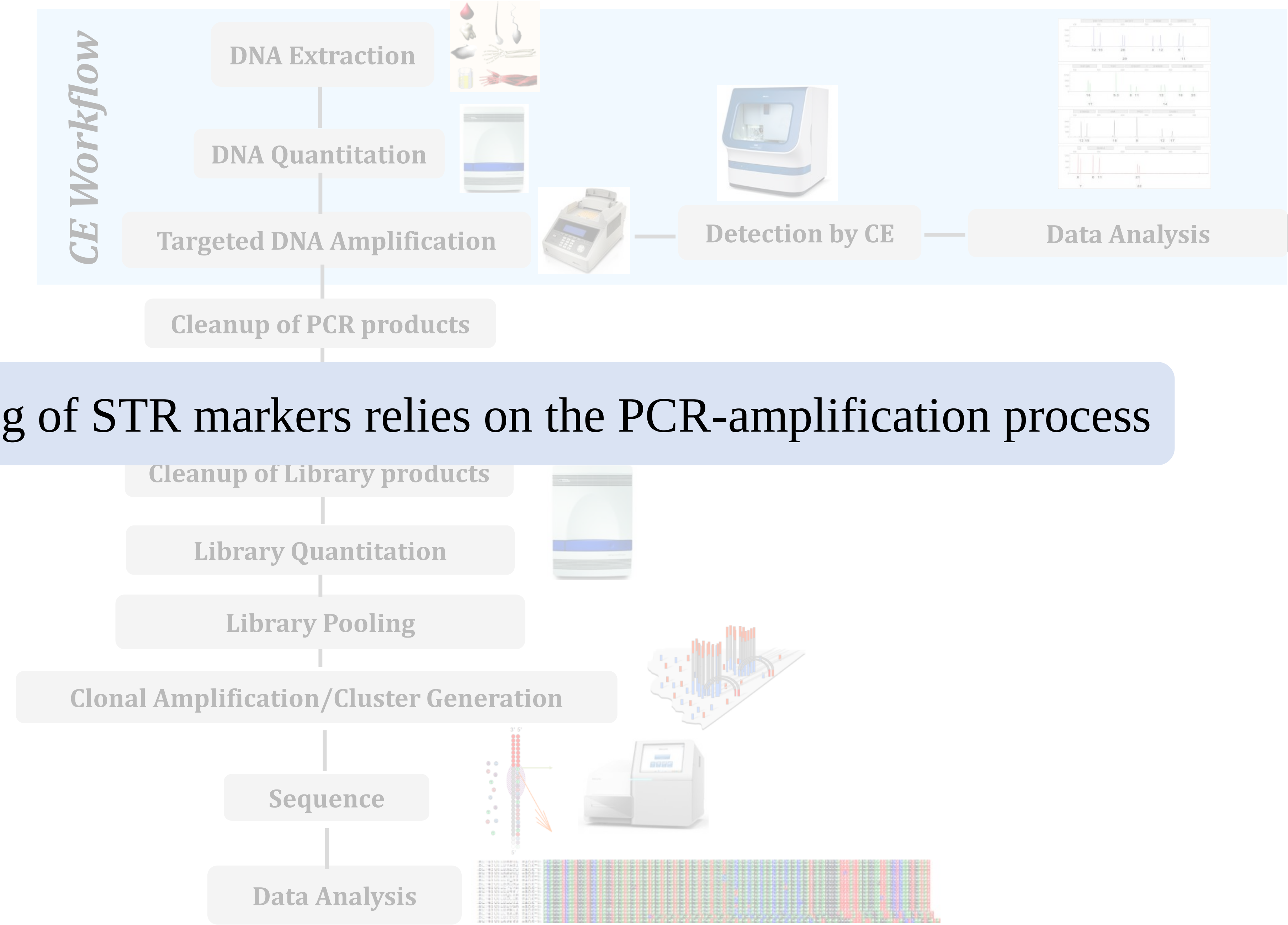


CE Workflow

- DNA Extraction
- DNA Quantitation
- Targeted DNA Amplification
- Detection by CE
- Data Analysis

- Cleanup of PCR products
- Library Construction
- Cleanup of Library products
- Library Quantitation
- Library Pooling
- Clonal Amplification/Cluster Generation
- Sequence
- Data Analysis

General Workflow for Next Generation Sequencing (NGS)



CE Workflow

DNA Extraction

DNA Quantitation

Targeted DNA Amplification

Detection by CE

Data Analysis

Cleanup of PCR products

Cleanup of Library products

Library Quantitation

Library Pooling

Clonal Amplification/Cluster Generation

Sequence

Data Analysis

Targeted sequencing of STR markers relies on the PCR-amplification process

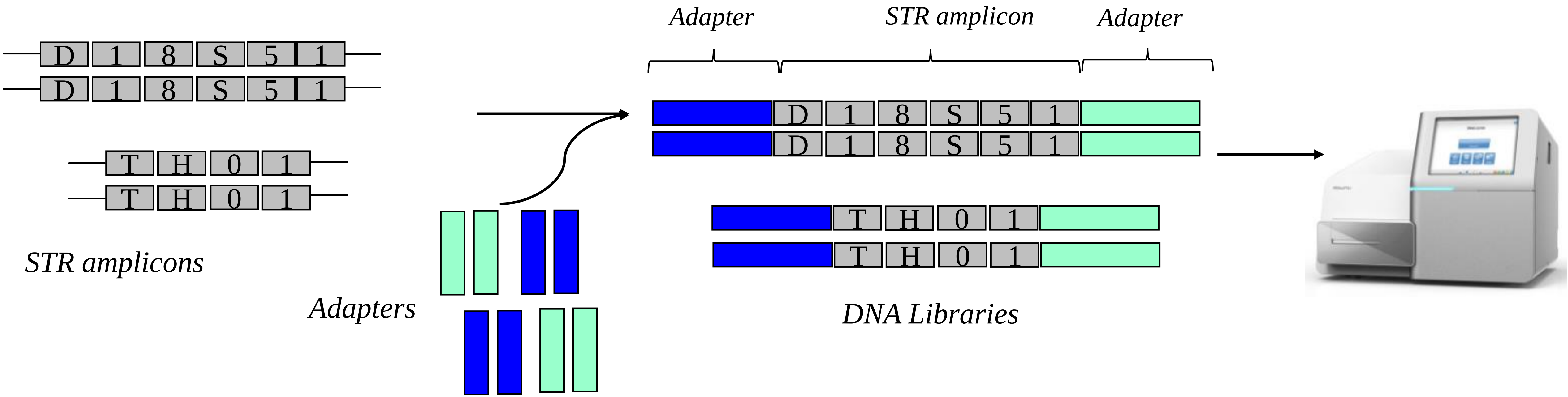
What is Library Construction ?

STR typing using CE

Dye-labeled STR amplicons

The aim of library preparation is to flank amplified STR products with adapters on both ends

