American Academy of Forensic Sciences

HYBRID WORKSHOP **W2**

February 21, 2022

The National Institute of Standards and Technology (NIST)

Forensic DNA Activities: Foundations, Research, and Standards

Chair

John M. Butler, PhD

Special Programs Office

Co-Chair

John Paul Jones, MBA

Special Programs Office

Presenter

Melissa K. Taylor, MA

Special Programs Office

Co-Chair

Peter M. Vallone, PhD

Applied Genetics Group

Presenter

Katherine B. Gettings, PhD

Applied Genetics Group

Presenter

Sarah Riman, PhD

Applied Genetics Group

Presenter

Carolyn R. Steffen, MS

Applied Genetics Group

NIST FORENSIC SCIENCE

RESEARCH. STANDARDS. FOUNDATIONS.

NIST APPLIED GENETICS

NIST DNA Workshop Schedule

Module	Time (Pacific)	Topic *Supporting Articles Supplied*	Presenter
1	8:30am (15 minutes)	Introduction to Workshop and NIST Forensic Science Activities	JP Jones
2	8:45am (75 minutes)	Scientific Foundation Study on DNA Mixture Interpretation	John Butler
3	10:00am (30 minutes)	Examining Probabilistic Genotyping Systems	Sarah Riman
	10:30am	**BREAK (15 minutes)**	
4	10:45am (30 minutes)	DNA Mixture Standards on the OSAC Registry	JP Jones
5	11:15am (45 minutes)	DNA Process Map & Human Factors Working Group on DNA Interpretation	Melissa Taylor
	12:00pm to 1:00pm 60-minute LUNCH BREAK		
6	1:00pm (30 minutes)	DNA Sequencing Research Overview	Pete Vallone
7	1:30pm (45 minutes)	STR Sequence Nomenclature Activities	Katherine Gettings
8	2:15pm (30 minutes)	NIST DNA Standard Reference Materials	Becky Steffen
	2:45pm	**BREAK (15 minutes)**	
9	3:00pm (30 minutes)	DNA Training Standards on the OSAC Registry and Educational Materials	JP Jones
10	3:30pm (15 minutes)	STRBase Updates	Pete Vallone
11	3:45pm (30 minutes)	DNA Most Valuable Publications List	John Butler
12	4:15pm (30 minutes)	**PANEL: Questions and Answers**	All Presenters
13	4:45pm (15 minutes)	Wrap-up and Workshop Conclusions	John Butler

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

Introduction to Workshop and NIST Forensic Science Activities

National Institute of Standards and Technology (NIST)

John Paul Jones
Special Programs Office

Module 1

1

Acknowledgments and Disclaimer

Points of view are the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities 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 National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

2

Workshop Overview

- **Title:** The National Institute of Standards and Technology (NIST) Forensic DNA Activities: Foundations, Research, and Standards
- **Learning Overview:** Presenters will review activities at NIST involving forensic DNA foundational studies, research, and standards.
- **Impact Statement:** Presentations in this workshop will impact the forensic science community by contributing to an understanding of NIST activities in advancing knowledge and practice of forensic DNA through foundation studies, focused research, and development of documentary standards.
- **Program Description:** Three sessions will focus on DNA mixture interpretation, DNA sequencing research, and DNA training materials that will benefit students, practitioners, and stakeholders. Participants will gain an understanding of principles involved in DNA analysis and interpretation, knowledge of core foundational literature supporting these principles, and information that can strengthen training programs for DNA analysts.

3

Planned Workshop Schedule (morning)

Time (Pacific)	Topic	Presenter(s)
8:30am (15 minutes)	Introduction to Workshop and NIST Forensic Science Activities	JP Jones
8:45am (75 minutes)	Scientific Foundation Study on DNA Mixture Interpretation	John Butler
10:00am (30 minutes)	Examining Probabilistic Genotyping Systems	Sarah Riman
10:30am	BREAK (15 minutes)	
10:45am (30 minutes)	DNA Mixture Standards on the OSAC Registry	JP Jones
11:15am (45 minutes)	DNA Process Map and Human Factors WG	Melissa Taylor

12:00pm to 1:00pm 60-minute LUNCH BREAK

4

Planned Workshop Schedule (afternoon)

Time (Central)	Topic	Presenter(s)
1:00pm (30 minutes)	DNA Sequencing Research Overview	Pete Vallone
1:30pm (45 minutes)	STR Sequence Nomenclature Activities	Katherine Gettings
2:15pm (30 minutes)	NIST DNA Standard Reference Materials	Becky Steffen
2:45pm (15 minutes)	BREAK (15 minutes)	
3:00pm (30 minutes)	DNA Training Standards on the OSAC Registry and Educational Materials	JP Jones
3:30pm (15 minutes)	STRBase Updates	Pete Vallone
3:45pm (30 minutes)	DNA Most Valuable Publications List	John Butler
4:15pm (30 minutes)	**Question and Answers** (live Zoom)	All Presenters
4:45pm (15 minutes)	Wrap-up and Workshop Conclusions	John Butler

5

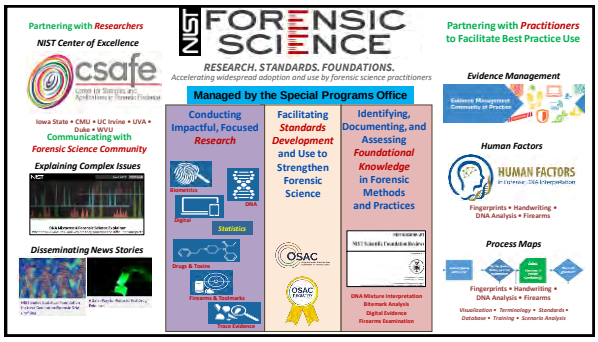
National Institute of Standards and Technology (NIST)

Unique Mission within the Federal Government ...
*to promote U.S. innovation and industrial competitiveness by advancing **measurement science, standards, and technology** in ways that enhance economic security and improve our quality of life.*

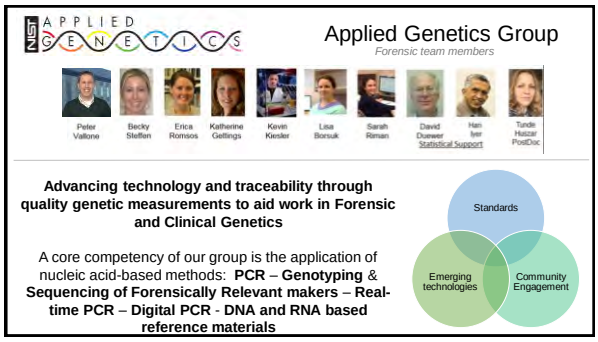
- **Deep research expertise underpins technological innovation** – e.g., new materials, advanced clinical diagnostics and therapies, advanced communications, **forensic science** etc.
- **Non-regulatory status** enables an important role as a **convener** that **facilitates collaboration** between agencies of the Federal Government, industry, private organizations, and state and local governments

Gaithersburg, MD Campus

6



7



8

Background and Qualification of Presenters

- **John M. Butler, PhD:** NIST Fellow in the Special Programs Office. Author of five textbooks on DNA (2001, 2005, 2010, 2012, 2015) and >180 research articles and has conducted dozens of workshops on forensic DNA.
- **Katherine Gettings, PhD:** Research biologist in the Applied Genetics Group at NIST, where she focuses on forensic applications of next generation sequencing technologies. Today she'll be sharing updates on STR sequence nomenclature.
- **John Paul Jones II, MBA:** Forensic Science Standards Program Manager in the Special Programs Office, where he manages the Organization of Scientific Area Committees (OSAC) for Forensic Science. He is active in forensic science standards development and implementation.
- **Sarah Riman, PhD:** Research Associate in the Applied Genetics Group. Riman's work is focused on understanding the factors that affect the measurement and interpretation of STR profiles. Today she will be discussing her recent study on examining performance and LR values of different LR systems.

9

Background and Qualification of Presenters

- **Becky Steffen, MS:** Research biologist for the Applied Genetics Group at NIST, where she focuses on Standard Reference Material development, capillary electrophoresis testing, and next generation sequencing. Today she'll be sharing recent and ongoing updates to SRM 2391d: PCR-Based DNA Profiling Standard.
- **Melissa K. Taylor, MA:** Senior Forensic Science Research Manager for the Forensic Science Program within the Special Programs Office. Her work focuses primarily on impression and pattern evidence-related research, process mapping, and integrating human-factors principles into forensic sciences. Today she will be discussing the progress of the NIST/NIJ Expert Working Group on Human Factors in DNA Interpretation and presenting the NIST-led DNA interpretation process map.
- **Peter Vallone, PhD:** Leader of the Applied Genetics Group at NIST, where he focuses on standards and methods to support forensic and clinical genetics. Today he'll be giving an overview of sequencing projects in his group as well as an update on the STRBase website

10

Thank you for your attention!

John M. Butler
john.butler@nist.gov

Peter M. Vallone
peter.vallone@nist.gov

John Paul Jones
john.jones@nist.gov

Katherine B. Gettings
katherine.gettings@nist.gov

Melissa Taylor
melissa.taylor@nist.gov

Carolyn R. Steffen
becky.steffen@nist.gov

Sarah Riman
sarah.riman@nist.gov

11

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

Scientific Foundation Study on DNA Mixture Interpretation

National Institute of Standards and Technology (NIST)

John M. Butler
Special Programs Office

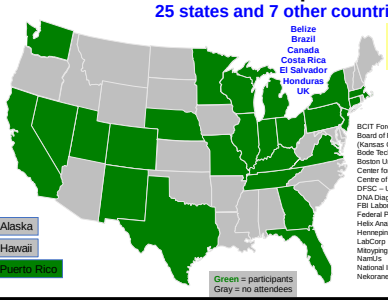
Module 2

1

NEW SLIDE

AAFS 2022 Workshop W2 Registrants

25 states and 7 other countries



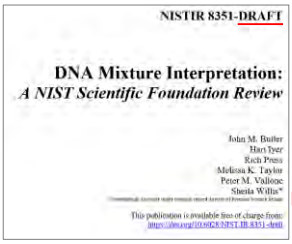
127 registered + 7 presenters
53 in-person & 74 virtual
(as of 2/15/2021)

In addition to state and local forensic laboratory analysts, we have representatives from:

BCIT Forensics
Board of Police Commissioners (Kansas City, MO)
Bode Technology
Boston University
Center for Forensic Sci. Res. & Ed.
Centre of Forensic Sciences
DFSC – USACIL
DNA Diagnostics Center
FBI Laboratory
Federal Public Defender's Office
Helix Analytical
Hennepin County Public Defender
LabCorp
Miloying Technologies
NamUs
National Institute of Justice
Nekoranec Psychology
Ohio Northern University
Pacific Architects and Engineers
Penn State University
Promega
RCMP
RTI International
Sixth District Public Defender
St. Mary's University
Superior Court (CA)
Texas Forensic Science Commission
Thermo Fisher Scientific
University of Illinois-Chicago
University of Kentucky
University of Nebraska
University of Nevada-Reno
University of New Haven
Verogen

2

NIST Draft Report Released in June 2021


NISTIR 8351-DRAFT

DNA Mixture Interpretation:
A NIST Scientific Foundation Review

John M. Butler
Hari Iyer
Rich Press
Melissa K. Taylor
Peter M. Vallone
Sheila Willis*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8351-draft

250 pages Executive Summary (9 pages)
6 chapters and 2 appendices
528 references cited
47 terms and acronyms defined
29 tables
12 figures
5 boxes
16 principles described
25 key takeaways
8 future considerations

Collected public comments on this draft report (June to November 2021)

3

Presentation Overview

1. Report Contents and Key Takeaways
   - Why NIST has undertaken this effort
   - Brief summary of our findings and principles for DNA mixture interpretation
2. Outreach and Public Comments Received
   - Public webinar given on July 21, 2021 (1,000 registrants) – 83 questions/comments
   - Presentations given to FBI SWGDAM (July 14) and NIST/NIJ Human Factors WG (July 28)
   - **63 sets of public comments received** (shared in December 2021)
3. Some Associated Topics
   - Data availability
   - Validation factor space

These handouts, which were due to AAFS by mid-January, do not contain the final slides; for a final version of the presentation, see https://strbase.nist.gov/NISTpub.htm after February 21

4

Disclaimer & Acknowledgments

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 endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Acknowledgments (*page i*): Members of the DNA Mixture Resource Group (listed in Table 1.2) contributed helpful feedback and assistance in the early stages of drafting this report. **Katherine Gettings**, **Nikola Osborne**, and **Sarah Riman** provided valuable input on the text, including the data summaries used in Chapter 4. **Jason Weixelbaum, Susan Ballou, Christina Reed, and Kathy Sharpless** assisted with copy editing. **Kathryn Miller** from the NIST Library helped finalize the document for public release.

***Acknowledgments*: NIST team members and Resource Group for their insights; all those who provided public comments**

FORENSIC SCIENCE
RESEARCH. STANDARDS. FOUNDATIONS.

5

Why is NIST involved in forensic science?

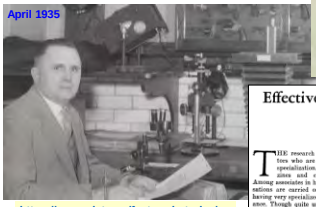
The assistance and technical expertise of NIST was requested by the U.S. Congress, Department of Justice, and others

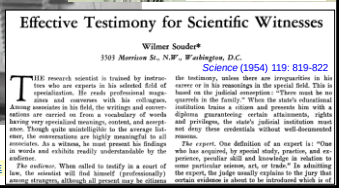
- **Establishment of FBI Laboratory** (early 1930s)
- Automated fingerprint detection (1960s to present)
- Law Enforcement Standards Laboratory (established in 1971)
- "Starch Wars" (1977 to 1978)
- Input on TWGDAM/SWGDAM (1988 to present)
- DNA reference materials (early 1990s to present)
- FBI's DNA Advisory Board (1995 to 2000)
- Digital forensics (late 1990s to present)
- National Institute of Justice (NIJ) funding (1970s to present)
- White House Subcommittee on Forensic Science (2009-2012)
- Memorandum of Understanding with DOJ leading to NCFS and OSAC (2013-2017)
- **Specific Congressional funding (ongoing)**

6

Dr. Wilmer Souder

(first NIST forensic scientist, 1911-1954)





Effective Testimony for Scientific Witnesses

Wilmer Souder*

3503 Morrison St., N.W., Washington, D.C.

Science (1954) 119: 819-822

THE research scientist is trained by instructors who are experts in his selected field of specialization. He reads professional magazines and converses with his colleagues. Among associates in his field, the writings and conversations are carried on from a vocabulary of words having very specialized meanings, content, and accepted uses. Though quite unintelligible to the average listener, the conversations are highly meaningful to all associates. As a witness, he must present his findings in words and exhibits readily understandable by the audience.

The audience. When called to testify in a court of law, the scientist will find himself (professionally) among strangers, although all present may be citizens the testimony, unless there are irregularities in his career or in his reasonings in the special field. This is based on the judicial conception: "There must be no quarrels in the family." When the state's educational institution limits a citizen and presents him with a diploma guaranteeing certain attainments, rights and privileges, the state's judicial institution must not deny these credentials without well-documented reasons.

The expert. One definition of an expert is: "One who has acquired, by special study, practice, and experience, peculiar skill and knowledge in relation to some particular science, art, or trade." In admitting the expert, the judge usually explains to the jury that certain evidence is about to be introduced which is of

https://www.nist.gov/featured-stories/who-was-detective-x

See video: https://www.youtube.com/watch?v=a97A44ORnrE

7

https://www.nist.gov/feature-stories/who-was-detective-x

Wilmer Souder is seen using an early comparison microscope to compare the rifling marks left on two bullets, a technique for determining whether the bullets were fired from the same gun. This technique for comparing bullets is still used today in much the same way. *Credit: Photo by NBS/NIST; source: NARA*

Dr. Wilmer Souder

- Obtained his PhD in physics at the University of Chicago in 1916
  - Under Nobel Prize winners Albert Michelson and Robert Millikan
- **Early pioneer in precision measurements** for handwriting, typewriting, and ballistics
  - Shown here using a comparison microscope in a 1929 photo
  - Learned from Calvin Goddard (who started Chicago's Scientific Detection Laboratory in 1930)
- Assisted in starting the FBI Laboratory in November 1932

8

FBI Laboratory Began Operations with NIST Input

Started on November 24, 1932 with Assistance of Dr. Wilmer Souder

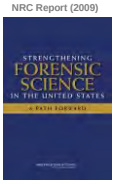
Page 47: "**The development of the [FBI] Laboratory has been carefully planned by the Division with the assistance and advice of Dr. Wilmer Souder**, a well-known and recognized authority in the field of scientific endeavor. Dr. Souder, who is at present acting in an advisory capacity in the further development of the Laboratory, has been engaged as **a scientist by the Bureau of Standards** for a period of eighteen years and has devoted the principle portion of his time to handwriting, typewriting and ballistics identification. **His advice and experience have rendered invaluable service to the Division in the training of the Laboratory personnel and in obtaining equipment which is considered the most desirable and essential for the performance of its work**."