Library preparation is essential for successful sequencing



STR amplicons

Adapters

Adapter

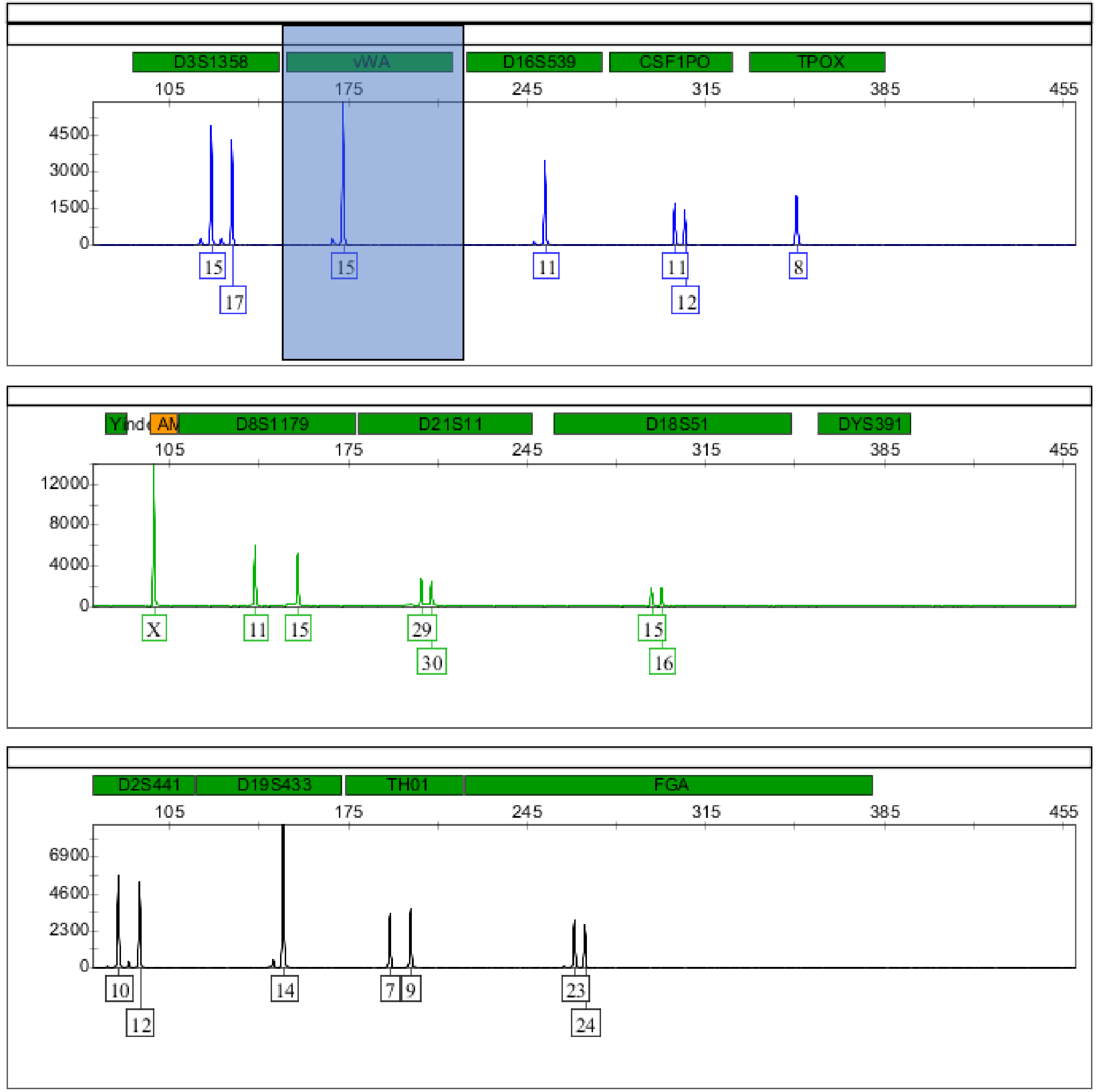
STR amplicon

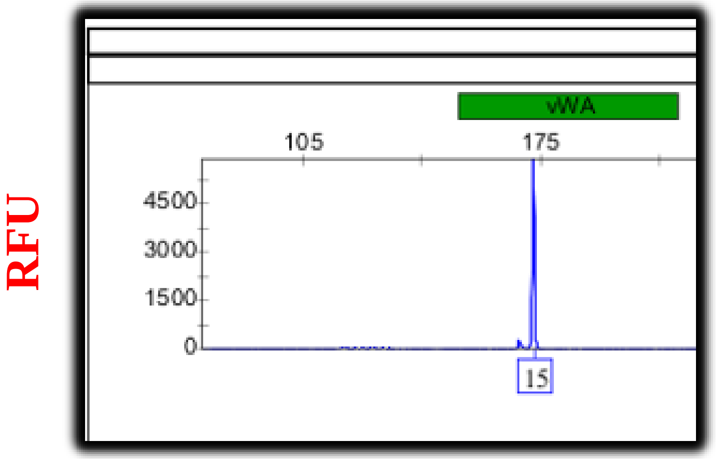
Adapter

DNA Libraries

Data Analysis using CE technology vs. NGS technology

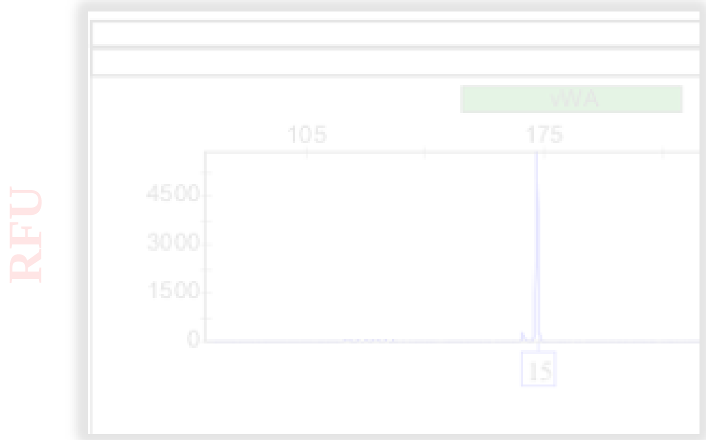
Data Analysis by CE





- ❖ Separation by size
- ❖ Length variation: 15
- ❖ Peak Height in RFU

Data Analysis by CE

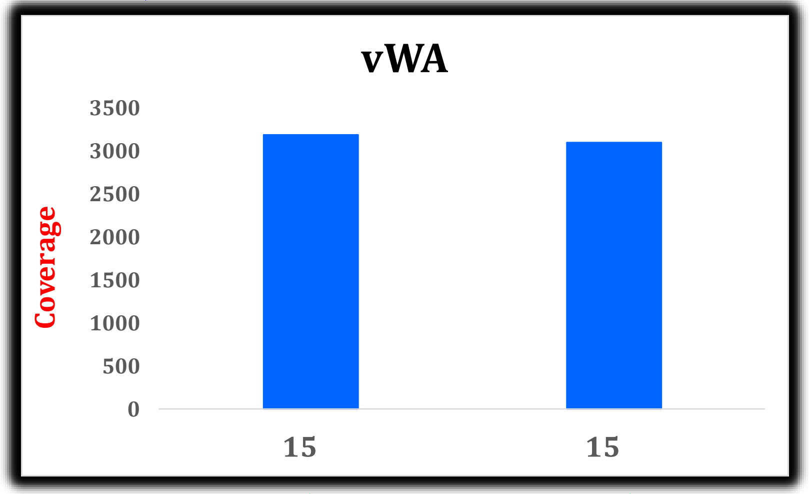


❖ Separation by size
❖ Length variation: 15
❖ Peak Height in RFU

Data Analysis by NGS

ATCCTGCAGATGCATCC
GTCTGTGCTGTTGCCTG
GTTATTGTAAGTCTCC
GATTCCCTTTTAGTTGC
TCTCATTTGCACTGGTT
CTGGGCCAACAAAAGCA
CAGCAGTTTTTCCCTCC
TTTCTTTATGGTGCTTG

Bioinformatics pipeline



[TAGA]11 [CAGA]3 TAGA

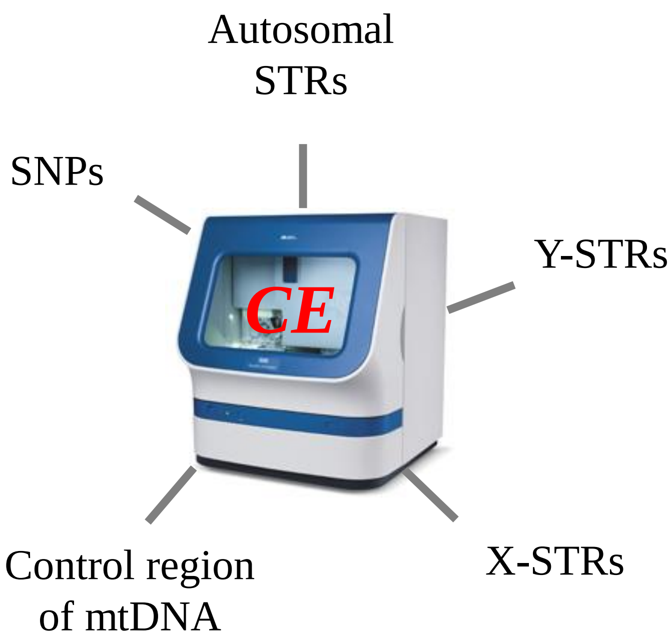
[TAGA]10 [CAGA]4 TAGA

❖ ~~Separation by size~~
✓ Length variation: 15
❖ ~~Peak Height in RFU~~ Depth of coverage

+ **Sequence variation**

Why implement NGS if you can accomplish DNA typing by CE?