From "A Digest of the Early History of the FBI Laboratory" (prepared by Fred M. Miller January 1956 for use by Don Whitehead in writing Chapter 16 of his 1956 book *The FBI Story*); a copy provided to NIST by FBI Historian John Fox on July 9, 2015

9

Requests for **Understanding What Data Exists** Supporting Forensic Science Methods and Practices

NRC Report (2009)



"demonstrating the validity of forensic methods" (Recommendation #3)

NCFS Recommendation (2016)

"technical merit evaluation"

Congressional funding uses NCFS language

PCAST Report (2016)

"establishing foundational validity"

NISTIR 8225 (2020)

NIST: a "Scientific Foundation Review"

10

NEW SLIDE

Trustworthy Results: A Shared Common Interest

- Obtaining **reliable (trustworthy, consistently accurate) results is an important goal** for forensic science, which NIST, as part of the forensic science ecosystem, shares in all our activities
- With NIST scientific foundation reviews, we are
  1. Documenting the **key scientific principles** that underpin current methods and practices
  2. Cataloging **available literature and information** that describe the state of the field
  3. Recommending strategies so that the community and its stakeholders **can have confidence in the results obtained** from a particular method or practice

11

NIST Scientific Foundation Reviews Underway in 2022

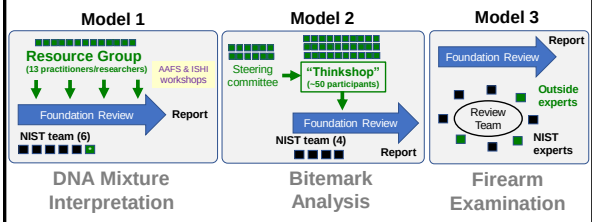
1. **DNA Mixture Interpretation** (initial pilot study)
   - Began in September 2017
   - AAFS 2019, ISHI 2019, ISHI 2020, AAFS 2021, AAFS 2022 workshops conducted
   - ***250-page report released*** *for public comment* ***on June 9, 2021****, with a 3-hour webinar held on July 21*
2. **Bitemark Analysis**
   - Began in October 2018
   - Workshop held in October 2019
3. **Digital Investigation Techniques**
   - Began in February 2019
   - Interlaboratory "black box" study conducted from June to November 2020 → *published Feb 2022*
4. **Firearm Examination**
   - Began in October 2019
   - Gathering literature and focusing on error rate studies

Reports will be provided with each foundation study and made available for a public comment period (usually 60 days)

https://www.nist.gov/topics/forensic-science/interdisciplinary-topics/scientific-foundation-reviews

12

Community Involvement and Input



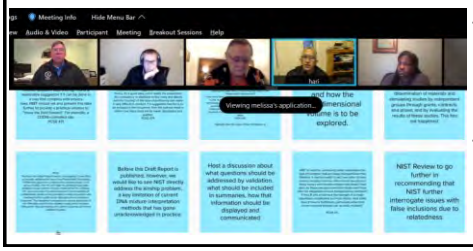
Model 4 Digital Evidence
Incorporated an interlaboratory study

Public Comment will be sought on our reports
(they will be initially released as "DRAFT")

13

NEW SLIDE

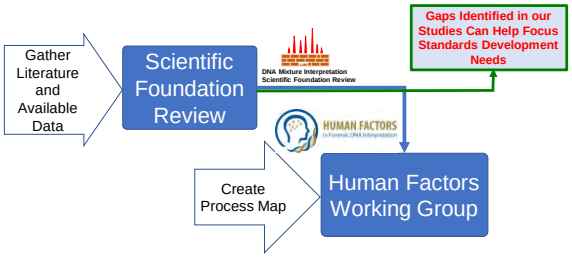
NIST Team Meets Regularly and Discusses Potential Responses to Public Comments Received



- Public comments have been reviewed and grouped into topics on a MURAL virtual bulletin board
- Each topic is carefully considered for potential revisions as we work to finalize our report

14

Inputs and Relative Relationships of Efforts



15

How Our Reports May Overlap with Other NIST Activities

- **Forensic Science Research**
  - information gaps identified can provide focus for forensic science research at NIST and elsewhere in the community
- **Forensic Science Standards (OSAC)**
  - information gaps identified can provide focus for standards needs
- **Forensic Science Foundations**
  - lessons learned to assist future efforts
- **Human Factors Studies and Process Maps**
  - e.g., Human Factors in Forensic DNA Interpretation Working Group (started Feb 2020)

Inform Future Forensic Science **Training Efforts**

16

Report Contents and Key Takeaways

17

DNA Mixture Report Content

In six chapters and two appendices:

- Chapter 1 **introduces the topic and challenges of DNA mixture interpretation**
- Chapter 2 provides background information on DNA, describes **principles and practices underlying mixture measurement and interpretation**, and introduces the likelihood ratio (LR) framework and probabilistic genotyping software (PGS)
- Chapter 3 lists **data sources used** in this study and strategies to locate them
- Chapter 4 and Chapter 5 cover **reliability and relevance**
- Chapter 6 explores the **potential of new technologies** to assist mixture interpretation and considerations for implementation
- Appendix 1 reviews the **history of how the field has progressed**
- Appendix 2 discusses strengthening the field with **training & continuing education**
- Bibliography includes 528 references cited in the report

18

How Was This Effort Organized?

1. **Team Efforts (regular discussion and report writing)**:
   - **John Butler** (SPO), **Peter Vallone** (MML), **Hari Iyer** (ITL), **Sheila Willis** (SPO Int Assoc), **Melissa Taylor** (SPO), **Rich Press** (PAO) – *report authors (see Table 1.1)*
2. **Resource Group (practitioner and researcher input)**:
   - **13 individuals** from federal, state, and local forensic laboratories or universities who met 12 times with NIST team from Dec 2017 to June 2019 *(see Table 1.2)*
3. **Additional Input (review and report editing):**
   - Content review and copy editing – *multiple individuals listed in acknowledgements*
   - NIST Office of Chief Counsel legal review
   - NIST Editorial Review Board examination

19

Who Conducted this NIST Review?

Name	NIST Operating Unit	Areas of Expertise
John M. Butler	Special Programs Office	Forensic DNA methods and scientific literature
Hari K. Iyer	Statistical Engineering Division, Information Technology Laboratory	Mathematics and statistics
Rich Press	Public Affairs Office	Communication and science writing
Melissa K. Taylor	Special Programs Office	Human factors (previous efforts in latent fingerprints and handwriting analysis)
Peter M. Vallone	Applied Genetics Group, Material Measurements Laboratory	DNA technology, research, rapid DNA, next-generation DNA sequencing
Sheila Willis	Special Programs Office (hired under contract as an **International Research Associate**)	Forensic laboratory management and trace evidence (retired director of Forensic Science Ireland)

Table 1.1 (p. 15)

NIST Team and the co-authors of this report

20

Our Desire with This Report is to Help Move the Field Forward to Improved Practices in DNA Mixture Interpretation

From the Executive Summary (page 1):
"As with any field, the scientific process (research, results, publication, additional research, etc.) continues to lead to advancements and better understanding. Information contained in this report comes from the authors' technical and scientific perspectives and review of information available to us during the time of our study. Where our findings identify opportunities for additional research and improvements to practices, we encourage researchers and practitioners to take action toward strengthening methods used to move the field forward. **The findings described in this report are meant solely to inform future work in the field**."

21

Clarification on What NIST Is and Is Not

- NIST is a Federal government **science agency** and does not comment on legal admissibility
- NIST is **not a regulatory agency**, which is why <u>key takeaways</u> are provided in our draft report rather than formal recommendations
- NIST **focuses on research and assisting with developing standards** (e.g., OSAC or SRMs); NIST does not conduct forensic science casework

22

Chapter Mapping

25 Key Takeaways (KT) and
8 Future Considerations (FC)

Chapter 1 INTRODUCTION	Chapter 2 PRINCIPLES	Chapter 3 SOURCES	Chapter 4 RELIABILITY	Chapter 5 RELEVANCE	Chapter 6 TECHNOLOGY	Appendix 1 HISTORY	Appendix 2 TRAINING
(none)	⇨ KT #2.1	(none)	⇨ KT #4.1	KT #5.1	KT #6.1	KT #A1.1	FC #A2.1
	⇨ KT #2.2		KT #4.2	KT #5.2	KT #6.2	KT #A1.2	FC #A2.2
	⇨ KT #2.3		KT #4.3	KT #5.3		KT #A1.3	FC #A2.3
	⇨ KT #2.4		KT #4.4	KT #5.4			FC #A2.4
	⇨ KT #2.5		KT #4.5	KT #5.5			FC #A2.5
	⇨ KT #2.6		KT #4.6	KT #5.6			FC #A2.6
			KT #4.7				FC #A2.7
	⇨ 16 Principles		KT #4.8				FC #A2.8
2 Tables	4 Tables 4 Figures	3 Tables	9 Tables 1 Box	5 Tables 3 Figures	3 Tables 5 Figures	3 Tables 4 Boxes	

Glossary & Acronyms: 47 terms **Bibliography: 528 references**

23

Chapter 2: Principles and Practices

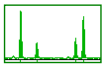
24

All DNA Samples Are Not Equivalent

KEY TAKEAWAY #2.1: DNA mixtures, where the DNA of more than one individual is present in a sample, are inherently more difficult to interpret than single-source DNA samples.

Math Analogy to DNA Evidence

2 + 2 = 4	$2x^2 + x = 10$	$\int_{x=0}^{\infty} f(x)dx$
Basic Arithmetic	Algebra	Calculus



Higher levels of uncertainty in determining contributing genotypes

Single-Source DNA Profile (DNA databasing)

Sexual Assault Evidence (2-person mixture with high-levels of DNA)

Touch Evidence (>2-person, low-level, complex mixtures perhaps involving relatives)

25

Steps in Processing a DNA Evidence Sample

(Single-Source or Mixture)

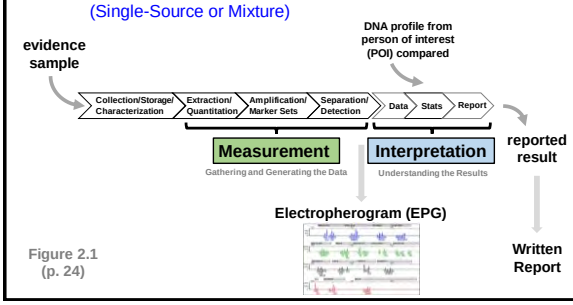


Figure 2.1 (p. 24)

26

Measurement and Interpretation Differ

KEY TAKEAWAY #2.2: Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves assigning values that are not inherent to the sample. To do this, the DNA analyst uses their judgment, training, tools (including computer software), and experience, and considers factors such as case context.

Table 2.1 (pp. 27-28)

Table 2.1. Measurable factors and features in a short tandem repeat (STR) DNA profile electropherogram (EPG) that influence DNA mixture interpretation with binary or probabilistic genotyping software (PGS) approaches.

Figure 2.2 (p. 29)

Figure 2.2. Steps in DNA mixture interpretation first outlined by the UK Forensic Science Service (Clayton et al. 1998) and endorsed by the ISFG DNA Commission (Gill et al. 2006b).

27

2.3.1 Factors that Contribute to Increased Complexity in DNA Mixtures

(p. 30)

There are at least three challenges that are fundamental to DNA mixture interpretation: (1) ***stochastic variation***, which impacts recovered quantities of alleles from contributors and can lead to uncertainty in assigning alleles to genotypes and uncertainty in assigning genotypes to contributor profiles when examining small amounts of DNA, (2) ***stutter products***, which create uncertainty through minor contributor(s) with alleles in the stutter positions of major contributor(s) alleles, and (3) ***sharing of common alleles***, which influences the ability to estimate the number of contributors, particularly when combined with stochastic variation and the existence of stutter products that create uncertainty in deconvoluting mixture components.

KEY TAKEAWAY #2.3: The process of generating a DNA profile can produce stochastic or random variation and artifacts that contribute to the challenge of DNA mixture interpretation.

28

Multiple Factors Influence Mixture Complexity

(p. 31)

The challenge of genotype assignment increases with the number of contributors in a mixture due to the possibility of allele sharing (Alfonse et al. 2017). In addition, estimating the number of contributors in a DNA mixture becomes more uncertain when there are more contributors as noted in several publications (Paoletti et al. 2005, Buckleton et al. 2007, Coble et al. 2015). The frequency of occurrence for an allele from population data correlates to the degree of allele sharing that is expected if that allele is present in the crime scene DNA mixture. If mixture contributors are related, then even more allele sharing between contributors is expected. Thus, with more contributors to a mixture, more allele sharing occurs, which increases the complexity and ambiguity of interpretation (e.g., Dembinski et al. 2018, Lynch & Cotton 2018).

KEY TAKEAWAY #2.4: DNA mixtures vary in complexity, and the more complex the sample, the greater the uncertainty surrounding interpretation. Factors that contribute to complexity include the number of contributors, the quantity of DNA from each contributor, contributor mixture ratios, sample quality, and the degree of allele sharing.

29

Table 2.2 (p. 33)

Different Statistical Approaches Answer Different Questions

This point is emphasized in **Principle 15**
"Different statistical approaches can produce different numerical results as they utilize different information and/or models and answer different questions."

Table 2.2. Different approaches used in statistical analysis of DNA and the questions addressed. RMP and MP are calculated for single source DNA profiles (or deduced major profiles). CPI and LR are calculated for mixtures.

	Question	Approach (Reference)	Specific Requirements
1	What is the probability of observing this profile in the population? (i.e., what is the rarity of the profile?)	Profile Probability (or random match probability, RMP) (NRC 1996 for single-source samples; Bille et al. 2013 for mixtures)	For mixtures, an assumption that the major contributor can be distinguished from minor components so that specific genotypes in the major can be inferred
2	What is the probability of observing this profile in the population if we have already observed one person with this profile in this population?	Match Probability (MP) (Balding & Nichols 1994, Weir 2001)	Use of conditional probabilities and a subpopulation correction
3	What is the probability that a person selected randomly in the population would be included (or not excluded) as a possible donor of the DNA typing result?	Combined Probability of Inclusion (CPI) (Bieber et al. 2016) **CPI**	All alleles for all contributors are all present in the reported loci (i.e., cannot cope with allele drop-out that is expected with low quantities of DNA)
4	By how much do the DNA typing results support the person of interest (POI) being the donor under specific assumptions and propositions?	Likelihood Ratio (LR) (Evett & Weir 1998) **LR**	An assumption as to the number of contributors and a specific pair of propositions

30

DNA Mixture Interpretation Approaches Compared

Table 2.3
(p. 35)

This point is emphasized in **Principle 14** "…continuous models use more information than discrete or binary approaches."

Table 2.3. Comparison of approaches used in DNA mixture interpretation. CPI = combined probability of inclusion, mRMP = modified random match probability, LR = likelihood ratio. Adapted from ISFG 2015 workshop by John Butler and Simone Gittelson available at https://strbase.nist.gov/training/ISFG2015-Basic-STR-Interpretation-Workshop.pdf.

	Takes into account		*Mathematically models*	
	Presence/ absence of alleles	Possible genotypes based on peak heights	Allele drop-out and allele drop-in	Peak heights
Binary Approaches				
CPI	X			
mRMP	X	X		
LR (binary)	X	X		
Probabilistic Genotyping				
LR (discrete)	X		X	
LR (continuous)	X	X	X	X

KEY TAKEAWAY #2.5: Continuous probabilistic genotyping software (PGS) methods utilize more information from a DNA profile than binary approaches.