Current Markers used in Forensic Genetics



Autosomal STRs

SNPs

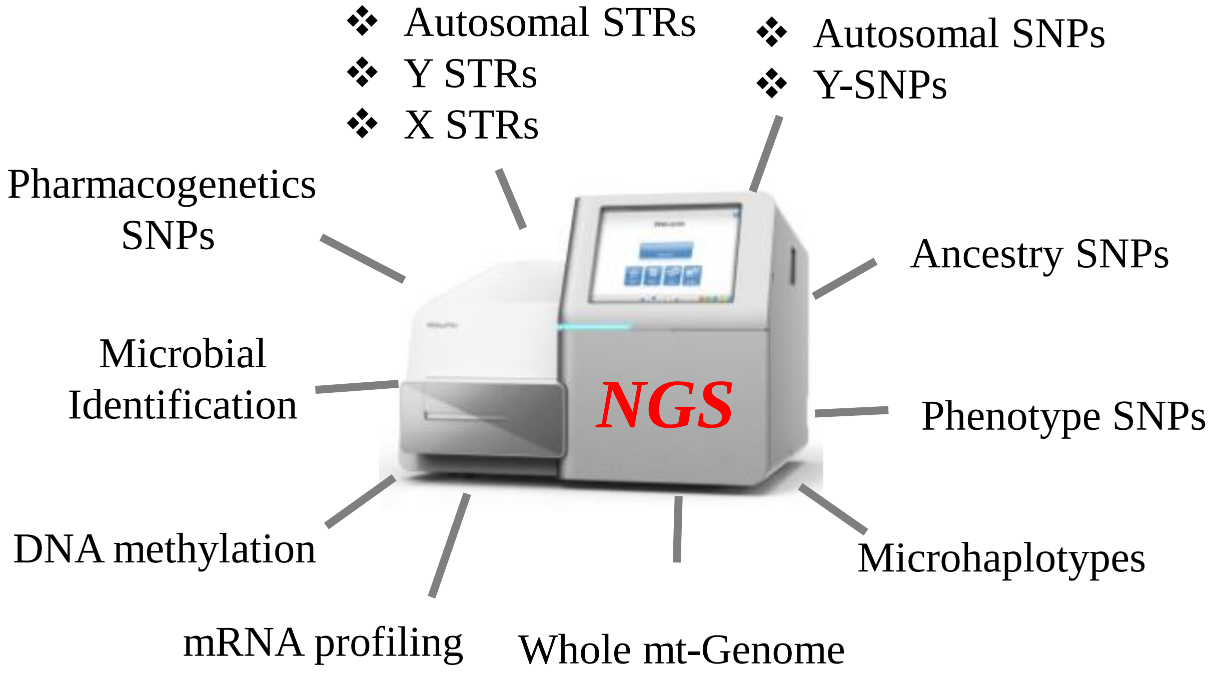
Y-STRs

Control region of mtDNA

X-STRs

- Examine one marker type at a time in one sample

NGS Sequencing Application and Markers



- Autosomal STRs
- Y STRs
- X STRs

- Autosomal SNPs
- Y-SNPs

Pharmacogenetics SNPs

Ancestry SNPs

Microbial Identification

Phenotype SNPs

DNA methylation

Microhaplotypes

mRNA profiling

Whole mt-Genome

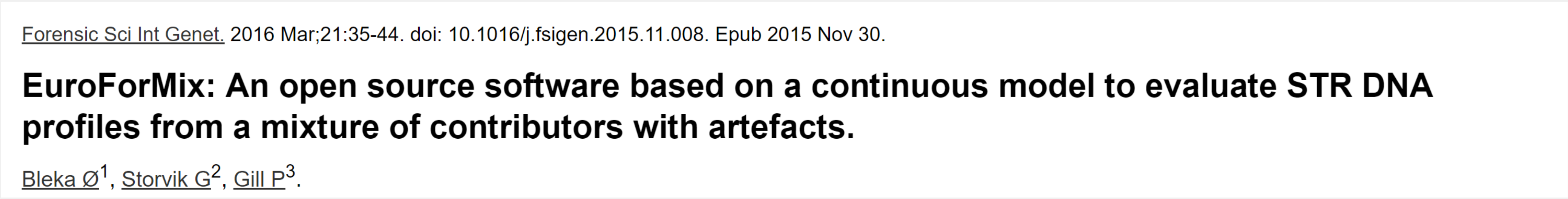
- Multiplex samples
- Multiplex markers
- **Distinguish between alleles identical by length but different in sequence content**

Forensic labs are moving from threshold based systems towards fully continuous and probabilistic DNA interpretation systems

Forensic Sci Int Genet. 2016 Jul;23:226-239. doi: 10.1016/j.fsigen.2016.05.007. Epub 2016 May 12.

Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles.

Bright JA[1], Taylor D[2], McGovern C[3], Cooper S[3], Russell L[3], Abarno D[4], Buckleton J[3].



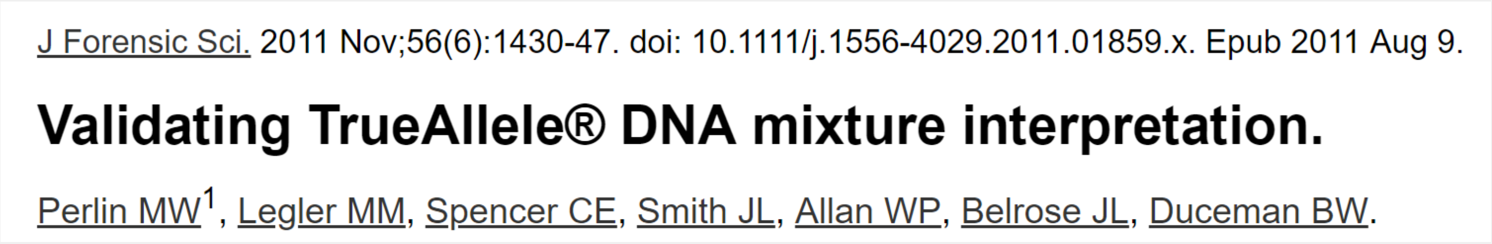
Forensic Sci Int Genet. 2016 Mar;21:35-44. doi: 10.1016/j.fsigen.2015.11.008. Epub 2015 Nov 30.

EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts.

Bleka Ø[1], Storvik G[2], Gill P[3].

What is DNA•VIEW®?

An integrated software package for DNA identification

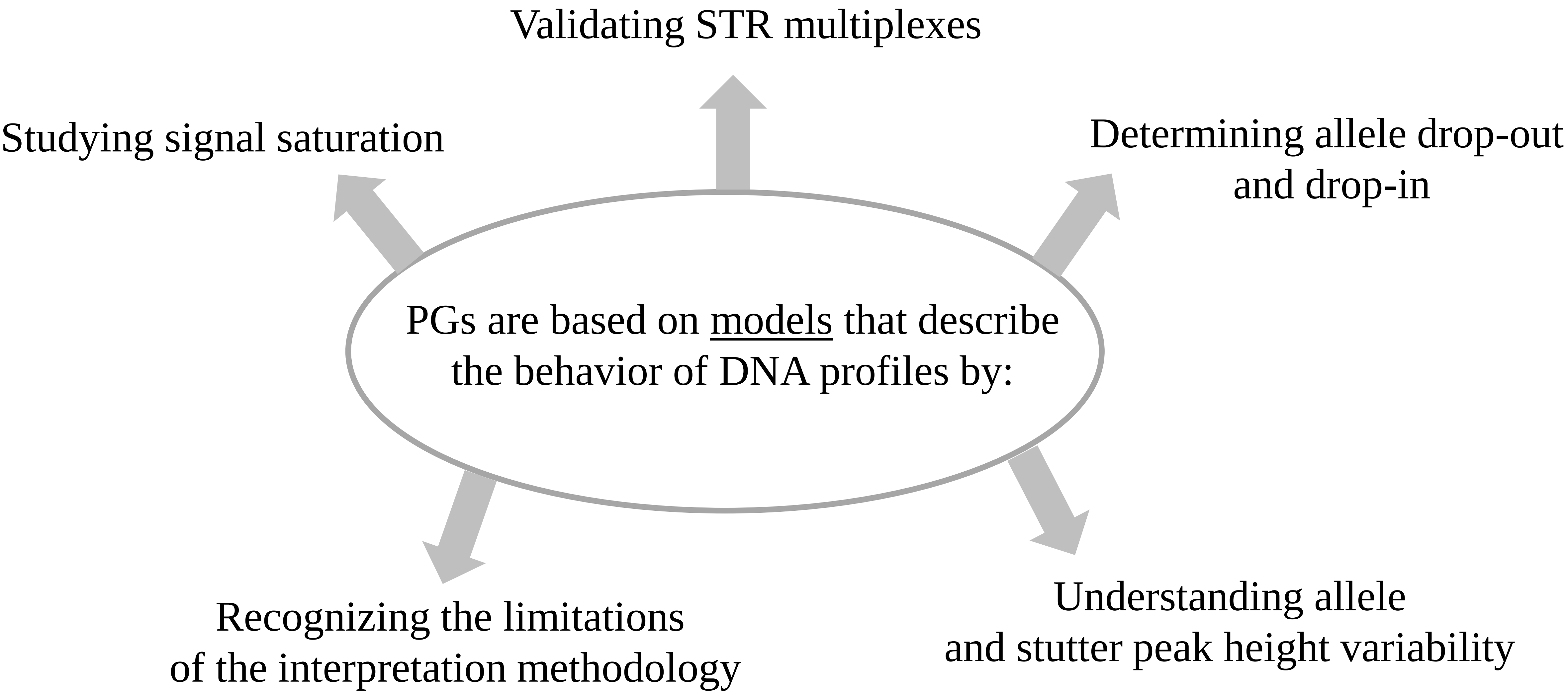


J Forensic Sci. 2011 Nov;56(6):1430-47. doi: 10.1111/j.1556-4029.2011.01859.x. Epub 2011 Aug 9.

Validating TrueAllele® DNA mixture interpretation.

Perlin MW[1], Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW.

Current considerations of the CE probabilistic genotyping (PG) systems



PGs are based on models that describe the behavior of DNA profiles by:

Validating STR multiplexes

Studying signal saturation

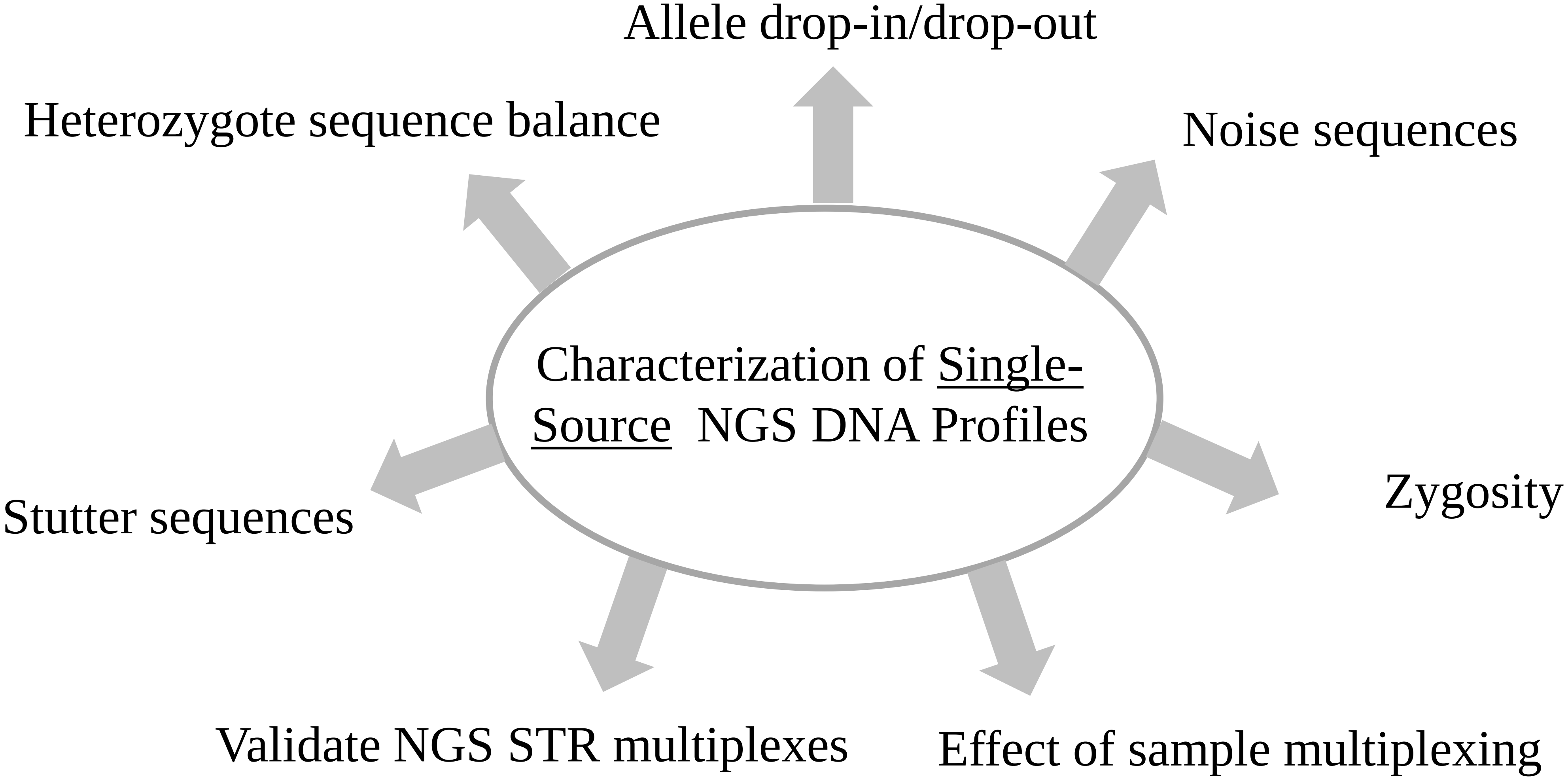
Determining allele drop-out and drop-in

Recognizing the limitations of the interpretation methodology

Understanding allele and stutter peak height variability

What do we need to understand to establish STR NGS interpretation systems?