31

Principles Described in Chapter 2

Principle 1 [Biology]: **Our DNA generally remains unchanged across time and cell type.**

- *This principle enables meaningful comparison of DNA from a reference sample to an evidence sample deposited and/or collected at a different time and to verify identity in a "biometric" sense, where a previously analyzed DNA profile is checked against a new one for "authentication" purposes.*

Principle 2 [Biology]: **DNA transfers and persists and can be collected and analyzed.**

- *This principle of direct or primary transfer enables results to be generated from evidentiary DNA profiles to assist in crime-to-crime and crime-to-individual associations.*

Principle 3 [Biology]: **Forensic DNA profiles examine a limited number of specific sites in the human genome.**

- *This principle is a reminder that the entire DNA sequence is not examined with forensic tests. Statistical assessments of profile rarity are used based on inheritance patterns and population genetics.*

32

REVISED

Principles Described in Chapter 2

Principle 7 [Relevance]: **Answers [derived] from DNA results depend on questions asked and circumstances of the evidence.**

The FBI DNA Advisory Board stated: "Proper statistical inference requires careful formulation of the question to be answered. Inference must take into account how and what data were collected, which, in turn, determine how the data are analyzed and interpreted" (DAB 2000). DNA results typically address questions at the sub-source level of the hierarchy of propositions (i.e., who could be the source of the DNA or is the DNA from the person of interest, Taroni et al. 2013). *This principle is a reminder to users that DNA information by itself can only answer "who" questions, that is, questions of source not activity.*

PC49: "The last sentence in italics is not true. DNA cannot answer 'who' questions by itself, ever…with proper evaluation, DNA can be used to give information related to activity."

Taroni F, Biedermann A, Vuille J, Morling N (2013) Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Science International: Genetics* 7:467-470.

33

These 16 Principles Form the Foundation for DNA Mixture Interpretation

- **P14**: Continuous models use more information
- **P15**: Results can differ with various approaches
- **P16**: Propositions impact strength of evidence
- **P11**: Stochastic variation impacts mixture ratios
- **P12**: Stutter peaks impact interpretation
- **P13**: Impacts on number of contributors estimate
- **P7**: Answers depend on questions asked
- **P8**: PCR can introduce artifacts
- **P9**: Peak positions and heights
- **P10**: Peak height variance
- **P4**: Established genetic inheritance patterns
- **P5**: Strength of evidence calculations use pop. gen.
- **P6**: Related DNA more similar than unrelated
- **P1**: DNA stability across time and cell type
- **P2**: DNA transferability
- **P3**: Forensic profiles only examine a portion of the human genome

34

NEW SLIDE

Public Comment on Principle 10

Draft Report:

- "Principle 10 [Measurement]: Relative fluorescence unit (RFU) variance (uncertainty) is inversely proportional to DNA profile peak height."

Proposed (by PC31):

- "Principle 10 [Measurement]: The variability (uncertainty) of peak height ratios (and heterozygote imbalance) increases as peak height decreases."
  - *PC31 notes*: …it is not the variability of peak heights that increases as peak height decreases. Peak height variability may increase with peak heights. It is the coefficient of variation (standard deviation divided by the mean) that increases as peak height decreases. This is reflected in the variability of peak height ratios and heterozygote imbalance…

35

Likelihood Ratios Are Not Measurements

(p. 42)

DNA mixture interpretation is performed in the face of uncertainty. As noted by Ian Evett and Bruce Weir in their 1998 book:

> "The origins of crime scene stains are not known with certainty, although these stains may match samples from specific people. The language of probability is designed to allow numerical statements about uncertainty, and we need to recognize that *probabilities are assigned by people rather than being inherent physical quantities*" (Evett & Weir 1998, p. 21, emphasis added).

KEY TAKEAWAY #2.6: Likelihood ratios are not measurements. There is no single, correct likelihood ratio (LR). Different individuals and/or PGS systems often assign different LR values when presented with the same evidence because they base their judgment on different kits, protocols, models, assumptions, or computational algorithms. Empirical data for assessing the fitness for purpose of an analyst's LR are therefore warranted.

36

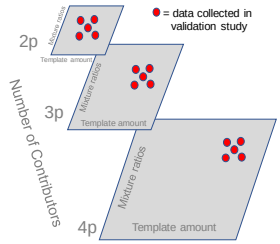
Chapter 4: Reliability of DNA Mixture Measurements and Interpretation

(4.1.1) System Reliability vs Component Reliability
(4.1.2) Definitions of Measurement, Uncertainty, Assessment, and Interpretation
(4.1.3) Empirical Assessments of Reliability
(4.1.4) Factor Space and Factor Space Coverage
(4.1.5) Provider-User Responsibilities and Examples
(4.2) Data Sources Used to Examine Reliability
(4.3) Review of Publicly Available Data and Factor Space Coverage
(4.4) Discussion
(4.5) Thoughts on a Path Forward

KEY TAKEAWAY #4.1: The degree of reliability of a component or a system can be assessed using empirical data (when available) obtained through validation studies, interlaboratory studies, and proficiency tests.

37

Factor Space and Factor Space Coverage



Factor Space Expands with the Number of Contributors (and degree of allele sharing typically increases)

- **Is a new term but not a new concept**
  - FBI QAS 8.3.2.1 requires laboratories to "include samples with a range of the number of contributors, template amounts, and mixture ratios expected to be interpreted in casework"
- **Table 4.1** lists influencing factors with DNA mixture measurements and interpretations using PGS systems
- Factor space coverage is summarized for 3 STR kit developmental validation studies (**Table 4.2**), 60 published PGS studies (**Table 4.3**), 11 publicly available internal validation summaries (**Table 4.5**), 83 proficiency test data sets (**Tables 4.6 and 4.7**), and 18 interlaboratory studies (**Table 4.8**)

38

NEW SLIDE

"Factor Space" in the Quality Assurance Standards

https://www.fbi.gov/file-repository/qas-guidance-document-070120.pdf/view

- Guidance Document for Quality Assurance Standards for Forensic DNA Testing and DNA Databasing Laboratories
  - This document clarifies standards and provides guidance to assist forensic DNA testing and DNA databasing laboratories and auditors in determining compliance. (Effective July 1, 2020)

Under **Forensic Standard 8.3.1**

- **Mixture studies:** Mixed DNA samples that are representative of those typically encountered by the testing laboratory shall be evaluated. Forensic mixture studies should use known samples that **represent the number of contributors and the range of general mixture types** for which the procedure will be used in casework (**e.g., mixture proportions, template quantities**) and must be used to develop interpretation guidelines.

"Factor Space Coverage" is not a New Concept, just a New Term!

39

Table 4.3 (pp. 66-69)

Factor Space Coverage for Published PGS Validation Studies

8 PGS studies were available and cited in the 2016 PCAST report

We examined and summarized 60 published PGS studies

Table 4.3. Factor space coverage for published PGS validation data from peer-reviewed literature. Studies are grouped by PGS system and publication date. Studies listed on row #6, #7, #10, #11, #12, #13, #14, and #49 were part of the PCAST 2016 review: Nikola Osborne and Sarah Riman (NIST Associates) assisted with early versions of these summaries. NoC = number of contributors; N.E.S. = not explicitly stated in the referenced publication; N/A = not applicable; [a]comparison of multiple PGS systems are discussed in Table 4.4. [b]inclusion of ranges is not meant to imply that all combinations of DNA quantities and mixture ratios were covered. [c]a 31-laboratory compilation (Bright et al. 2018) contained data from eight different STR kits: GlobalFiler, Identifiler Plus, NGM SElect, PowerPlex Fusion 5C, PowerPlex Fusion 6C, PowerPlex ESI17 Pro, PowerPlex ESI17 Fast, and PowerPlex 16 HS.

#	Reference	PGS System STR Kit	NoC Range	# samples by NoC	Total DNA Quantity Range (pg)[b]	Mixture Ratio Range[b]
1	Perlin & Szendrovi 2009	TrueAllele *PowerPlex 16*	2	40	125 to 1000	1:1 to 9:1
2	Perlin et al. 2011	TrueAllele *Pro+Cofiler*	2	16 *adjudicated cases*	N.E.S.	N.E.S.
3	Perlin et al. 2013	TrueAllele *Pro+Cofiler*	2; 3	*7*; 14 *adjudicated cases*	N.E.S.	N.E.S.
4	Balliauryne et al. 2013 (proof of concept)	TrueAllele *Identifiler*	2	2	N.E.S.	1:1
5	Perlin et al. 2014	TrueAllele *PowerPlex 16*	2; 3; 4	40; 65; 8 *adjudicated cases*	N.E.S.	N.E.S.
6	Perlin et al. 2015	TrueAllele *Identifiler Plus*	2; 3; 4; 5	10; 10; 10; 10 (5 donors)	200, 1000	1:1 to :32:16:15:2:1
7	Greenspoon et al. 2015	TrueAllele *PowerPlex 16*	1; 2; 3; 4	11; 18; 15; 7 (11 donors)	10 to 1000	1:1 to 1*7:1:1:1
59	You & Balding 2019 (data from Steele et al. 2018)	*multiple *NGM SElect*	1; 2; 3	36; 24; 12 (36 donors)	4 to 328	1:1 to 16:1; 1:1:1 to 16:4:1
60	Riman et al. 2021	*multiple *GlobalFiler*	2; 3; 4	154; 147; 127 (PROVEDIt data)	30 to 750	1:1 to 1:9; ... 1:1:1:1 to 1:9:9:1

40

Published PGS Comparison Studies

11 + 1 NIST study (conducted during our review)

Table 4.4 (pp. 69-72)

PGS Systems Compared	Samples Tested	Observations Made
Reference	*Supplementary Tables 4 and 5 contain all LR values and provide an example*	
A pre-print version is available at https://www.biorxiv.org/content/10.1101/2021.05.26.445891v1 EuroForMix (v2.1.0) STRmix (v2.6) Riman et al. 2021	Examined **154 two-person, 147 three-person, and 127 four-person mixtures** from the PROVEDIt dataset; see Supplemental Table 4 in their article	Provided LR values for 1279 Hp-true tests (Supplemental Table 4) and 1279 Hd-true tests (Supplemental Table 5) for each software; explored LR distributions observed and used ROC plots, scatter plots, histograms with distribution of differences; evaluated apparent discrepancies between PGS models, adventitious exclusionary and inclusionary support, and verbal equivalent discordance; the authors reported: "in certain cases differences in numerical LR values from both software resulted in differences in one or more than one verbal categories (Table 8). These differences were substantially more with low template minor contributors and higher [number of contributors]. . ."

Riman S, Iyer H, Vallone PM (2021) Examining performance and likelihood ratios for two different likelihood ratio systems using the PROVEDIt dataset. ***PLoS ONE (published since draft report released)***

41

Riman et al. 2021
Published Since Draft Report Released

(published Sept 17, 2021)

PLOS ONE

Update citation to Table 4.3 (#60) and Table 4.4

RESEARCH ARTICLE

Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset

Sarah Riman[1]*, Hari Iyer[2], Peter M. Vallone[1]

1 Applied Genetics Group, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America. 2 Statistical Design, Analysis, Modeling Group, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America

42

Recent Publication Comparing Two PGS Systems

(published Sept 22, 2021)

Will add to Table 4.3 (#61) and **Table 4.4**

Received: 3 June 2021 | Revised: 17 July 2021 | Accepted: 18 August 2021
DOI: 10.1111/1556-4029.14886

PAPER
Criminalistics

FORENSIC SCIENCES

A comparison of likelihood ratios obtained from EuroForMix and STRmix™

Kevin Cheng MSc[1,2] | Øyvind Bleka PhD[3] | Peter Gill PhD[3,4] | James Curran PhD[2] | Jo-Anne Bright PhD[1] | Duncan Taylor PhD[5,6] | John Buckleton DSc[1,2]

Explores similarities and differences between software with single-source profiles and 129 mixtures (from the PROVEDIt dataset). "Likelihood ratios (LR) differences between the probabilistic genotyping software EuroForMix and STRmix are examined. After considering differences in the allele probabilities, the LRs from both software for an unambiguous single-source profile were identical (four significant figures)… **LRs for 84% of the comparisons for known contributors without rare alleles were within two orders of magnitude**."

43

Recent PGS Review Article (published Sept 30, 2021)

genes

Open access (freely available) at https://www.mdpi.com/2073-4425/12/10/1559

MDPI

Review

A Review of Probabilistic Genotyping Systems: *EuroForMix, DNAStatistX* and *STRmix*™

Peter Gill [1,2,*], Corina Benschop [3], John Buckleton [4,5], Øyvind Bleka [1] and Duncan Taylor [6,7]

[1] Forensic Genetics Research Group, Department of Forensic Sciences, Oslo University Hospital, 0372 Oslo, Norway; oeybleb@ous-hf.no
[2] Department of Forensic Medicine, Institute of Clinical Medicine, University of Oslo, 0315 Oslo, Norway
[3] Division of Biological Traces, Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands; c.benschop@nfi.nl
[4] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand; john.buckleton@esr.cri.nz
[5] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand
[6] Forensic Science SA, GPO Box 2790, Adelaide, SA 5001, Australia; Duncan.Taylor@sa.gov.au
[7] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
* Correspondence: peter.d.gill@gmail.com

44

Outreach and Public Comments Received

45

NIST Process

46

Resource Group

- Question (from July 21 webinar): **Does the report reflect a consensus across the scientists mentioned in Table 1.2 (the Resource Group members)?**
- "**Acknowledgments:** Members of the DNA Mixture Resource Group (listed in Table 1.2) contributed helpful feedback and assistance in the early stages of drafting this report." [DRAFT, p. i]
- "A DNA Mixture Resource Group (see Table 1.2), with extensive experience in public and private forensic DNA laboratories, reviewed an early draft of our report and provided valuable feedback, insights, and suggestions. However, **they were not asked to sign off on our final report or endorse its conclusions**. The NIST team is grateful for their dedication and contributions to our efforts." [DRAFT, p. 2]

47

Public Comments Shared in December 2021

Public comments received

on NISTIR 8351-DRAFT

DNA Mixture Interpretation: ***A NIST Scientific Foundation Review***

Published December 3, 2021

- Public comment periods
  - June 9 to August 23, 2021
  - October 22 to November 19, 2021
- **63 public comments (PC1 to PC63)**
  - Plus 83 questions/comments from July 21 webinar
  - Six (6) provided feedback in both comment periods
  - 30 practitioners
  - 11 researchers
  - 10 stakeholder organizations
  - 5 consultants/other
  - 4 lawyer (prosecution and defense) perspectives
  - 3 PGS vendor perspectives

https://www.nist.gov/dna-mixture-interpretation-nist-scientific-foundation-review

48

There Are Many Different Perspectives and Lenses on Our DNA Mixture Interpretation Report…

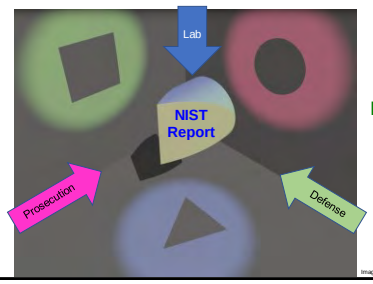


Image source: https://imgur.com/gallery/1zZ8V5e

This is Why Public Comment is so Important!

49

Expectations for Our Report Appear to Vary Based on Public Comments Received

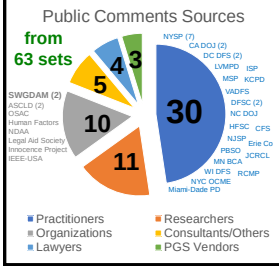
Summary of Some Expressed Expectations

1. Scope of this report and study should only be on PGS reliability
   Chapters 5 & 6 are not appropriate and should be removed or in a separate study
2. Practitioners from accredited forensic laboratories must be co-authors and if their perspectives are not represented, then the report findings cannot be useful
3. Stronger statements are needed
   e.g., on specific validation requirements, racial justice issues, or even calling for a moratorium on using PGS for complex mixtures until full reliability assessments can be performed by NIST or some other group
4. Full and complete solutions should be provided to every issue raised

Writing this report has been challenging (and taken much longer than expected)

50

Public Comments Received on Our Draft Report



- We are extremely grateful for the detailed feedback provided during our public comment periods and acknowledge the significant time and effort of those who carefully read and provided valuable written feedback on our draft report
- **We are carefully considering each comment as part of the NIST process to finalize this report and working to clarify language regarding data**
- A final report will be issued when we have completed this process

51

Some Associated Topics

52

NEW SLIDE

Common Themes/Questions Received

- How is the NIST team considering the public comments received?
- Why only seek information available in the public forum (disagree with KT4.3)?
- What is missing in data or data summaries that are available?
- Why is Chapter 5 in the draft report (relevance is not an analyst's job)?
- Have perspectives changed from Butler's 2006 Urban Legends of Validation?
- Why use "factor space"?
- Why use ROC plots?
- Why is the proficiency test information included?
- Will new references/data be added in the final report?

53

NEW SLIDE

NISTIR 8225 Principle of Information Retrievability

Retrievability is among the principles and criteria because transparency and openness are hallmarks of good science [5]. Therefore, we believe that for something to be considered foundational, it must be reasonably accessible to anyone who wishes to review it.