We need to understand and analyze the STR NGS sequence data



Characterization of Single-Source NGS DNA Profiles

Allele drop-in/drop-out

Noise sequences

Zygosity

Effect of sample multiplexing

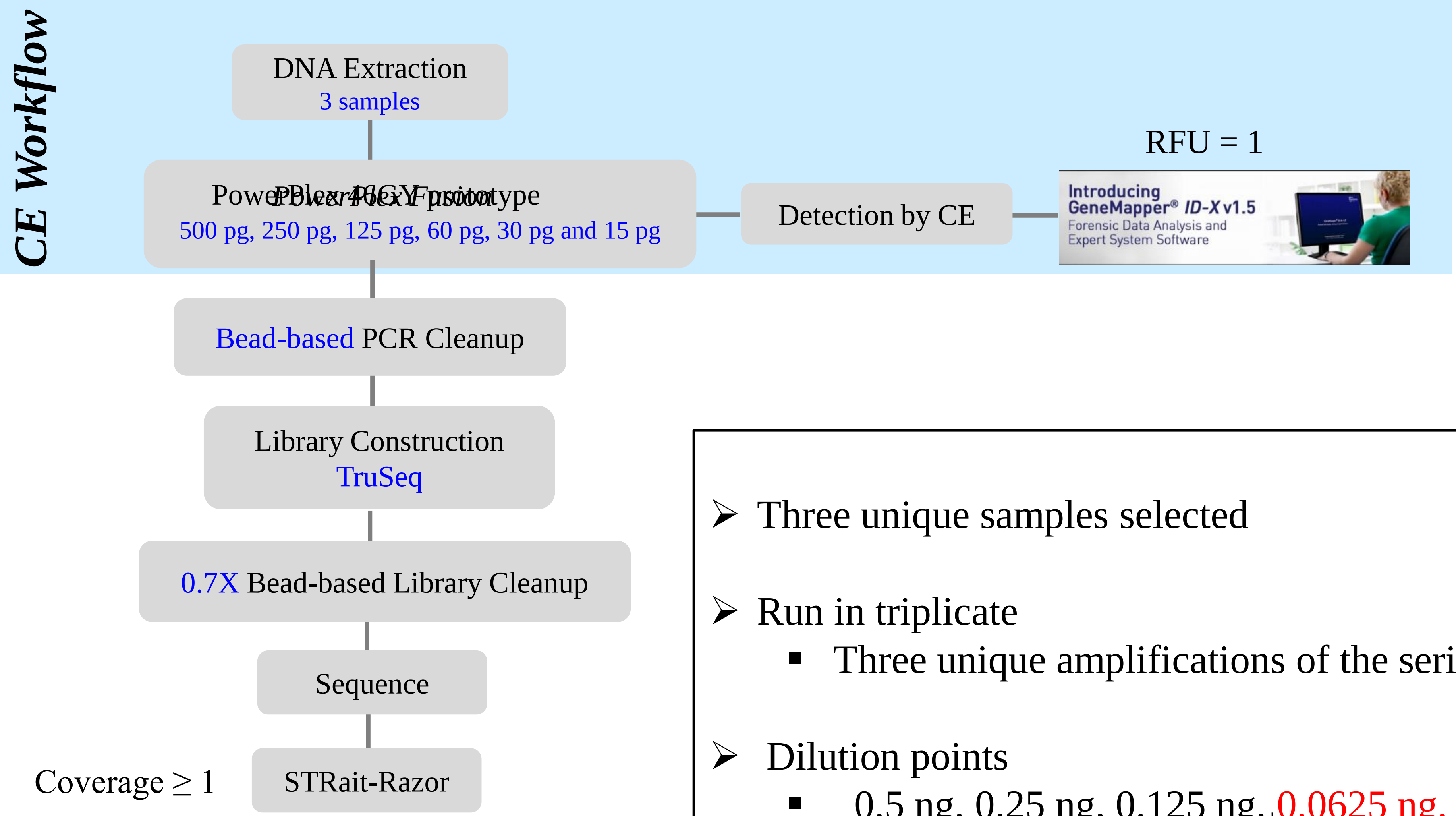
Validate NGS STR multiplexes

Stutter sequences

Heterozygote sequence balance

CE and NGS sensitivity experimental design

CE and NGS sensitivity experimental design



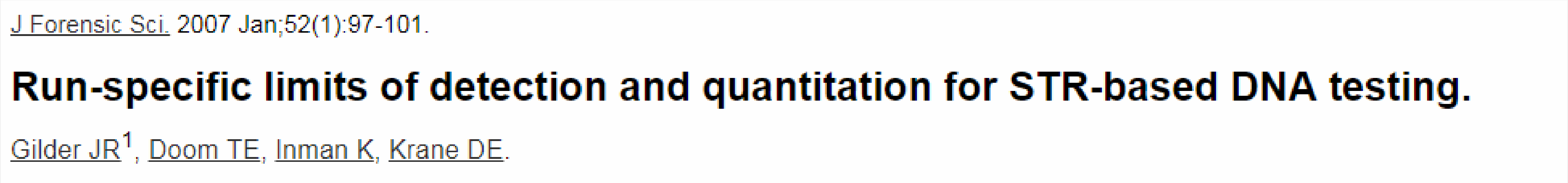
- Three unique samples selected
- Run in triplicate
  - Three unique amplifications of the serial dilutions
- Dilution points
  - 0.5 ng, 0.25 ng, 0.125 ng, $\underbrace{\text{0.0625 ng, 0.03 ng, and 0.015 ng}}_{\text{Stochastic effects}}$

Noise Thresholds for CE Data

Forensic Sci Int Genet. 2012 Dec;6(6):723-8. doi: 10.1016/j.fsigen.2012.06.012. Epub 2012 Jul 12.

Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out.

Rakay CA[1], Bregu J, Grgicak CM.



J Forensic Sci. 2007 Jan;52(1):97-101.

Run-specific limits of detection and quantitation for STR-based DNA testing.

Gilder JR[1], Doom TE, Inman K, Krane DE.

J Forensic Sci. 2013 Jan;58(1):120-9. doi: 10.1111/1556-4029.12008. Epub 2012 Nov 6.

Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis.

Bregu J[1], Conklin D, Coronado E, Terrill M, Cotton RW, Grgicak CM.

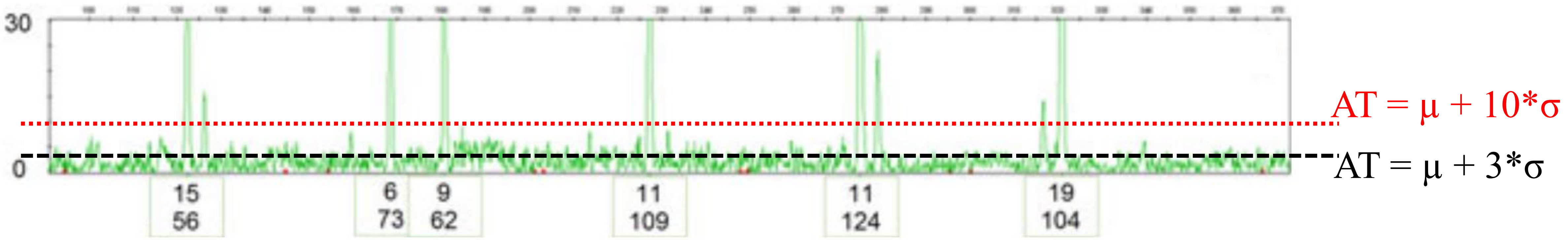
Analytical Threshold Most Commonly Determined by:

$$AT_{M1} = \overline{Y}_{bl} + k s_{bl}$$

- *k = Numerical factor (e.g. k=3)*

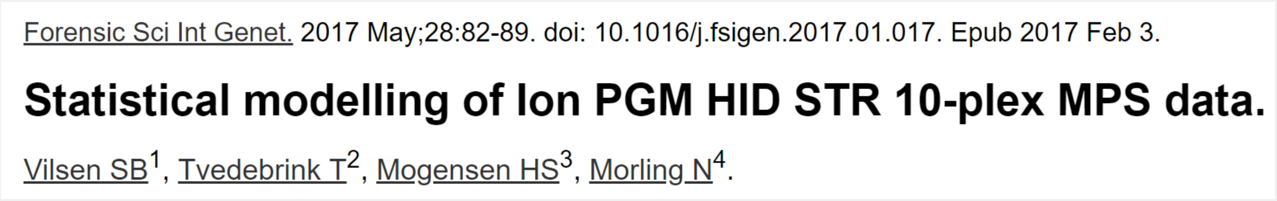
Average RFU signal

STDEV of the signal



J. Bregu, D. Conklin, E. Coronado, M. Terrill, R.W. Cotton, C.M. Grgicak, Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis, Journal of forensic sciences 58(1) (2013) 120-9.

Noise Thresholds for NGS Data

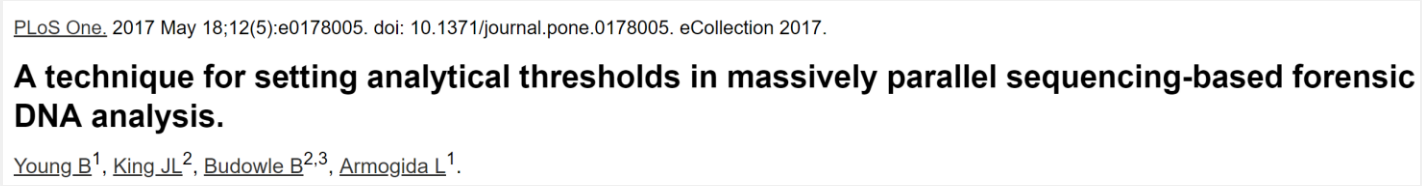


Forensic Sci Int Genet. 2017 May;28:82-89. doi: 10.1016/j.fsigen.2017.01.017. Epub 2017 Feb 3.

Statistical modelling of Ion PGM HID STR 10-plex MPS data.

Vilsen SB[1], Tvedebrink T[2], Mogensen HS[3], Morling N[4].

Removal of general noise using thresholds created by fitting the distribution of general noise sequences.

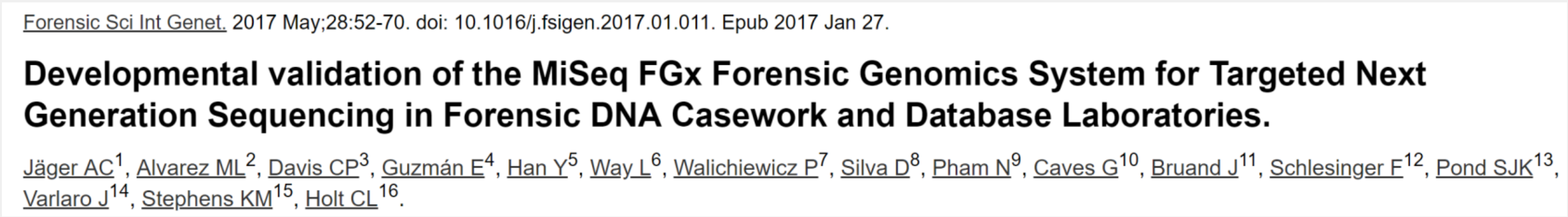


PLoS One. 2017 May 18;12(5):e0178005. doi: 10.1371/journal.pone.0178005. eCollection 2017.

A technique for setting analytical thresholds in massively parallel sequencing-based forensic DNA analysis.

Young B[1], King JL[2], Budowle B[2,3], Armogida L[1].

$AT = c * (Max_{noise} - Min_{noise})$

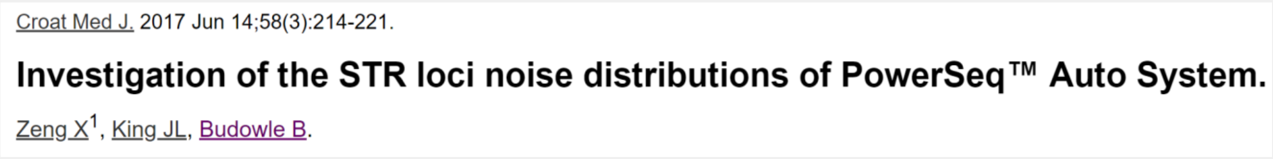


Forensic Sci Int Genet. 2017 May;28:52-70. doi: 10.1016/j.fsigen.2017.01.011. Epub 2017 Jan 27.

Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories.

Jäger AC[1], Alvarez ML[2], Davis CP[3], Guzmán E[4], Han Y[5], Way L[6], Walichiewicz P[7], Silva D[8], Pham N[9], Caves G[10], Bruand J[11], Schlesinger F[12], Pond SJK[13], Varlaro J[14], Stephens KM[15], Holt CL[16].

AT level is set at 1.5% of total locus coverage



Croat Med J. 2017 Jun 14;58(3):214-221.

Investigation of the STR loci noise distributions of PowerSeq™ Auto System.

Zeng X[1], King JL, Budowle B.

Characterization of sequences in STR profiles generated on MiSeq platform using the PowerSeq 46GY prototype kit

We grouped the generated sequences intro three categories:

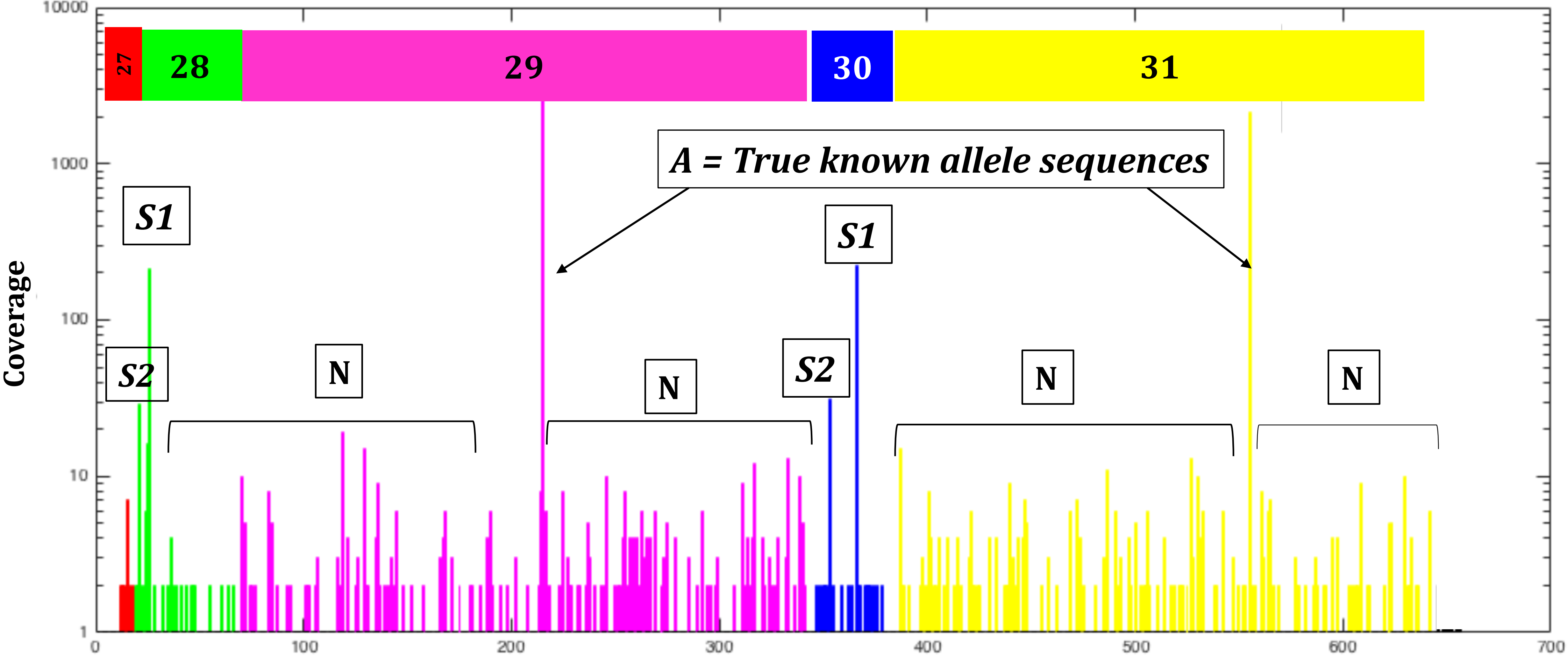
S1 = Back stutter of the longest uninterrupted stretch of the basic repeat motifs within an allelic sequence

S2 = Back stutter sequences not attributed to S1

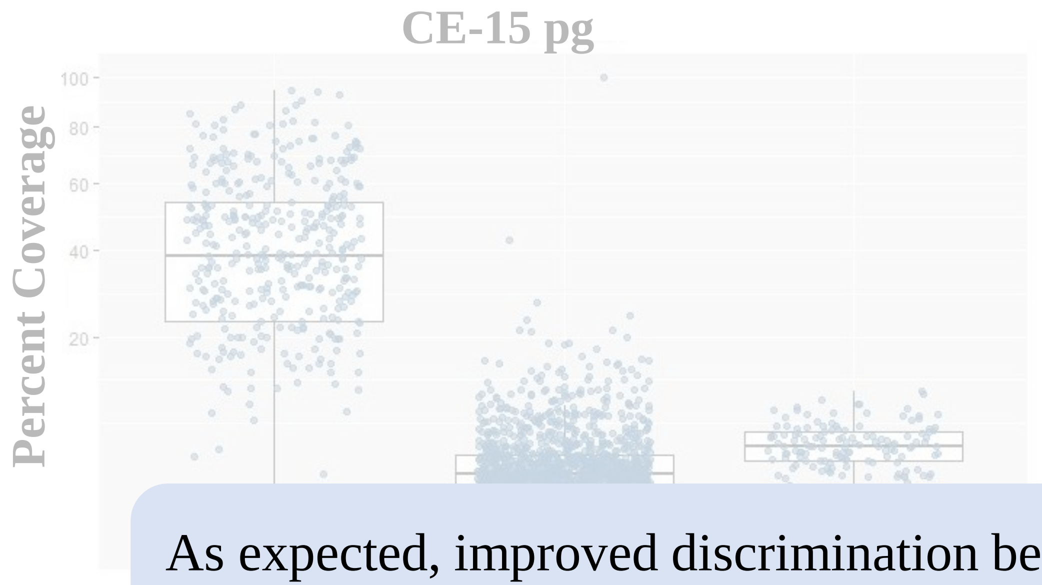
N = Noise sequences

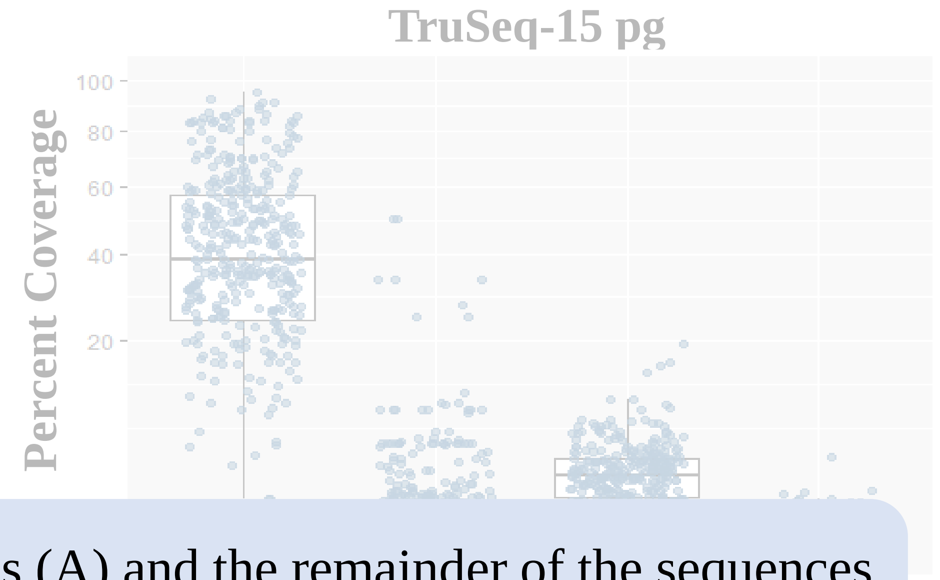
Total Type of Sequences = 646
Locus Coverage = 6949

D21S11 (29, 31)