Where peer-reviewed publications are not available, transparency and accessibility can help fill the gap. For instance, publishing validation data from forensic laboratories online would allow for "open peer review" [6].

We recognize that there is a degree of subjectivity and judgment in determining what retrievable information has foundational merit in terms of being reliable and respected. A rigid checklist or algorithmic evaluation is not planned for every publication describing a forensic method or practice under review. Authors of NIST scientific foundation reviews will strive to use their collective experience in assessing sources of information as defined above. Since it is possible that an important source of retrievable information may be missed in conducting a foundational review, we plan to initially release reports as drafts for public comment. It must be kept in mind that some publications may be commonly cited to illustrate a problem, so frequent citation alone does not provide an adequate criterion for foundational information. In areas of rapid development, a body of literature might be the foundation rather than a single article.

- Information needs to be available for others to independently assess claims of scientific validity
- Journals currently do not require underlying PGS data to be available when an article is accepted (as noted by PC2)

https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8225.pdf (see page 2)

54

NEW SLIDE

Some Public Comments on Chapter 5

- **PC11**: "One of the central themes of Chapter 5 is relevance, but the overall point seemed to have gotten lost due to the large amount of information that was being provided."
- **PC14**: "chapter 5…is a complete deviation from the intended scope of this report"
- **PC24**: "This chapter [Chapter 5] is a welcome addition to the draft report. In our view, it addresses some important elements but should acknowledge that context is the key driver in the relevance of everything we do as forensic scientists."
- **PC50**: "Chapters 5 and 6 of the draft report contain interesting comments, observations and recommendations but may be more appropriate as a separate publication for forensic science practitioners as they are not directly relevant to the foundation review, nor mixture interpretation generally."
- **PC56**: "First, I want to acknowledge the importance of writing about this topic [Chapter 5] in this foundation review… My brain is stuck on the word 'relevance.' I'd love to see this chapter called what it is: *The Evaluation of Findings Given Activities: To Consider Transfer, Persistence, Prevalence, and Background.* … I would like to reiterate that the DNA analyst cannot determine relevance – but what we can do is assess the results given the disputed facts (activities) – which can help the jury/factfinder make a decision about the 'relevance' of the DNA."

Shorten?

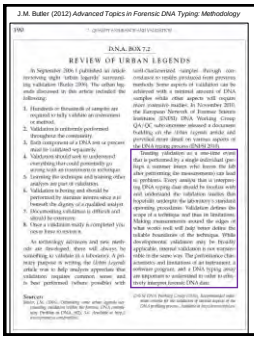
Emphasize Context More?

New Title?

These comments all come from active forensic DNA analysts

55

J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*


190

DNA BOX 7.2

REVIEW OF URBAN LEGENDS

In September 2006, I published an article involving eight 'urban legends' surrounding validation (Butler 2006). The urban legends discussed in this article included the following:

1. Hundreds or thousands of samples are required to fully validate an (extraction) or method.
2. Validation is uniformly performed throughout the community.
3. Each component of a DNA test or process must be validated separately.
4. Validation should work to understand everything that could potentially go wrong with an examination or technique.
5. Learning the technique and training other analysts are part of validation.
6. Validation is boring and should be performed by summer interns or at least with the degree of a peer-validated analysis.
7. Documenting validation is difficult and should be extensive.
8. Once a validation study is completed you are done forever.

As technology advances and new methods are developed, there will always be something to validate in a laboratory. A primary purpose of writing the 'Urban Legends' article was to help analysts appreciate that validation requires common sense and is best performed (where possible) with such characteristics as samples through considered to results produced from previous methods. Some aspects of validation can be achieved with a minimal amount of DNA samples while other aspects will require more extensive studies. In November 2010, the European Network of Forensic Science Institutes (ENFSI) DNA Working Group QA/QC subcommittee released a document (building on the Urban Legends article and provided more detail on various aspects of the DNA testing process (ENFSI 2010).

Treating validation as a one-time event that is performed by a single individual (perhaps a summer intern who leaves the lab after performing the measurements) can lead to problems. Every analyst that is interpreting DNA typing data should be familiar with and understand the validation studies that hopefully underpin the laboratory's standard operating procedures. Validation defines the scope of a technique and thus its limitations. Making measurements around the edges of what works well will help better define the reliable boundaries of the technique. While developmental validation may be broadly applicable, internal validation is not transferrable in the same way. The performance characteristics and limitations of an instrument, a software program, and a DNA typing assay are important to understand in order to effectively interpret forensic DNA data.

Sources:

My Comments on My Urban Legends

"Treating validation as a one-time event that is performed by a single individual (perhaps a summer intern who leaves the lab after performing the measurements) can lead to problems. **Every analyst that is interpreting DNA typing data should be familiar with and understand the validation studies that hopefully underpin the laboratory's standard operating procedures.** Validation defines the scope of a technique and thus its limitations. Making measurements around the edges of what works well will help better define the reliable boundaries of the technique. While developmental validation may be broadly applicable, internal validation is not transferrable in the same way. "

"The performance characteristics and limitations of an instrument, a software program, and a DNA typing assay are important to understand in order to effectively interpret forensic DNA data."

56

NEW SLIDE

Validation - What question are you answering?

- Conventional STR typing with single-source samples
  - E.g., if adopting a new STR kit, then genotype concordance is crucial to demonstrate
  - **This was the focus and context of the Urban Legends article (2006) – mentioned by PC33, PC50**
- Change in method (sensitivity improvements)
  - Interpretation approaches may more complicated
  - Setting interpretation guidelines in SOPs is different from method demonstration (KT2.2)
- LR systems with PGS
  - Complex mixtures with low-level DNA (at least some of the components of mixtures)
  - Not perfectly reproducible (in part due to MCMC and in part to the stochastic effects)
  - The more complex the method, the more testing is needed to understand the limitations

57

NEW SLIDE

DNAmix 2021 — Request for participation

- DNAmix 2021 is a large-scale independent study being conducted to evaluate the extent of consistency and variation among forensic laboratories in interpretations and statistical analyses of DNA mixtures, and to assess the effects of various potential sources of variability.
- The study is being conducted by Noblis and Bode Technology, under NIJ grant # 2020-R2-CX-0049.
- Participation is open to all forensic laboratories that conduct DNA mixture interpretation as part of their SOPs
  - Non-U.S. laboratories are welcome to participate if they report interpretations in English.
- The study will be composed of four phases:
  - Questionnaire to assess laboratory policies and procedures relevant to DNA mixture interpretation
  - Questionnaire to assess how participants would conduct analyses for various casework-derived scenarios
  - Assessment of suitability and number of contributors, given electropherogram data for 14 mixtures
  - Interpretations and statistical analyses, given electropherogram data for 7 mixtures, each provided with DNA profiles of potential contributors
- ***Registration will be open through 6 March 2022!***
- Interested? Register at https://dnamix.edgeaws.noblis.org
- Questions? Contact DNAmix@noblis.org

This slide was provided by **Austin Hicklin** from Noblis, who will present some preliminary results from the laboratory policies and procedures questionnaire here at AAFS on **Friday at 2pm-2:15pm (E124)**

58

NEW SLIDE

New References to Consider for Report Bibliography

New articles continue to come out on a regular basis…

- Additional PGS publications
  - Review article on EuroForMix, DNAStatistX, and STRmix (Gill et al. 2021)
  - STRmix and EuroForMix comparisons (Riman et al. 2021, Cheng et al. 2021)
  - Mixture Solution from Charles Brenner (Lucassen et al. 2021)
  - MaSTR from Soft Genetics (Holland et al. 2022)
  - A mixed DNA profile controversy revisited (Kalafut et al. 2022)
  - Probabilistic genotyping of single cell replicates (Huffman et al. 2022)
- Additional DNA transfer publications
  - Recent progress towards meeting challenges (van Oorschot et al. 2021)
  - Shedder status categorization (Goray & van Oorschot et al. 2021)
  - DNA transfer without contact (Thornbury et al. 2021)

We May Develop a Most Valuable Publication (MVP) List for Each Forensic Method Studied

59

Thank you for your attention!

John Butler
john.butler@nist.gov

https://www.nist.gov/topics/forensic-science

NIST FORENSIC SCIENCE

RESEARCH. STANDARDS. FOUNDATIONS.

Questions?

60

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

Examining Probabilistic Genotyping Systems

National Institute of Standards and Technology (NIST)

Sarah Riman
Applied Genetics Group

Module 3

1

Acknowledgments and Disclaimer

Points of view are the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities 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 National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

2

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

3

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

4

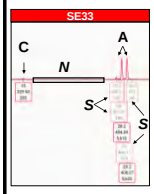
Different approaches to assign LR values

Probabilistic genotyping software

Binary | Semi-Continuous | Continuous | ***LR***

5

Continuous PGS use quantitative information contained within a profile



A = True known allelic peaks
S = Different stutter types (greyed-out)
C = Drop-in
N = Noise peaks

biology, statistics, mathematical models

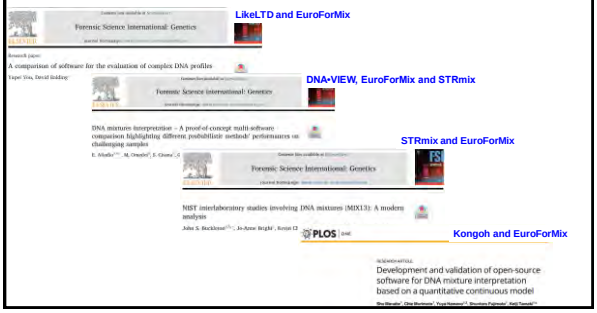
(Proprietary) **STRmix**
(Open source) **EuroForMix**

Statistefix 4.0
DNAxs/DNAStatistX
CEESIt
TrueAllele
Kongoh

Resolve genotypes
Or
Assign *LR values*

6

Few studies explored the degree of variability in LR values across various fully continuous PGS



7

Few studies explored the degree of variability in LR values across various fully continuous PGS

These studies:

- Had limited number of samples
- Did not quantify the differences in LRs
- Concluded that the models yielded similar LRs

8

Inter-model variability impact numerical values and verbal expression of the LRs

9

Motivation of the work

To perform an **independent study** and understand the amount of variability expected when mixtures are interpreted using different systems

Published: September 17, 2021

Highlights:

- first study that includes large-scale comparison of LRs
- uses publicly available data
- uses 2 reputable and well-cited fully continuous PG models
- evaluates the extent to which different models disagree (e.g. by a factor of 10, factor of 100, more than a factor of 1000)
- outlines the steps that may be used by other labs to assess the performance of different LR systems and analyze the resulting data
- shares the generated LR values in the interest of transparency and literature-to-literature comparisons by other researchers

The results are expected to vary if other parties conduct a similar analysis but use different software versions and protocols.

10

Another comparison study of LRs between STRmix and EuroForMix was published by the software developers

Published: September 22, 2021

Highlights

- "A comparison of likelihood ratios (LR) between two probabilistic genotyping software – EuroForMix and STRmix
- Similarities and differences between software were assessed with single-source profiles and 129 mixtures
- Results demonstrate that even though there are differences, both software can be useful in assigning an LR".

"…….PCAST advocated that this comparison should be carried out by independent groups (i.e. not the developers of the software. **An independent comparison of EuroForMix (version 2.1) and STRmix™ (version 2.6) was recently published out by Riman et al. [18]. We believe that our concurrent study reinforces the findings from Riman et al**……"

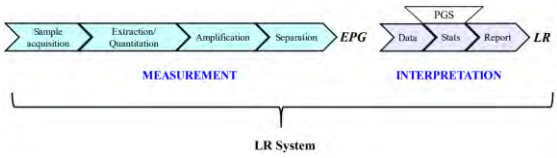
11

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

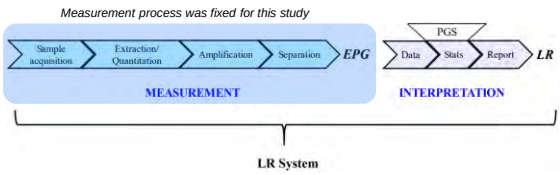
12

Definition of LR System



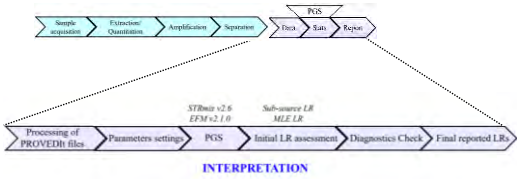
13

Definition of LR System



14

The performance assessment in this study encompasses the interpretation process only



The assigned LR values are a product of the decisions that went into the interpretation process of the LR system and not solely the PGS.

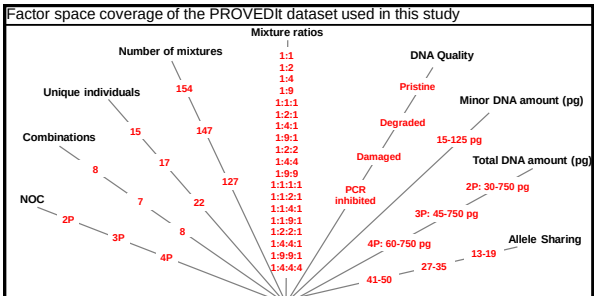
15

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

16

Factor space coverage of the PROVEDIt dataset used in this study



[1] L.E. Alfonse, A.D. Garrett, D.S. Lun, K.R. Duffy, C.M. Grigicak, A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt, Forensic science international. Genetics 32 (2018) 62-70.

17

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

18

Parameters of the interpretation process

Thursday—2:45 p.m. – 3:00 p.m. B82

Interpretation summary	STRmix v2.6	EuroForMix v2.1.0
Analytical thresholds (ATs) settings	Per dye channel ATs	Allows only one overall AT value
Drop-in	Drop-in frequency = 0.0015; drop-in cap = 180 RFU; uniform distribution	Drop-in probability = 0.0015; Drop-in hyper-parameter (λ) = 0.032
Stutter models applied	N-1, N-2 and N+1	N-1
Model maker parameters	333 single source profiles	NA
Diagnostic statistics	Per locus LR, deconvolution, genotypic weights, Gelman-Rubin statistics, & log likelihood	Per locus LR, deconvolution, genotypic weights, & model selection
Sub-source LR values	Labeled as sub-source LRs	Labeled as MLE based LRs
Input files	Same/fixed mixture EPG features (filtered CSV files from the PROVEDIt) Analyzed using the per dye ATs (B=35; G=65; Y=45; R=50; P=60)	
Mixture vs POI	Same combination of comparisons per each analysis	
Allele frequencies	NIST 1036-Caucasian	
F_{ST} (θ)	0.01	
Propositions	Same defined pair of propositions	
Number of contributors (NOCs)	Ground truth	

Determination and impact of NOC (ongoing project)

19

Parameters of the interpretation process

Interpretation summary	STRmix v2.6	EuroForMix v2.1.0
Analytical thresholds (ATs) settings	Per dye channel ATs	Allows only one overall AT value
Drop-in	Drop-in frequency = 0.0015; drop-in cap = 180 RFU; uniform distribution	Drop-in probability = 0.0015; Drop-in hyper-parameter (λ) = 0.032
Stutter models	N-1, N-2 and N+1 stutter peaks modeled	N-1 stutter peaks only modeled
Model maker parameters	333 single source profiles	NA
Diagnostic statistics	Per locus LR, deconvolution, genotypic weights, Gelman-Rubin statistics, & log likelihood	Per locus LR, deconvolution, genotypic weights, & model selection
Sub-source LR values	Labeled as sub-source LRs	Labeled as MLE based LRs
Input files	Same/fixed mixture EPG features (filtered CSV files from the PROVEDIt) Analyzed using the per dye ATs (B=35; G=65; Y=45; R=50; P=60)	
Mixture vs POI	Same combination of comparisons per each analysis	
Allele frequencies	NIST 1036-Caucasian	
F_{ST} (θ)	0.01	
Propositions	Same defined pair of propositions	
Number of contributors (NOCs)	Ground truth	

The assigned LR values are a product of the decisions that went into the interpretation process of the LR system and not solely the PGS. The results are expected to vary if other parties conduct a similar analysis but use different software versions and protocols.