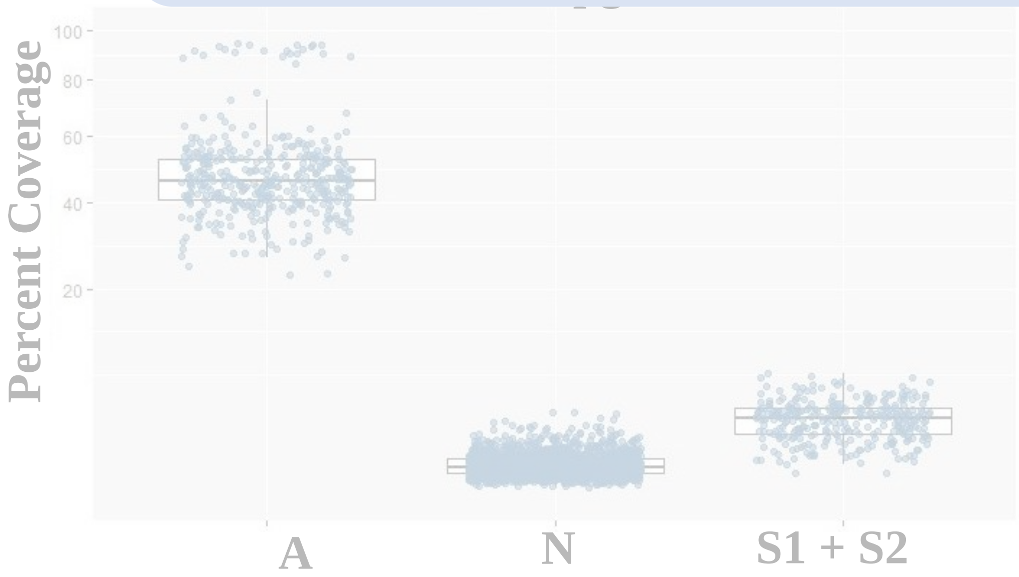


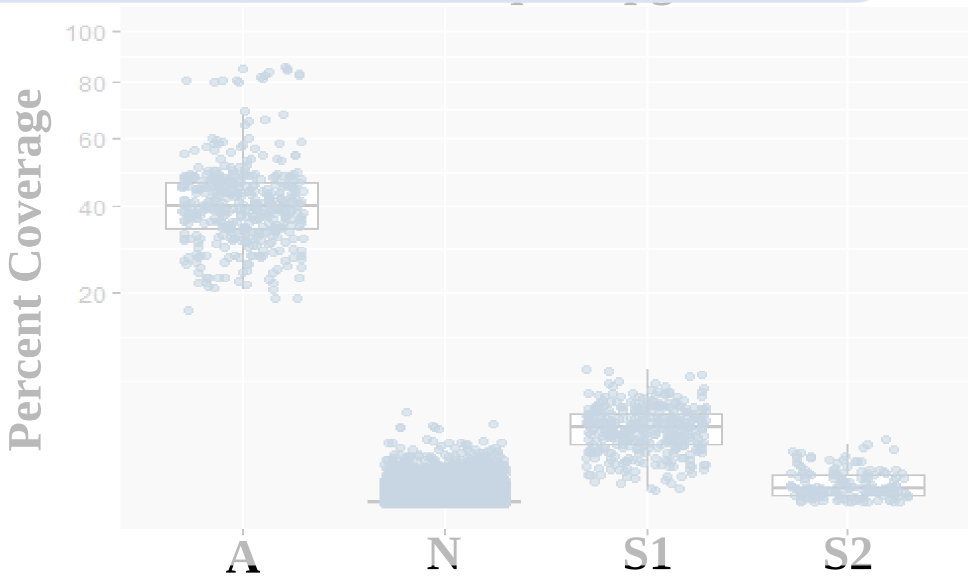
Distribution of Known Allele, Stutter, and Noise Sequences



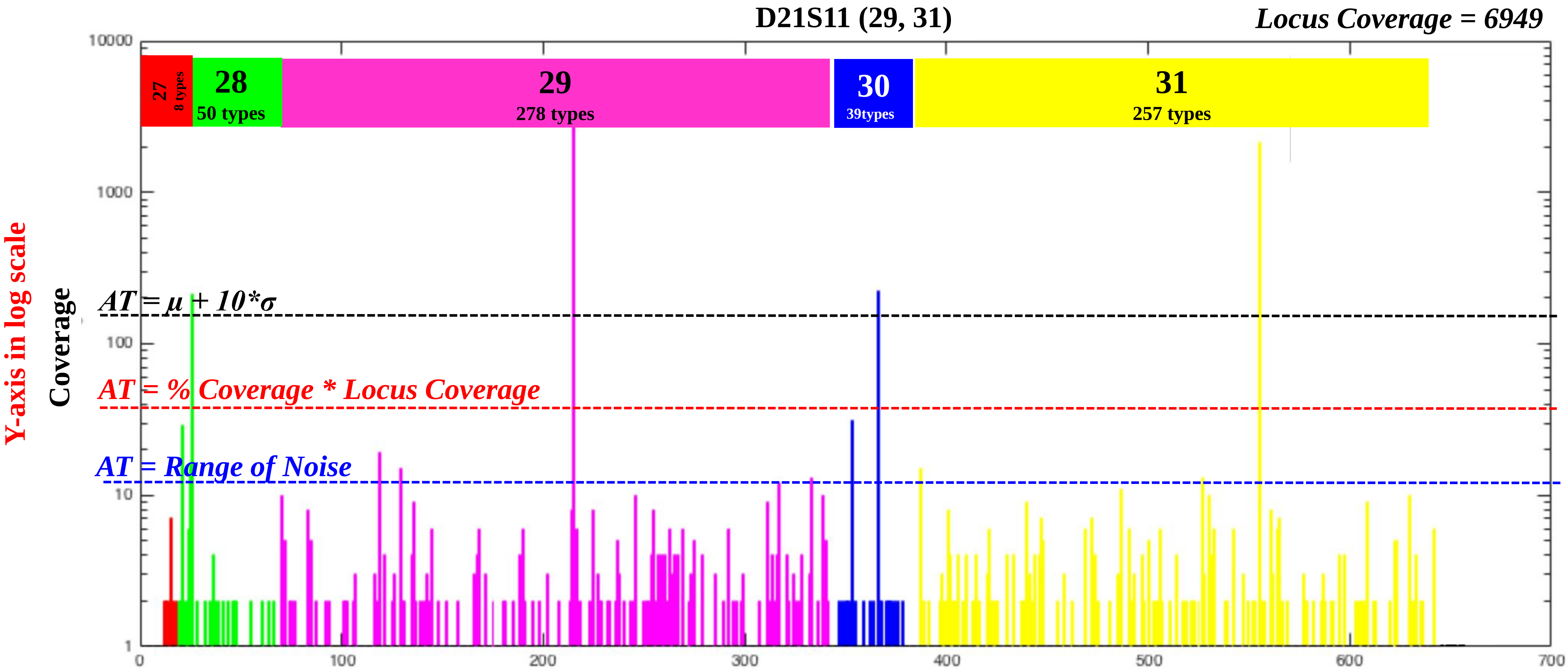


As expected, improved discrimination between known alleles (A) and the remainder of the sequences (N, S1, and S2) is observed as the amount of DNA template increases.