20

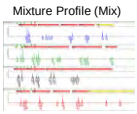
Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

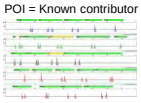
21

LR assessment workflow of the H1-true (known contributor) tests

Mixture Profile (Mix)



POI = Known contributor



Assign an LR using

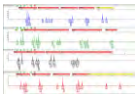
Mix	POI	NOC	$Log_{10}(LR)_{STRmix}$
Mix1	48	2	15.42
Mix2	33	3	21.37
⋮	⋮	⋮	⋮
Mix100	29	4	12.93

Mix	POI	NOC	$Log_{10}(LR)_{EFM}$
Mix1	48	2	15.38
Mix2	33	3	24.27
⋮	⋮	⋮	⋮
Mix100	29	4	10

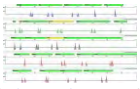
22

LR assessment workflow of the H2-true (known non-contributor) tests

Mixture Profile (Mix)



POI = Known non-contributor



(random person from NIST 1036)

Assign an LR using

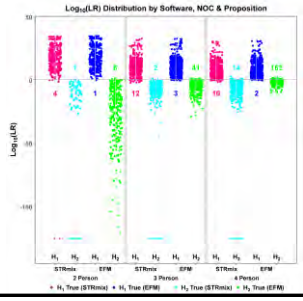
Mix	POI	NOC	$Log_{10}(LR)_{STRmix}$
Mix1	Y1	2	-1.75
Mix2	C20	3	-8.71
⋮	⋮	⋮	⋮
Mix100	Y33	4	-26.63

Mix	POI	NOC	$Log_{10}(LR)_{EFM}$
Mix1	Y1	2	-1.25
Mix2	C20	3	-4.23
⋮	⋮	⋮	⋮
Mix100	Y33	4	-6.46

23

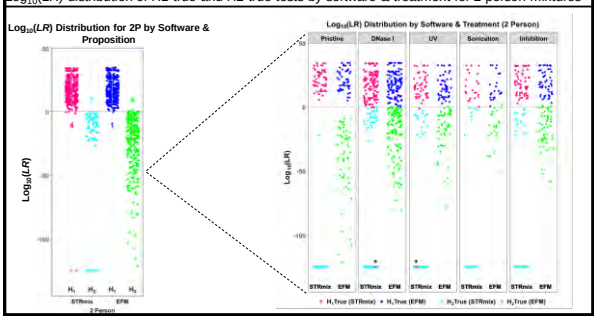
Qualitative assessment of the discrimination performance

Overall distribution plots



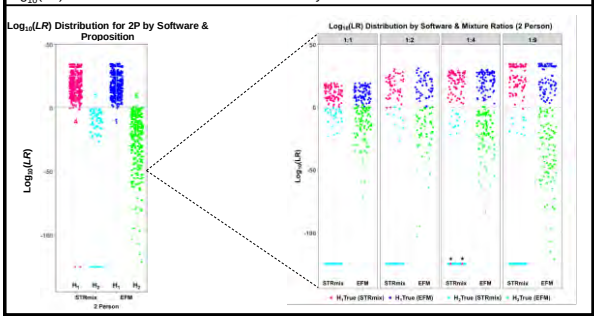
24

Log_{10}(LR) distribution of H1-true and H2-true tests by software & treatment for 2 person mixtures



25

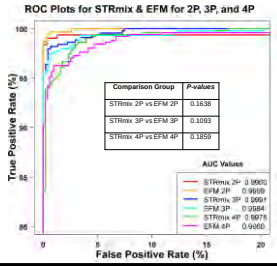
Log_{10}(LR) distribution of H1-true and H2-true tests by software & mixture ratios for 2P mixtures



26

Quantitative assessment of the discrimination performance by Receiver Operating Characteristic (ROC) plots

Sensitivity and specificity at every LR decision threshold

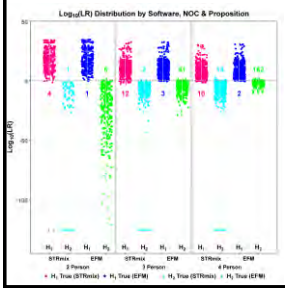


Comparison Group	P-values
STRmix 2P vs EFM 2P	0.1638
STRmix 3P vs EFM 3P	0.1093
STRmix 4P vs EFM 4P	0.1859

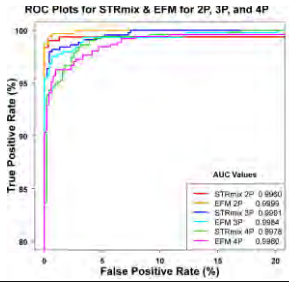
27

Qualitative and quantitative assessment of the discrimination performance

Overall distribution plots



Receiver Operating Characteristic (ROC) plots



28

Qualitative and quantitative assessment of the discrimination performance

Overall distribution plots

Receiver Operating Characteristic (ROC) plots

- The ability of the two LR systems to discriminate between known contributors and known non-contributors in aggregate are statistically indistinguishable for the data we considered.
- However, that did not imply that STRmix and EFM assigned equal LR values on a case-by-case basis.

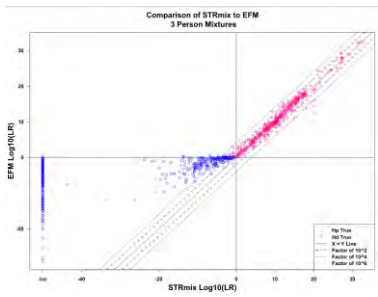
29

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

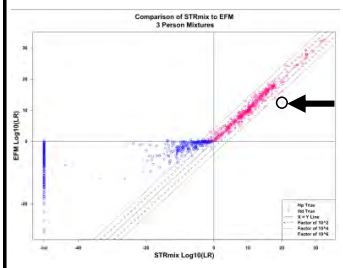
30

Case-by-case assessment of the assigned LR values between the two LR systems



31

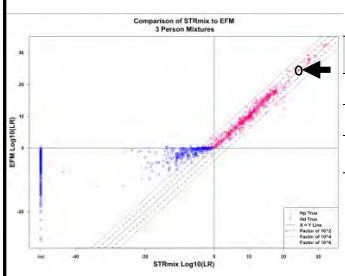
LR (STRmix) > LR (EFM)



Software	$Log_{10}(LR)$
STRmix	20.11
EFM	12.61
STRmix - EFM	***7.5***

32

LR (STRmix) = LR (EFM)

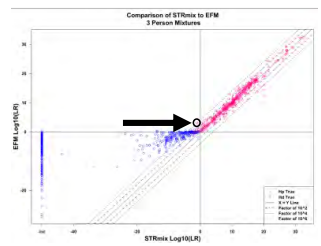


Software	$Log_{10}(LR)$
STRmix	28.91
EFM	28.91
STRmix - EFM	***~ 0***

33

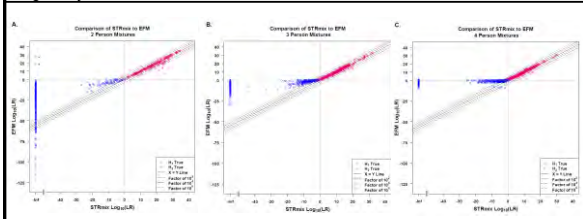
LR (EFM) > LR (STRmix)

Software	$Log_{10}(LR)$
EFM	3.16
STRmix	-1.14
EFM - STRmix	***4.30***



34

Case-by-case assessment of H1-true tests and H2-true tests profile log10(LR) values assigned by STRmix and EFM



- Differences in log10(LR) values assigned by STRmix and EFM at the profile level covered a wide range from 0 to over a million.
- Differences between the two software were higher with H2-true tests than the ones observed with H1-true tests.

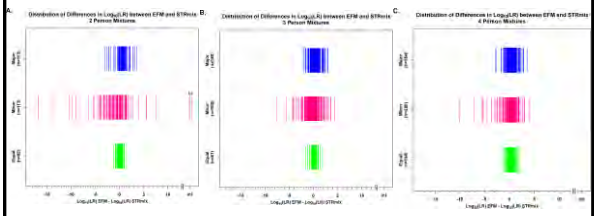
35

Overview

- Motivation of the work
- Definition of the LR system
- Factor space coverage
- Parameter settings of the interpretation process
- Discrimination performance of the two LR systems
- Comparison of LRs obtained by the two systems on a case-by-case basis
- Distribution of differences in LRs between the two LR systems

36

The magnitude of the differences in LR values between EFM and STRmix was greater for minor donors than major and equal contributors



37

Summary

Within our defined LR systems and for the data considered:

- STRmix and EFM had similar discrimination performance
- STRmix and EFM did not always assign equal LR values on a case-by-case basis
- Differences in LR values were observed in both directions (e.g., when LR STRmix ≥ LR EFM or when LR EFM ≥ LR STRmix
- The magnitude of the differences was greater with minor donors than with equal or major contributors

Examining more than one PGS with similar discrimination power especially with low-template profiles or minor contributor cases can be beneficial and an additional empirical diagnostic check even if software in use does contain certain diagnostic statistics as part of the output.

38

Our intent from this study is to:

- Highlight the importance of examining more than one PGS with similar discrimination power that can
  - lead to improving one or both models
  - be an additional empirical diagnostic check even if software in use does contain certain diagnostic statistics as part of the output
- Understand the variability in LR values across different PG models.
- Demonstrate the value of using a publicly available ground truth known mixture data to assess performance of any LR system.
- Outline steps to assess the performance of different software and analyze the resulting data.
- Share the generated LR values in the interest of transparency and literature-to-literature comparisons.

- The focus of this study is not to suggest that any one of the software is based on a true or best model.
- The results are expected to vary if other parties conduct a similar analysis but use different software versions and protocols.

39

Acknowledgement

Pete Vallone (NIST)

Hari Iyer (NIST)

John Butler (NIST)

Sicen Liu (JHU)

Øyvind Bleka (Oslo University Hospital)

Zane Kerr (ESR)

Steven Myers (CAL DOJ)

Contact: sarah.riman@nist.gov

Funding
NIST Special Programs Office: *Forensic DNA*

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

40

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

DNA Mixture Standards on the OSAC Registry

National Institute of Standards and Technology (NIST)

John Paul Jones
Special Programs Office

Module 4

1

Agenda

1. Introduction to Organization of Scientific Area Committees (OSAC) for Forensic Science
2. OSAC Registry
3. Standards to Discuss
   - ASB 20 (Validation & Verification)
   - ASB 40 (Interpretation & Comparison Protocols)
   - ASB 18 (Validation of Prob Gen)
   - OSAC Proposed 2020-N-0007 (DNA Elimination Databases)
   - OSAC Proposed 2020-S-0004 (Failed Controls & Contamination)
4. Staying Connected

2

OSAC's Objective & Core Principles

HARMONIZATION

BALANCE

To create a sustainable organizational infrastructure dedicated to identifying and **fostering the development of technically sound, consensus-based documentary standards** and guidelines for **widespread implementation** throughout the forensic science community

OPENNESS

CONSENSUS

3

OSAC's Structure

Forensic Science Standards Board (FSSB)

Seven Scientific Area Committees (SACs)

22 Subcommittees (SCs)

Four Resource Task Groups:

- Human factors
- Legal
- Quality
- Statistics

4

OSAC Membership Snapshot & What They Do...

478 members
324 active affiliates
3,300+ applications received

Employer Classification

- Federal: 20%
- State: 21%
- Local: 19%
- Academic: 21%
- Private: 17%
- FFRDC: 1%

Job Classification

- Practitioner: 51%
- Researcher: 18%
- Educator: 10%
- Lab Mgr/Director: 8%
- Other: 4%
- Quality: 3%
- Lawyer: 3%
- Judge: 2%
- R&D Tech: 1%

✓ **Facilitate development** of science-based standards through the formal SDO processes
✓ **Evaluate OSAC proposed and SDO published standards** for placement on the OSAC Registry
✓ **Promote** implementation of standards on the OSAC Registry
✗ Publish standards
✗ Have the authority to enforce standards

5

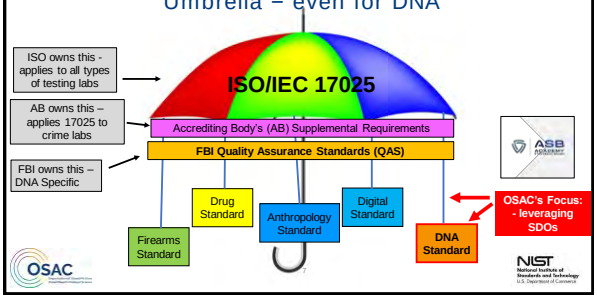
ISO/IEC 17025:2017 General Requirements for the Competence of Testing and Calibration Laboratories

- This document specifies the general requirements for the competence, impartiality and consistent operation of laboratories.
- This document is applicable to all organizations performing laboratory activities, regardless of the number of personnel.
- Laboratory customers, regulatory authorities, organizations and schemes using peer-assessment, accreditation bodies, and others use this document in confirming or recognizing the competence of laboratories.

https://www.iso.org/obp/ui/#iso:std:iso-iec:17025:ed-3:v1:en

6

Adding More Specificity Under ISO's 17025 Umbrella – even for DNA



7

OSAC Registry

- Repository of high-quality, technically sound published and proposed standards and guidelines for forensic science.
- All standards on the OSAC Registry have passed a rigorous technical and quality review by OSAC members, including forensic science practitioners, research scientists, statisticians and legal experts.
- OSAC encourages the forensic science community to implement published and proposed standards.

8

8

OSAC Registry: Current Snapshot

77

67 published
10 OSAC Proposed

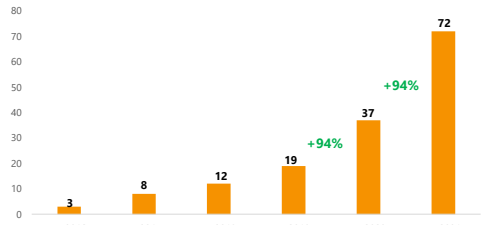
- **1** Anthropology
- **10** Biology/DNA (8 published & 2 OSAC Proposed)
- **1** Bloodstain Pattern Analysis
- **2** Crime Scene Investigation & Reconstruction (2 OSAC Proposed)
- **3** Digital Evidence*
- **1** Dogs & Sensors
- **4** Facial Identification (3 published & 1 OSAC Proposed)*
- **2** Fire & Explosion Investigation
- **6** Fire Debris
- **2** Firearms & Toolmarks
- **2** Footwear & Tire
- **2** Gunshot Residue Analysis (1 published & 1 OSAC Proposed)
- **5** Medicolegal Death Investigation (3 published & 2 OSAC Proposed)
- **3** Odontology
- **7** Seized Drugs
- **4** Toxicology
- **11** Trace Materials
- **1** Video/Imaging Technology & Analysis*
- **6** Wildlife Forensics (4 published & 2 OSAC Proposed)
- **5** Interdisciplinary

https://www.nist.gov/osac/osac-registry

*ASTM E2916-19e1 drafted in collaboration with OSAC's Digital Evidence, Facial ID, and VITAL Subcommittees

9

OSAC Registry Growth



10

10

OSAC Current Standards Activities

Tier 1	Tier 2	Tier 3	Tier 4
• Standards on the **OSAC Registry** • Approved by OSAC – highest level of vetting	• OSAC supported standards **published** by an SDO and **eligible for the Registry** • Completed SDO consensus process	• OSAC drafted standards **sent to an SDO** • Drafted with input from RC and approved by SAC	• **Under development** • Working draft document inside OSAC development process and not yet publicly available
77 standards	80 standards	129 standards	161 standards

11

11

ANSI/ASB 20 Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol, First Edition 2018

1 Scope

1.1 This standard sets forth the requirements for the design and evaluation of internal validation studies for mixed DNA samples and the development of appropriate interpretation protocols for mixtures based on the validation studies performed. This standard includes a requirement that the laboratory verify and document that the mixture interpretation protocols developed from the completed validation studies generate reliable and consistent interpretations and conclusions for the types of mixed DNA samples typically encountered by the laboratory.

1.2 This standard applies to any type of DNA testing technology and methodology used, including but not limited to, STR testing, DNA sequencing, SNP testing, haplotype testing, traditional and rapid protocols, etc., where mixtures of DNA may be encountered, analyzed and interpreted.

Laboratories are advised to review their previous validation for compliance with this standard, supplement validation where necessary, and modify existing protocols accordingly.

12

12

ANSI/ASB 020 cont.

Internal Validation

4.2 The laboratory shall perform DNA mixture studies as part of the internal validation to support interpretation protocols prior to their use for casework samples in the laboratory. The mixture studies shall include, at a minimum, mixed DNA samples that:

4.2.1 Are representative of those typically encountered and interpreted by the testing laboratory.

4.2.2 Span the dynamic range of the detection platform.

4.2.3 Include each number of contributors to be interpreted by the laboratory.

4.2.4 Are constructed from extracted DNA samples of known origin (having known genotypes or sequences, etc.) combined: a) in varied input ratios based on the estimated DNA template amounts of the individual contributors; and b) with varied degrees of allele sharing.

13

13

ANSI/ASB 020 cont.

Verification

4.4 The laboratory shall verify and document that the mixture interpretation protocols developed from the validation studies generate reliable and consistent interpretations and conclusions for the types of mixed DNA samples typically encountered by the laboratory.

4.4.1 Verification of the mixture protocols shall be performed on mixed DNA samples of known origin that are different from those in the initial validation studies used to establish the protocol.

Verification of the mixture interpretation protocol shall demonstrate that its use results in the correct inclusion of true contributors, exclusion of non-contributors, and the parameters considered in the interpretation protocols.

NOTE Parameters may include, but are not limited to, assessment of the number of contributors, and evaluation of contributor ratios.

14

14

ANSI/ASB 40
Standard for Forensic DNA Interpretation and Comparison Protocols, First Edition 2019

1 Scope
This document provides requirements for a laboratory's DNA interpretation and comparison protocol. A protocol is needed for any DNA testing methodology that includes data interpretation and/or comparison. The protocol should encompass all variables permitted in the technical protocols that may have an impact on the data generated and the variety and range of test data anticipated in casework based on the types of samples routinely accepted and tested in the laboratory.

ANSI/ASB Standard 040, First Edition 2019

Standard for Forensic DNA Interpretation and Comparison Protocols

Added to Registry: May 12, 2020

15

15

ANSI/ASB 040 cont.

Interpretation

4.2 The laboratory shall maintain and follow documented DNA interpretation protocols that address the following.

4.2.1 Criteria for assessing the DNA data as originating from a single source or multiple sources.

4.2.2 Criteria upon which assumptions may be made and the types of assumptions that may be used in data interpretation including, but not limited to, the number of contributors and the presence of assumed contributors.

4.2.4 The limitations of the interpretation methods used such as characterizing and defining the maximum number of contributors, and issues associated with low-level data, low-level contributors and potential contamination events.

4.2.5 Criteria for defining what are interpretable data versus data that cannot be interpreted.

4.2.6 Criteria for defining data that are suitable for comparison versus data that are unsuitable for comparison.

16

16

ANSI/ASB 040 cont.

Conclusions

4.4 The laboratory shall maintain and follow documented protocols for drawing conclusions from the comparison of suitable evidentiary data derived from single source, mixed, and limited quality/quantity samples to reference (or other evidentiary) data.

4.4.1 Laboratory protocols shall describe the criteria used for concluding that the source of the reference data is included, excluded, or inconclusive when compared to evidentiary data when those terms are used by the laboratory. If a comparison is deemed inconclusive, the reason(s) shall be documented in the case record.

4.4.2 All re-evaluations of, and changes to, the original evidentiary data interpretation shall be thoroughly documented within the case record. The laboratory shall have protocols that address re-evaluation of evidentiary data after the comparison to reference (or other evidentiary) data has been performed.

17

17

ANSI/ASB 018
Standard for Validation of Probabilistic Genotyping Systems, First Edition 2020

1 Scope:
1.1 This standard sets forth the requirements to be used by laboratories for the validation of probabilistic genotyping systems related to interpreting autosomal STR results. Amelogenin is not covered by this standard.

1.2 Laboratories are advised to review their validation for compliance with this standard, supplement validation where necessary, and modify existing protocols accordingly.

ANSI/ASB Standard 018, First Edition 2020

Standard for Validation of Probabilistic Genotyping Systems

Added to Registry: May 4, 2021

18

18

ANSI/ASB 18 cont.

4 Requirements

NOTE Refer to Annex A, Requirements - Supporting Information, for additional information on the requirements in this section.

4.1 The laboratory shall validate a probabilistic genotyping system prior to its use for casework samples in the laboratory.

4.1.1 Validations shall include both developmental and internal studies. Developmental validation may be conducted by the manufacturer/developer of the application or another laboratory/agency. Developmental validation shall not replace internal validation.

4.1.2 Developmental validation studies shall address the following: accuracy, sensitivity, specificity, and precision. These studies shall include case-type profiles of known composition that represent (in terms of number of contributors, mixture ratios, and total DNA template quantities) the range of scenarios that would likely be encountered in casework. Studies shall not be limited to pristine DNA samples but shall also include compromised DNA samples (e.g., low template, degraded, and inhibited samples).

OSAC 19 NIST

19

ANSI/ASB 18 cont.

Internal Validation

4.1.3 Internal validation studies shall address the following: accuracy, sensitivity, specificity, and precision. These studies shall include internally generated case-type profiles of known composition that represent (in terms of number of contributors, mixture ratios, and total DNA template quantities) the range of actual casework samples intended for analysis with the system at the laboratory. Studies shall not be limited to pristine DNA samples but shall also include compromised DNA samples (e.g., low template, degraded, and inhibited samples). The internal validation shall not exceed the scope of the conditions tested in the developmental validation. Case type profiles that fall outside the range of conditions explored in developmental validation shall require additional developmental validation studies. See Annex A.

4.1.4 Internal validation studies shall include evaluating user input parameters that vary run to run. The effects of artifacts (e.g., stutter) and parameters that relate to the statistical algorithm (e.g., run time parameters for the software system that can vary from system to system) shall also be evaluated. The parameters may vary depending upon the approach or intended use of the software. Therefore, the specific parameters to be tested shall be determined by the laboratory.

4.1.5 Internal validation studies shall also include the evaluation of multiple propositions for case type samples to aid in the development of propositions. Such studies shall also consider the effect of overestimating and underestimating the number of contributors.

4.1.6 For internal validation, the laboratory shall evaluate both the appropriate sample types (i.e., number of contributors, mixture ratios, and template quantities) and the number of samples within each type to demonstrate the potential limitations and reliability of the software. The laboratory shall base this evaluation on the intended application of the software.

OSAC NIST

20

ANSI/ASB 18 cont.

Publication, Quality Assurance & Software Modifications

4.2 The underlying scientific principle(s) of the probabilistic genotyping model and associative method and software including the mathematical basis and underlying algorithms shall be published in peer-reviewed scientific journal(s).

4.3 Quality assurance parameters, analytical procedures, and interpretation protocols shall be derived from internal validation studies. Developmental and manufacturer recommendations may be used in addition to internal validation studies but shall not replace internal validation.

4.4 Software modifications, changes to computing platform or changes to upstream analytical processes (i.e., amplification processes, detection platforms) that may impact the interpretation or reported result(s) shall be evaluated to determine whether a validation or performance check is required prior to implementation. Such modifications shall require a validation or performance check of the affected software component. If neither is conducted after a software modification, changes to computing platform or changes to upstream analytical processes, the laboratory shall document the justification (e.g., software update simply enhances visual output or displays, therefore no performance check was conducted). See Annex A.

21

OSAC Proposed: OSAC 2020-N-0007 Best Practice Recommendations for the Management and Use of Quality Assurance DNA Elimination Databases in Forensic DNA Analysis

1 Scope:
This document provides best practice recommendations for the collection, storing, searching, and retention of DNA elimination samples and/or profiles in a quality assurance database. This document addresses the use of elimination databases as one component of a comprehensive approach to detect and monitor contamination.

OSAC Proposed Standard

OSAC 2020-N-0007
Best Practice Recommendations for the Management and Use of Quality Assurance DNA Elimination Databases in Forensic DNA Analysis

Added to Registry: April 6, 2021

OSAC 22 NIST

22

OSAC 2020-N-0007 cont.

Who Should Be In the Database?

4.3 An elimination database should be comprised of profiles from the following categories.

4.3.1 Individuals who have direct contact with the evidence, such as:
- a) Personnel in the forensic biology/DNA unit of the laboratory
- b) Laboratory personnel who may handle and/or examine evidence prior to the transfer of an item to the forensic biology/DNA unit
- c) Investigative or crime scene personnel who collect or handle evidence
- d) Medical examiner's office personnel, sexual assault nurses, and other hospital staff who may come in direct contact with and handle evidence
- e) Laboratory or investigating agency staff who may handle outer evidence packaging
- f) Laboratory custodial staff who may enter the forensic biology/DNA unit

4.3.2 Individuals with more limited contact with the evidence, such as:
- a) Laboratory staff regardless of work unit or access level to the forensic biology/DNA unit
- b) Investigative agency staff who may be present at the crime scene
- c) Visitors, maintenance staff, or vendor staff who may enter the forensic biology/DNA unit

4.3.3 Additional DNA profiles, such as:
- a) Profiles attributed to consumable manufacturing staff
- b) Unattributed contamination profiles
- c) Laboratory positive control profiles

NIST

23

OSAC Proposed: 2020-S-0004 Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events

1 Scope:
This standard provides requirements for the interpretation, comparison, and reporting of DNA data associated with control failures or contamination where re-testing is not performed. DNA data associated with a failed control or a contamination event may still be scientifically valid and may be relevant to an investigation.

2020-S-0004 Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events

Added to Registry: June 1, 2021

OSAC 24 NIST

24

OSAC 2020-S-0004 cont.

Documentation, Suitable & Compromised

4.2 The laboratory shall perform and document the assessment of the integrity of the associated DNA test results to determine the impact of the failed control or contamination. The assessment shall be based on scientifically valid principles in DNA analysis and include a determination of the possible cause and effect of the failed control or contamination, and an assessment of the risks associated with moving forward with data interpretation vs. those associated with re-testing.

4.2.1 If the DNA test results are determined to be suitable for interpretation within the constraints of the laboratory's internal validation studies and documented interpretation and comparison protocols and the laboratory does not retest, the laboratory shall perform and report the interpretation and comparison(s) with applicable statistical analysis.

4.2.2 If the DNA test results are determined to be compromised to the extent of being unsuitable for interpretation and retesting is not conducted, the results shall be reported as not suitable for interpretation according to laboratory policy.

NOTE If the DNA test results are determined to be compromised to the extent of being unsuitable for interpretation and retesting is conducted, it may be necessary to report results, interpretations, and comparisons from both the original and second tests.

25

Standards Implementation Options

- Implementing standards is NOT an all or nothing scenario.
- Different standards may exist on the Registry that address the same topic.
- Some labs may not implement a standard in its entirety – that's OK!

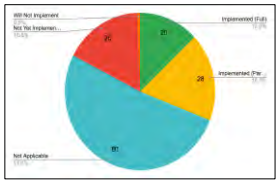
Labs should pick and choose to implement standards and/or portions of standards that work best for their situation.

See the "How-to Guide" on OSAC's Registry Implementation webpage!

26

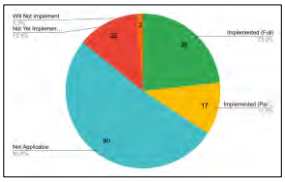
OSAC Implementation Survey Between June –August 2021: 155 Responses

ANSI/ASB 20 Validation & Verification



48 Implementors Full & Partial

ANSI/ASB 40 Interpretation & Comparison Protocols



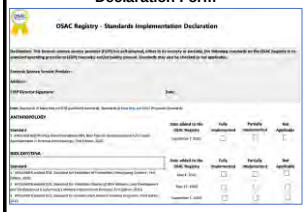
53 Implementors Full & Partial

80 Respondents Indicated Not Applicable

27

OSAC Registry Implementation

Declaration Form



Implementation Certificate

- **Download:** Detailed "**How To**" Guidelines with step-by-step instructions for labs
- https://www.nist.gov/osac/osac-registry-implementation

28

Stay Informed!

Website: www.nist.gov/osac

BULLETIN STANDARDS

- Provides monthly updates on forensic science standards moving through development process at SDOs and those moving through OSAC Registry process
- https://www.nist.gov/osac/osac-standards-bulletin

- Quarterly communication that provides updates on OSAC's program status, activities, accomplishments, and opportunities for public input with internal and external audiences.
- https://www.nist.gov/osac/osac-newsletter

- Follow us! https://www.linkedin.com/showcase/organization-of-scientific-area-committees-osac-for-forensic-science/

29

Questions/Discussion?

John Paul Jones II
OSAC Program Manager
Special Programs Office
National Institute of Standards and Technology
301-975-2782
john.jones@nist.gov

FOLLOW THE ORGANIZATION OF SCIENTIFIC AREA COMMITTEES FOR FORENSIC SCIENCE AT NIST

www.nist.gov/osac

30

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

DNA Process Map and Human Factors Working Group

National Institute of Standards and Technology (NIST)

Melissa K. Taylor
Special Programs Office

Module 5

1

Acknowledgments and Disclaimer

Points of view are the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities 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 National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

2

Agenda

- NIJ/NIST Expert Working Group on Human Factors in Forensic DNA Interpretation
  - Human Factors Overview
- Process Mapping- A Tool for Continuous Improvement
- Summary

3

NIJ/NIST Expert Working Group on Human Factors in Forensic DNA Interpretation

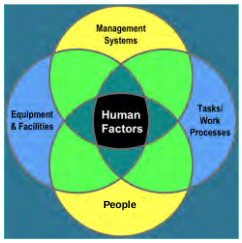
4

Working Group Charge

The Expert Working Group on Human Factors in Forensic DNA Interpretation is charged with **conducting a scientific assessment** on the effects of human factors in forensic DNA examination with the goal of **recommending approaches to improve its practice and reduce the likelihood of errors**. The Working Group will evaluate relevant bodies of scientific literature and technical knowledge to develop its recommendations and will publish a report of its findings.

5

Defining Human Factors



The scientific discipline concerned with the **understanding of interactions among humans and other elements of a system**

6

Traditional Error Prevention

- Make rules
- Enforce rules
- Punish violators
  - Fire them
  - Suspend them
  - Retrain them
  - Counsel them

If you follow the rules you cannot have an error

7

Lessons from Human Factors Research

- People are fallible and even the best make mistakes
- Error-likely situations are predictable, manageable, and preventable
- Drift happens!
- Fear of punishment for performance errors inhibits error reporting
- Error reporting is a critical aspect of a quality management system
- The Systems Approach offers a critical way to assess issues within the system and identify areas for improvement

8

What's in the System? Just remember P.E.A.R.

People **E**nvironment **A**ctions **R**esources

9

People

Physical
- Size
- Gender
- Age
- Strength
- Senses
- Perception

Physiological
- Health
- Nutrition
- Lifestyle
- Alertness/fatigue
- Chemical dependency

Psychological
- Experience
- Knowledge
- Training
- Attitude
- Emotional state

Psychosocial
- Interpersonal relations
- Ability to communicate
- Empathy
- Leadership

10

The Dirty Dozen

11

Environment

Physical
- Distractions/Interruptions
- Lighting
- Temperature extremes
- Location (in/out)
- Workspace
- Sound levels
- Housekeeping
- Safety issues

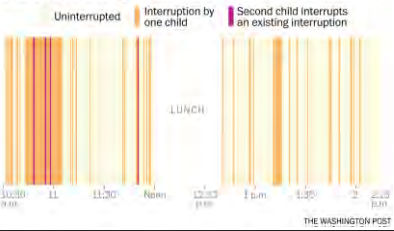
Socio-Technical
- Personnel
- Supervision
- Agency size
- Job security
- Morale
- Organizational culture
- Safety culture

12

Distractions/Interruptions

Balancing work and parenting in the pandemic

One day, two parents documented how often their two children interrupted a portion of the workday.