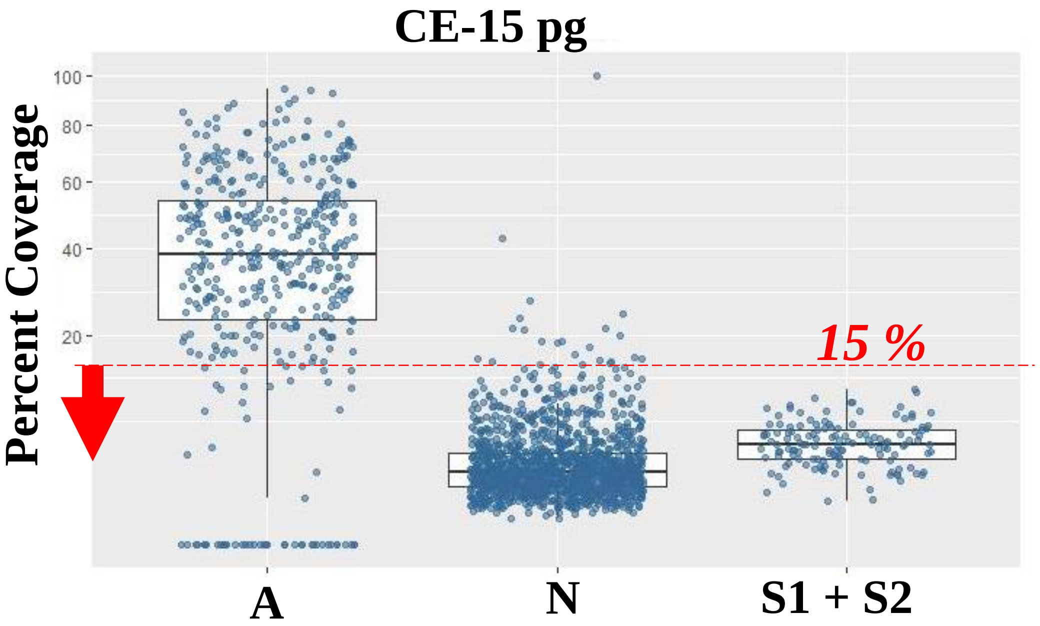




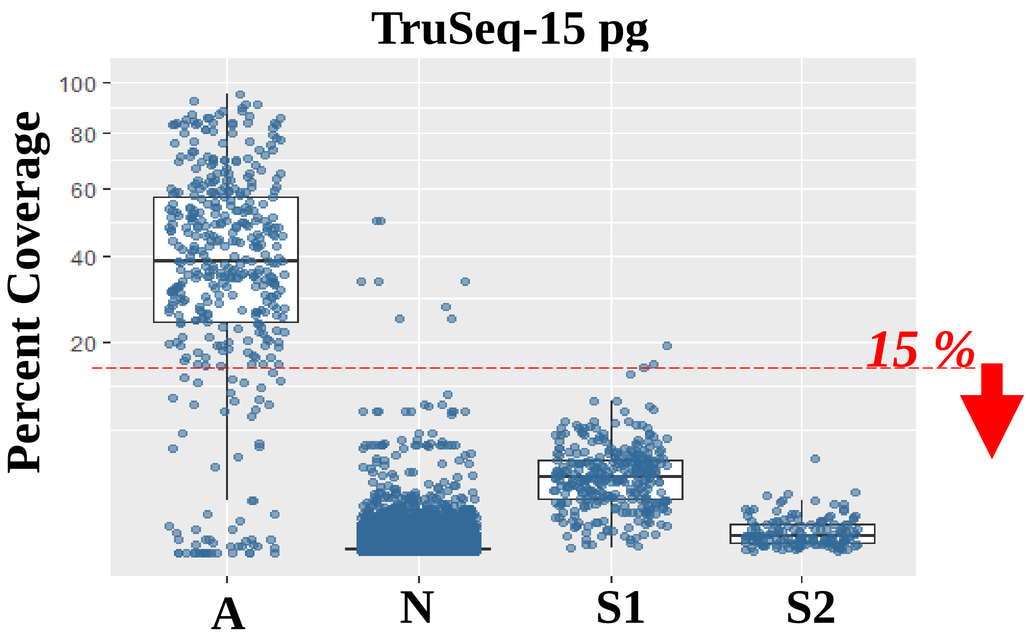
Observed Sequences and their coverage at a heterozygote D21S11 Locus



Evaluating the tradeoff between the allelic (true positives), stutter, and noise sequences (false positives)



CE-15 pg



TruSeq-15 pg

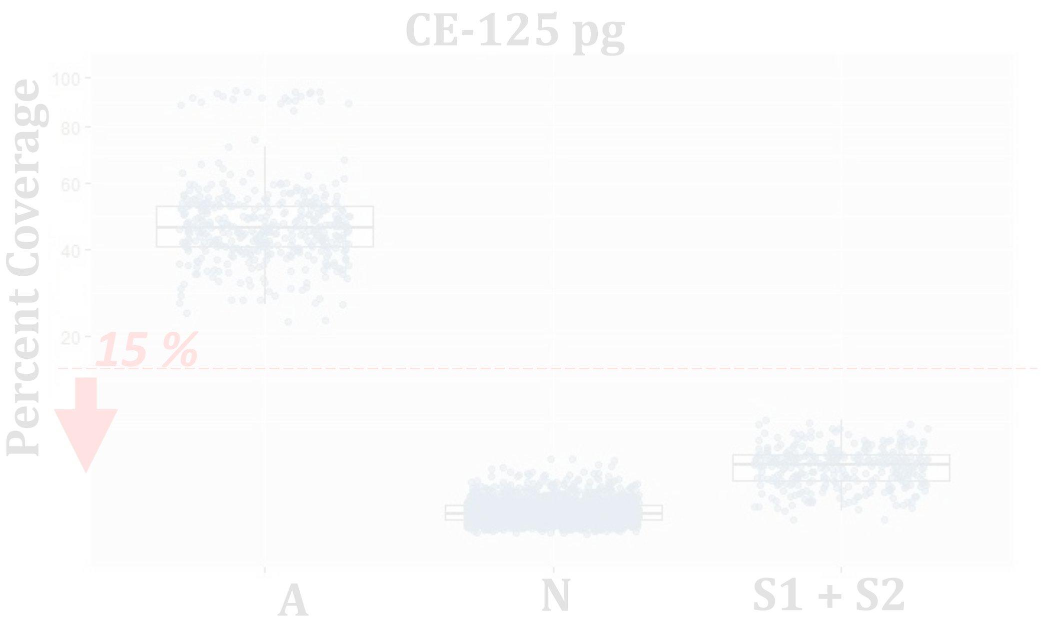
At a percent coverage of 15%:

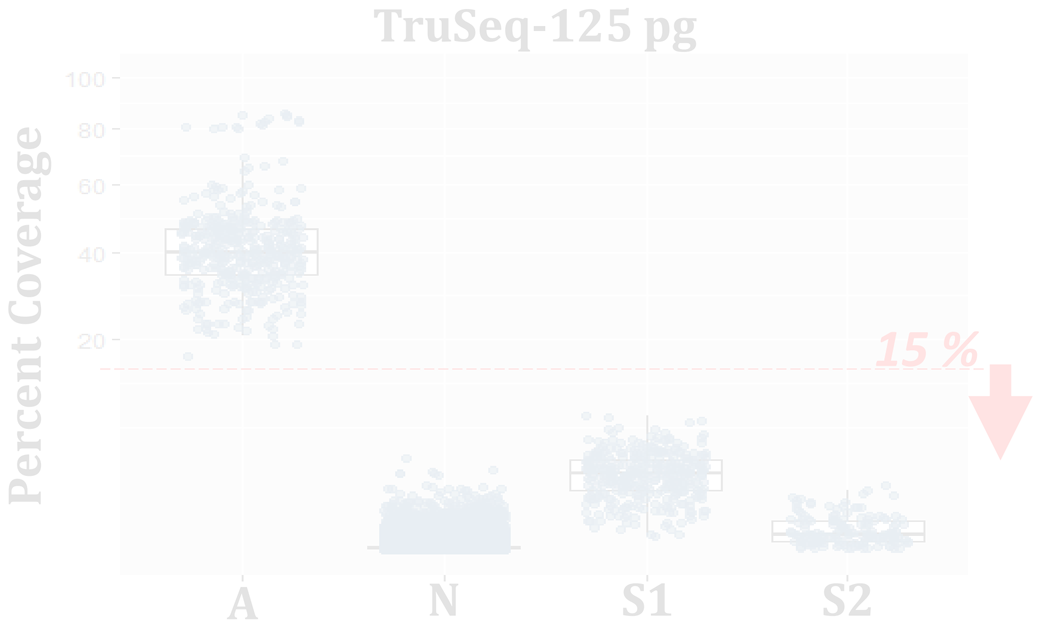
- ***363 peaks are called***
  - ***352** can be attributed to **A***
  - ***21** can be attributed to **N***
  - ***0** can be attributed to **S***

At a percent coverage of 15%:

- ***361 sequences are called***
  - ***350** can be attributed to **A***
  - ***8** can be attributed to **N***
  - ***3** can be attributed to **S1***
  - ***0** can be attributed to **S2***

Evaluating the tradeoff between the allelic (true positives), stutter, and noise sequences (false positives)





A value of 15 % is <u>ONLY</u> used for illustrative purposes and not as a recommended threshold. Each lab should perform sensitivity experiments and establish a threshold for interpretational purposes.

- ***0** can be attributed to **S***

- ***0** can be attributed to **S1***
- ***0** can be attributed to **S2***

Summary

- Understanding the behavior of STR NGS profiles can help in statistical modeling and probability distributions needed for establishing an STR NGS interpretation system.

- Future work will focus on analyzing more single source and mixture samples.

Presentation will be available for download from STRBase:
http://strbase.nist.gov/NISTpub.htm#Presentations

https://strbase.nist.gov/pub_pres/Sarah-ISHI2018-Poster-SR-final_pmv.pdf

Acknowledgement

NIST

Pete Vallone

Hari Iyer (Statistical Design, Analysis, and Modeling Group)

Lisa Borsuk

Becky Steffen

Erica Romsos

Katherine Gettings

Kevin Kiesler

Margaret Kline

Megan Cleveland

Promega
Doug Storts
Spencer Hermanson

Funding
NIST Special Programs Office: *Forensic DNA*
FBI Biometrics Center of Excellence: *Forensic DNA Typing as a Biometric tool.*

All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Contact: sarah.riman@nist.gov