13

13

Actions

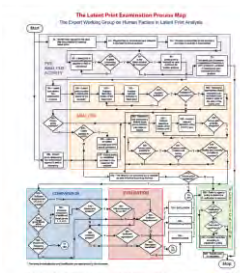
- **Understanding the objectives of the job**
- Sequence of actions
- Need for redundancy
- **Communication requirements**
- **Information requirements**
- **Review requirements**
- Certification requirements

14

14

Process Mapping

- Identify the steps required to complete a task
- For each step, identify the information, tools, communication links, procedures necessary to complete the step



15

15

Resources

- Clear technical documentation
- Appropriate equipment/tools
- Materials
- Enough time
- Enough people
- Continuous training
- Compensation

16

16

Expert Working Group Members - DNA

- ***Adele Quigley-McBride,*** Postdoctoral Associate, Duke University School of Law
- ***Alex Chaparro,*** Professor, Embry-Riddle Aeronautical University
- ***Angie Spessard,*** Forensic Scientist, Maryland State Police, Forensic Sciences Division
- *Ashley Hinkle,* Forensic Biologist, Georgia Bureau of Investigation
- ***Bas Kokshoorn,*** Principal Scientist, Netherlands Forensic Institute
- ***Brandon Garrett,*** Professor, Director, Wilson Center for Science and Justice, Duke University School of Law
- ***Brenda Danosky,*** Biology Program Manager, Illinois State Police
- ***Britton Morin,*** Laboratory Director, Union County Prosecutor's Office Forensic Laboratory
- ***Catherine Grgicak,*** Associate Professor, Chemistry, Rutgers University,
- ***Clinton Hughes,*** Forensic DNA Attorney, Brooklyn Defender Services
- ***Craig O'Connor,*** Assistant Director / Technical Leader, New York City Office of Chief Medical Examiner
- ***David Kaye,*** Distinguished Professor Emeritus, Pennsylvania State University,
- ***Erica Romsos,*** Research Biologist, NIST
- *Gabriel Lopez,* Assistant Chief, Phoenix Police Department
- ***Glenn Langenburg,*** Forensic Scientist, Elite Forensic Services
- ***Jarrah Kennedy,*** Senior DNA Criminalist / Forensic Specialist, Kansas City Police Crime Laboratory
- ***Jocelyn Carlson,*** DNA Quality Assurance Program Manager, FBI Laboratory
- ***Jody Wolf,*** Crime Lab Administrator, Phoenix Police Department
- ***Kaye Ballantyne,*** Chief Scientist, Victoria Police Forensic Services Department
- ***Kayleigh Matook,*** Forensic Scientist, Colorado Bureau of Investigation
- ***Kristy Martire,*** Associate Professor, University of New South Wales Sydney
- ***Lynn Garcia,*** General Counsel, Texas Forensic Science Commission,

17

Expert Working Group Members - DNA

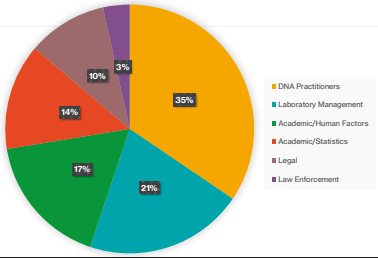
- ***Mandi Van Buren,*** Criminalist, Kern Regional Crime Laboratory
- ***Matthew Farr,*** Section Chief, Forensic Biology, Bureau of Alcohol, Tobacco, Firearms and Explosives
- ***Michael Ambrosino,*** Special Counsel for DNA & Forensics, U.S. Attorney's Office
- ***Michelle Madrid,*** Senior Criminalist, Los Angeles County Sheriff's Department
- ***Sandy Zabell,*** Professor, Northwestern University
- ***Tacha Hicks,*** Specialist in Interpretation, School of Criminal Justice & University Center of Legal Medicine, Lausanne – Geneva
- ***Tiffany Roy,*** Forensic DNA Expert, ForensicAid, LLC
- ***Tim Scanlan,*** Commander, Technical Services Bureau, Jefferson Parish Sheriff's Office
- ***Tom Busey,*** Professor, Indiana University

Steering Committee

- ***Melissa Taylor,*** Program Manager, NIST
- ***Niki Osborne,*** Forensic Research Scientist, NIST (Contractor)
- ***Donia Slack,*** Director, Forensic Sciences, RTI International
- ***Sarah Norsworthy,*** Senior Project Management Specialist, RTI International
- ***Mikalae Martin,*** Forensic Scientist, RTI International

18

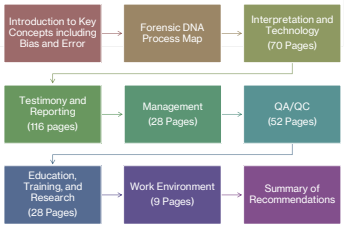
Member Composition (total = 29)



19

Report outline with current page count

(total 303 pages)



20

DNA Technical Leaders Survey

- Comprehensive survey of topics related to human factors in forensic DNA interpretation
- To be completed by the DNA technical leader (or equivalent)
- Approximately 85 questions
- Estimated 45 minutes to complete

21

Sample Survey Questions

- Who are your **primary customers**? (LE, Attorneys, Private clients, ...)
- What **forensic DNA services** are you providing to your primary customer(s)? (autosomal, Mito, Y-STR, ...)
- Does your laboratory **track the TYPE of DNA samples** that you routinely analyze? (e.g., track whether a sample is liquid blood, saliva stains, dried semen stains, touch DNA, ...)
- How often does your laboratory **perform the following tasks**? (presumptive tests for semen, blood, or saliva; microscopic search for sperm; confirmatory test for blood or saliva, ...)

22

Sample Survey Questions: Error Language

- Which of the following terms does your laboratory regularly use as part of your quality management system? (Select all that apply)

☐ Analyst Error	☐ Conflict	☐ Deviation from protocol	☐ Disagreement
☐ Error	☐ Incident	☐ Instrument Error	☐ Lapse
☐ Mistake	☐ Non-conformity	☐ Quality Issue	☐ Slip
☐ Systematic Error	☐ Technological Error	☐ Unexpected finding	☐ Other

- How does your laboratory define "error", "disagreement", "conflict", or any other related terms that it regularly uses? Please include the term(s) and definition(s) here or write N/A.

23

Sample Survey Questions: Thoughts on Bias

Please indicate your level of agreement with the following statements: Strongly disagree, ... Strongly agree

- Cognitive bias is a bigger issue for analysts in other forensic disciplines than those in forensic biology.
- Knowing about the reference profile before examining a complex DNA mixture can affect how a DNA analyst interprets the mixture.
- Knowing about a confession before examining a complex DNA mixture can affect how a DNA analyst interprets the mixture.
- Knowing one DNA analyst's Number-of-Contributor determination can affect another DNA analyst's Number-of-Contributor determination.

24

Sample Survey Questions: Thoughts sharing validation data

- An idea being discussed in the DNA community is to create a central repository of validation summaries that multiple laboratories could contribute data to and use. This repository could be accessible to all stakeholders/interested parties (including attorneys and researchers), or it could be password-protected and only available to other DNA laboratories (i.e., private). Please read the following statements and select the one that best applies to your laboratory:
- o Our laboratory would use a central repository, regardless of who can access it.
- o Our laboratory would only use a central repository if it was private.
- o I do not know if our laboratory would use a central repository.
- o Our laboratory would not use a central repository.
- o Validation summaries are not applicable to our laboratory.

25

Sample Survey Questions: Needs

- Some laboratories use internally-collected DNA samples for their validation studies (e.g., from staff members). Collecting samples in this way may restrict sharing data outside of the laboratory due to privacy concerns. Would your laboratory **benefit from access to appropriately consented, externally-collected DNA samples to use in your validation studies**?
- Has your laboratory encountered any **barriers to creating complex DNA mixture samples** for your internal validation exercises? Please discuss or type "not applicable".

26

Agenda

- NIJ/NIST Expert Working Group on Human Factors in Forensic DNA Interpretation
  - Human Factors Overview
- Process Mapping- A Tool for Continuous Improvement
- Summary

27

Process Defined

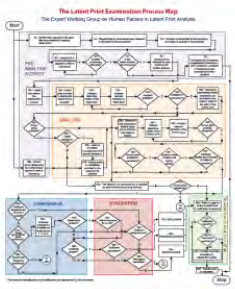
proc·ess | /ˈprä,ses, ˈprō,ses/

a series of actions or steps taken which transforms inputs into outputs of value to a customer (internal or external).

28

Why is it Important to Map the Current Process?

- Shows others **how a process is done**
- Helps users to analyze **how the process could be improved**
- **Improves communication** between individuals engaged in the same process



29

How Can it Help?

Once everyone has a shared understanding of the current state, we can now begin asking questions like:

- Why are we doing it this way? Is this the best way?
- What are we seeking to accomplish with a specific step?
- Are we getting the correct input to make this decision here?
- Is there peer-reviewed research that supports or contradicts this step?
- What are the potential risks/adverse consequences?
- What is training is required to be considered a qualified user of the technique/procedure?

30

How we do it: Constructing a Process Map

Step 1: Determine the Boundaries - Determine the start and stop points to your flow of process steps

Step 2: List and Sequence the Steps - Write down the process steps as they exist now.

- Use post-it notes for each process step
- If there are feedback arrows, make sure feedback loop is closed

Step 3: Check for Completeness (internal)

- "Walk the process", repeatedly
- Analyze/review from finish to start

Step 4: Finalize the Map (external)

- Are people following the process as charted?
- Did we miss anything?

31

DNA ANALYSIS PROCESS MAPPING PARTICIPANTS

- **Beth Ordeman**, Pinellas County Forensic Laboratory
- **Carl Sobieralski,** Indiana State Police Laboratory Division
- **Jason Befus**, Maryland State Police-Forensic Sciences Division, Biology Section
- **Eugene Lien**, NYC Office of Chief Medical Examiner
- **Ann Marie Gross,** Minnesota Bureau of Criminal Apprehension
- **Melissa Suddeth**, Florida Department of Law Enforcement
- **Amber Carr**, FBI
- **Jeanette Walin,** California Department of Justice
- **Jarrah Kennedy,** Kansas City Police Crime Laboratory

Process Map Team
Melissa Taylor – Facilitator
Heather Waltke - Visio
Sara Bitner – Notetaker/Visio
Blythe Toma - Visio
Niki Osborne - Assistant

32

33

How we do it: Constructing a Process Map

- There are **NO** right or wrong steps in the map.
- Everyone should see their process in the map.
- If someone does it, it gets mapped.

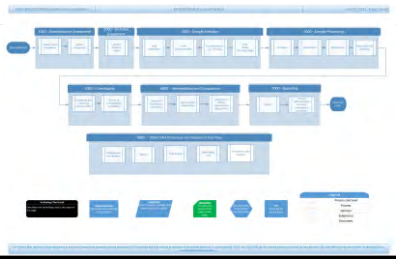
34

How we do it: Map Symbols

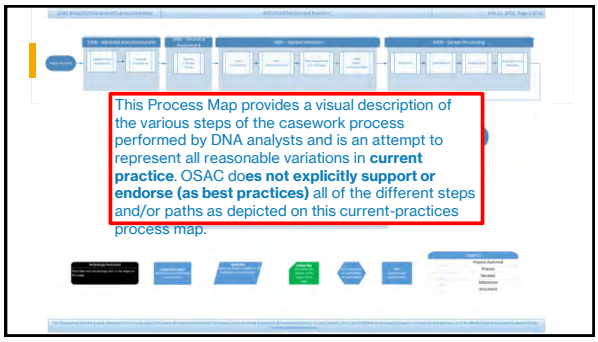
Symbol	Name	Description
	Process	The most frequently used flowchart shape shows an action, task or operation that needs to be done
	Subroutine	Shows a multistep action that may be predefined in a standard, by lab policy, and/or by examiners; it could also mean that there is already a flowchart that can be used as a reference
	Document	Indicates a process step that generates documentation
	Decision	The point at which a decision needs to be made; the arrows flowing from the decision shape will be labeled with yes or no
	Arrow	The arrows indicate the direction in which the flowchart should be read
	Connector	In order to connect to different page or section of the chart, and you can't draw a line
	Input/Output	Summarizes the material or information entering or leaving the process
	Terminator	Represents the entry and exit points of your flowchart

35

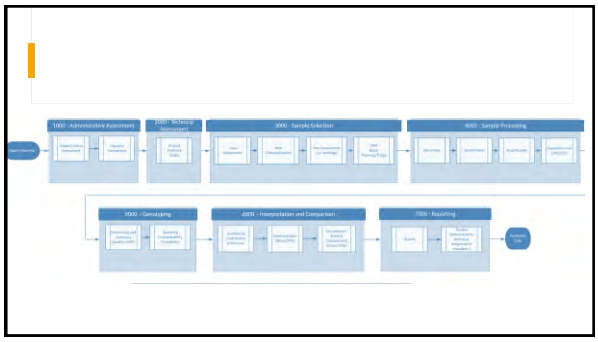
DNA Analysis Process Map Preview



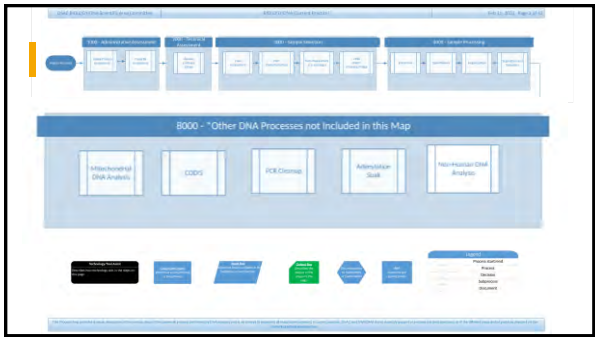
36



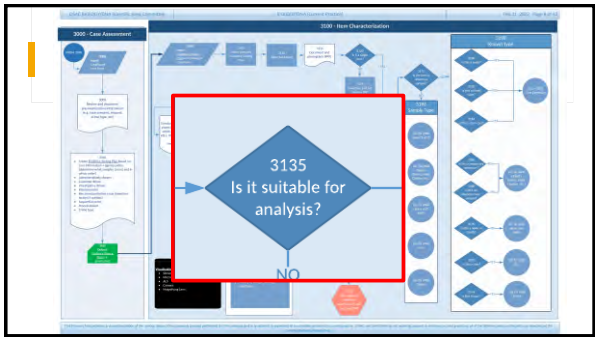
37



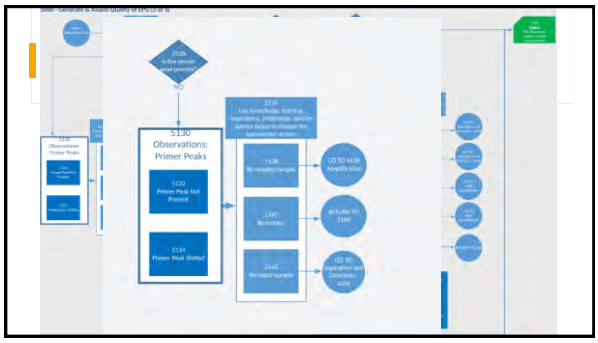
38



39



40



41

DNA Analysis Process Map

OSAC Human Forensic Biology Subcommittee
https://www.nist.gov/osac/human-forensic-biology-subcommittee

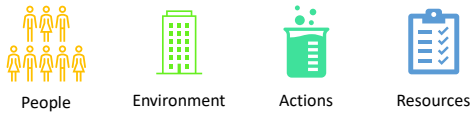
42

Agenda

- NIJ/NIST Expert Working Group on Human Factors in Forensic DNA Interpretation
  - Human Factors Overview
- Process Mapping- A Tool for Continuous Improvement
- Summary

43

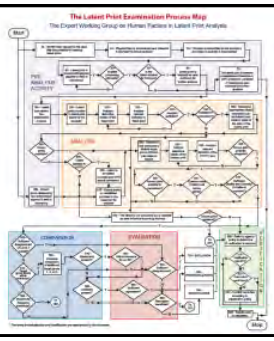
P.E.A.R.



44

The Dirty Dozen

45



Process Mapping-

A Tool for Continuous Improvement

46

Melissa Taylor
Senior Forensic Science Research Manager
NIST Forensic Science Program
301.975.6363
melissa.taylor@nist.gov

www.nist.gov/forensics/

47

LUNCH BREAK

60 minutes

48

DNA Technical Leader Survey

Thank you for your interest in the DNA Technical Leader Survey.

The National Institute of Standards and Technology (NIST), in collaboration with the National Institute of Justice (NIJ), is in the process of developing the Human Factors in DNA Interpretation report which will be published as the third instalment in the Human Factors in Forensic Sciences Expert Working Group Series. The DNA Technical Leader Survey is an important tool to obtain information regarding the current protocols and practices within DNA laboratories from a variety of forensic science service provider types including publicly funded local, county, state, and federal laboratories along with private practitioners and consultant groups or individuals within the U.S. and abroad.

This survey is to be completed by the **DNA Technical Leader (or equivalent).** This is the individual who is responsible for the technical oversight of the DNA laboratory, which may include (but is not limited to) day-to-day quality assurance and accreditation compliance, design and implementation of methods development, verification of analytical instrumentation function, and validation of new technologies.

OMB Statement
OMB Control #0693-0043
Expiration Date: 03/31/2022
NIST Generic Clearance for Usability Data Collections

A Federal agency may not conduct or sponsor, and a person is not required to respond to, nor shall a person be subject to a penalty for failure to comply with an information collection subject to the requirements of the Paperwork Reduction Act of 1995 unless the information collection has a currently valid OMB Control Number. The approved OMB Control Number for this information collection is 0693-0043. Without this approval, we could not conduct this survey/information collection. Public reporting for this information collection is estimated to be approximately 45 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the information collection. All responses to this information collection are voluntary. Send comments regarding this burden estimate or any other aspect of this information collection, including suggestions for reducing this burden to the National Institute of Standards and Technology (NIST) point of contact: Melissa Taylor, melissa.taylor@nist.gov.

Frequently Asked Questions about the DNA Technical Leader Survey

Who should complete the survey?

The DNA Technical Leader Survey is intended to be completed by one person in each participating laboratory – the DNA Technical Leader (or equivalent). This is the individual who is responsible for the technical oversight of the DNA laboratory, which may include (but is not limited to) day-to-day quality assurance and accreditation compliance, design and implementation of methods development, verification of analytical instrumentation function, and validation of new technologies.

What is the purpose of the survey?

The Forensic DNA Technical Leader Survey has been designed to assess consistency and variability between forensic DNA laboratories with respect to laboratory management, tasks performed, DNA data interpretation, cognitive bias, internal and external training and research, testimony and reporting practices, quality assurance and quality control measures, and stakeholder engagement opportunities. This survey will provide insight into where standardization of DNA practices is being utilized and the role of technology in forensic DNA interpretation. The results of this survey will inform standards and best practice recommendations for the discipline, aid in the identification of research needs, and assist NIST in its mission to support the forensic science community.

What will you do with the results of the survey?

Currently, there are few sources of information in existence focusing on the influence of human factors within the discipline of forensic DNA. This survey will serve as a starting point for gathering such data. Further, the resulting data obtained through this survey will be incorporated into the report produced by the NIST/NIJ Expert Working Group on Human Factors in Forensic DNA Interpretation to create recommendations for all activities related to, and impacted by, DNA interpretation.

SAMPLE SURVEY QUESTIONS

1. What type of crime laboratory or forensic science service provider (FSSP) do you represent?

o Publicly funded local crime laboratory (to include city or town)

o Publicly-funded county crime laboratory

o Publicly funded state crime laboratory

o Publicly funded federal crime laboratory

o Private laboratory

o Consultant

o Other (please specify)

2. What Forensic DNA services are you providing to your primary customer(s)? (Select all that apply)

- o Autosomal STR
- o Mitochondrial
- o Y-STR
- o Next Generation Sequencing
- o Mixture Interpretation
- o Probabilistic Genotyping
- o CODIS upload and search
- o Familial Searching
- o Forensic Genetic Genealogy
- o Paternity/parentage (criminal)
- o Paternity/parentage (non-criminal)
- o Phenotyping
- o Other (please specify) _____

3. What are the categories that your laboratory uses to track DNA samples? (Select all thatapply)

- o Bodily fluid type
- o Case scenario
- o Crime type
- o Number of contributors
- o Template amount
- o Evidence item type (e.g., gun, clothing)
- o Other (please list) _____
- o Not applicable

4. How do you monitor DNA analysts' abilities to perform complex tasks (excluding routineopen proficiency testing), and how often?

	Monthly	Quarterly	Biannually	Yearly	Biennially	When required	Never	Not sure
In-house testing/research	o	o	o	o	o	o	o	
Internal collaborative exercises	o	o	o	o	o	o	o	
Inter-laboratory exchange	o	o	o	o	o	o	o	
Training exercises	o	o	o	o	o	o	o	
Blind proficiency tests	o	o	o	o	o	o	o	
Other (please specify or select "never")	o	o	o	o	o	o	o	

5. How are your laboratory's reports formatted?

- o Narrative (written explanations or paragraphs that describe evidence/items tested andthe DNA results and opinions)
- o Tabular (lists and tables of the evidence/items tested and the DNA results and opinions)
- o Combination
- o Not sure

6. Is your DNA laboratory reporting a quantitative value only or a combination of quantitativeand qualitative statements?

- o Quantitative only (Likelihood Ratio or other numerical value)
- o Qualitative only (verbal equivalent or written explanation)
- o Quantitative and qualitative
- o Not sure

7. Does your laboratory have a procedure to monitor testimony?

- o Yes

- o No
- o Not sure
- o Not applicable

8. If any results or opinions are changed as a result of the review processes, how are thedisagreement/non-consensus and action documented? (Select all that apply)

- o Report
- o Case file
- o Personnel file
- o Not documented
- o Other (please specify) _____
- o Review process would not change results or opinions

9. Does your agency rely on external grants to provide DNA analysts training from outside your organization?

o Yes

o No

o Not sure

---

10. Some laboratories use internally-collected DNA samples for their validation studies (e.g., from staff members). Collecting samples in this way may restrict sharing data outside of the laboratory due to privacy concerns.

Would your laboratory benefit from access to appropriately consented, externally-collected DNAsamples to use in your validation studies?

- o We already obtain external DNA samples
- o We do not currently obtain external DNA samples but would benefit from such samples
- o No, we would not benefit
- o Not sure
- o Not applicable

11. Has your laboratory encountered any barriers to creating complex DNA mixture samplesfor your internal validation exercises? Please discuss or type "not applicable".

_____

_____

_____

_____

_____

12. How long does it usually take a DNA analyst to complete their training at your agency?

- o0-3 months
- o4-6 months
- o7-9 months
- o10-12 months
- o>12 months
- oNot sure

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

DNA Sequencing Research Overview

National Institute of Standards and Technology (NIST)
Peter M. Vallone
Applied Genetics

Module 6

1

Acknowledgments and Disclaimer

Points of view are the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities 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 National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

2

Outline

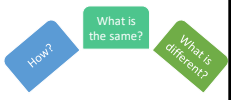
- Forensic applications of sequencing (general)
- Various sequencing platforms and kits
- Examples of sequencing research in the Applied Genetics Group

3

Forensic Applications of Sequencing

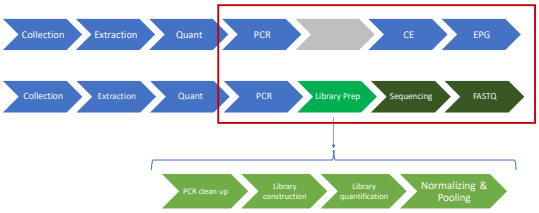
- There is an interest in sequencing for forensic analyses
  - More markers/marker types – higher multiplexing capability than CE
  - More information → sequence level resolution for STRs

    19 allele -> [GGAA]11 [GGCA]8 **or** [GGAA]12 [GGCA]7 **or** [GGAA]13 [GGCA]6
  - **The promise:** access to this additional information will support forensic casework applications
- Differs from the traditional PCR-CE workflow



4

Comparing workflows – targeted sequencing



5

Sequencing STRs

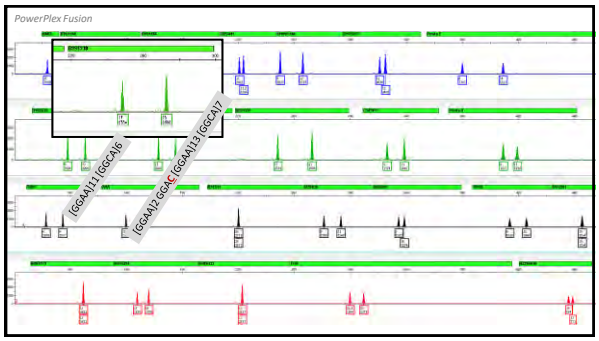
Targeted sequencing of STRs
- STR motif sequence variation; flanking region variation
- Further understand simple versus complex repeat motifs
- Characterize stutter, 'noise'

Applications
- One-to-one matching?
- With the new U.S. core loci we are already quite low (>10^{-20})
- Partial profiles
- Kinship

Greater degree of Multiplexing
- Not confined by dye colors; smaller PCR amplicons (for degraded samples)
- PCR for sample enrichment
- ***Still using PCR*** – stochastic effects, stutter

Mixtures
- Resolve alleles identical by length, but differ by sequence
- Resolve stutter from low-level contributors (based on sequence)
- A sequenced allele *may* have a lower frequency (lower RMP or higher LR)

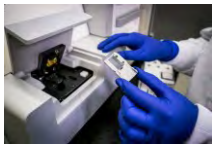
6

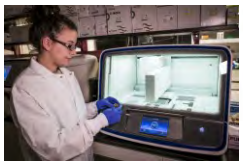


7

Sequencing platforms in our lab

- Illumina – MiSeq FGx
- Thermo Fisher S5xl and Ion Chef





8

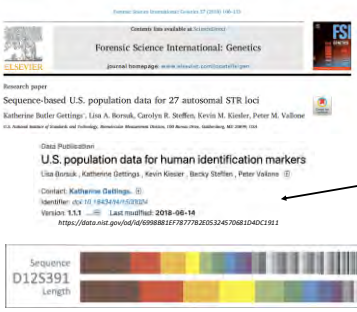
Select listing of commercial sequencing workflows

Assay	Platform	Associated Software	Markers
ForenSeq DNA Signature Prep Kit	MiSeq FGx	UAS	auSTRs, Y STRs, X STRs and SNPs
ForenSeq MainstAY	MiSeq FGx	UAS	auSTRs, Y STRs
ForenSeq mtDNA Control Region Kit (and whole mtGenome)	MiSeq FGx	UAS	Mitochondrial control region (WG soon?)
ForenSeq Kintelligence	MiSeq FGx	UAS/GEDmatch	10,230 SNPs
PowerSeq 46GY System	MiSeq	Open	auSTR and Y STRs
PowerSeq CRM Nested System, Custom	MiSeq	Open	Mitochondrial control region (and WG)
Precision ID SNP Identity Panel	S5	Converge	Identity SNPs
Precision ID SNP Ancestry Panel	S5	Converge	Ancestry SNPs
Precision ID STR GlobalFiler NGS STR Panel v2	S5	Converge	Autosomal STRs
Precision ID mtDNA Whole Genome Panel	S5	Converge	Whole mitochondrial genome
Precision ID mtDNA Control Region Panel	S5	Converge	Mitochondrial control region
Ion AmpliSeq SNP Phenotype Panel	S5	Converge	SNPs
GeneReader DNAseq Targeted Panels V2	Illumina/S5	CLCBio - open	Mito, SNPs

9

Selected sequencing projects

10



Sequencing of 1036 NIST Population Samples

- Platform Illumina/Verogen FGx
- Verogen ForenSeq Kit (Multiplex B)
- Detailed Materials and Methods describing sequencing and data analysis
- Reporting sequenced-based allele frequencies – 27 auSTR loci

89 unique alleles
24 unique alleles

11

Sequencing Forensic STRs in Population Samples

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

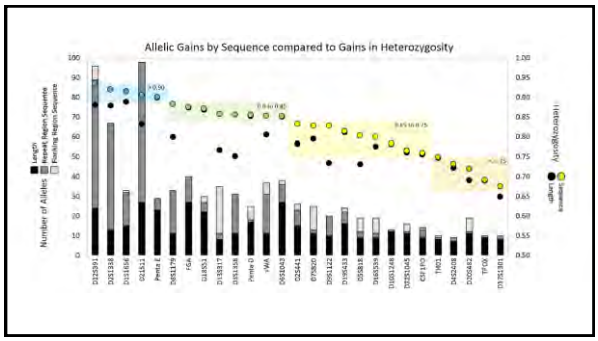
Example of length versus sequence-based frequency:

D4S2408

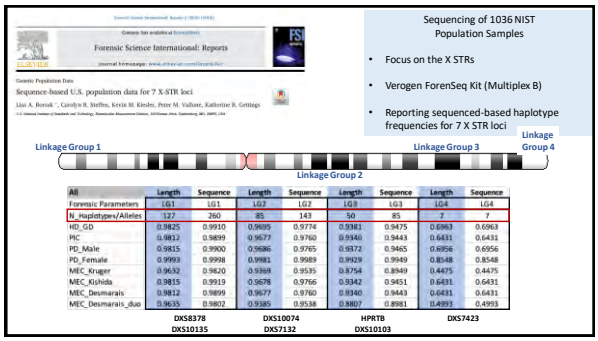
Allele	N	Freq	Sequence Allele	N	Freq
7	1	0.6%	[ATCT]7	1	0.6%
8	23	14.4%	[ATCT]8	23	14.4%
9	60	37.5%	[ATCT]9	18	11.3%
			[ATCT] GTCT [ATCT]7	42	26.3%

Length	Sequence
8,9	[ATCT]8, [ATCT]9
2pq	2pq
2*0.144*0.375	2*0.144*0.113
0.108	0.033
1 in 9.3	1 in 30.7

12

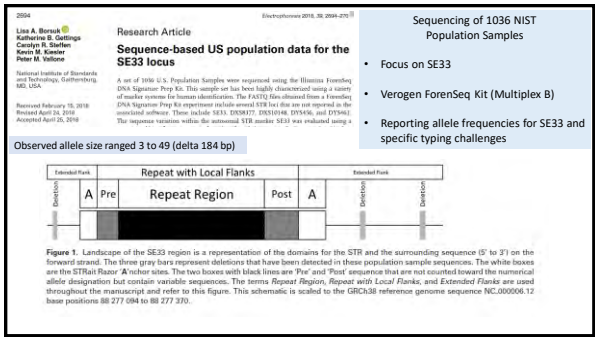


13

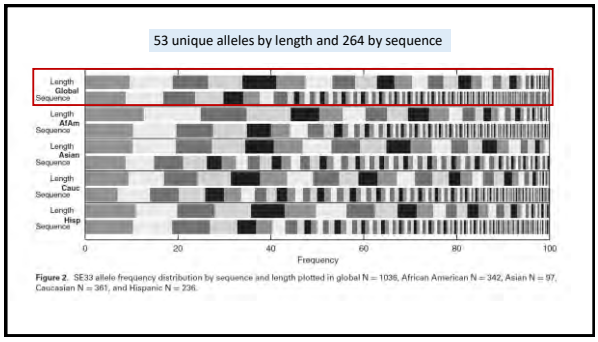


All	Length	Sequence	Length	Sequence	Length	Sequence	Length	Sequence
Forensic Parameters	LG1	LG1	LG2	LG2	LG3	LG3	LG4	LG4
N_Haplotypes/Alleles	127	260	85	143	50	85	7	7
HD_GD	0.9825	0.9910	0.9605	0.9774	0.9361	0.9475	0.6963	0.6963
PIC	0.9812	0.9899	0.9677	0.9760	0.9340	0.9443	0.6431	0.6431
PD_Male	0.9815	0.9900	0.9686	0.9765	0.9372	0.9465	0.6956	0.6956
PD_Female	0.9993	0.9998	0.9981	0.9989	0.9929	0.9949	0.8548	0.8548
MEC_Kruger	0.9632	0.9820	0.9369	0.9535	0.8764	0.8949	0.4475	0.4475
MEC_Kishida	0.9815	0.9919	0.9678	0.9766	0.9342	0.9451	0.6431	0.6431
MEC_Desmarais	0.9812	0.9899	0.9677	0.9760	0.9340	0.9443	0.6431	0.6431
MEC_Desmarais_duo	0.9635	0.9802	0.9385	0.9538	0.8807	0.8981	0.4993	0.4993

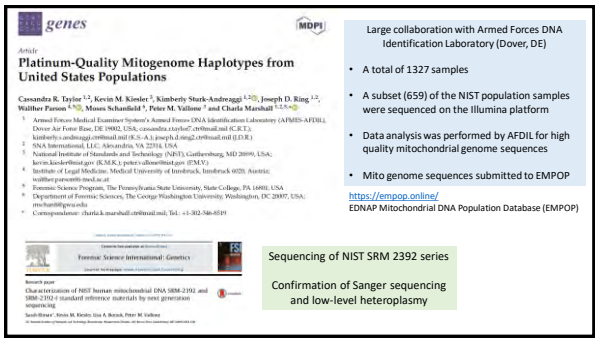
16



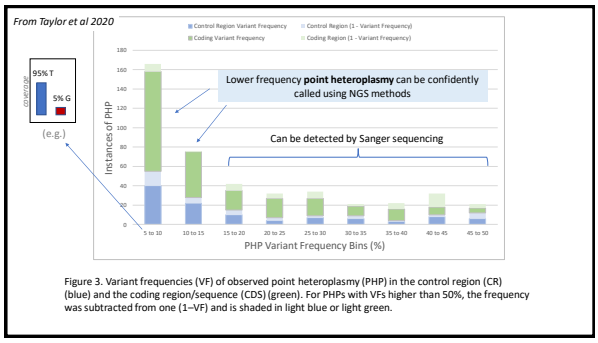
20



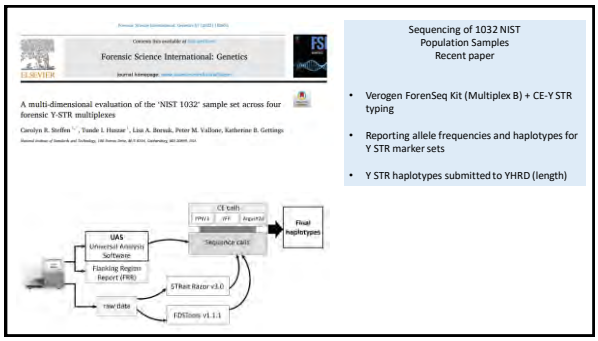
21



23



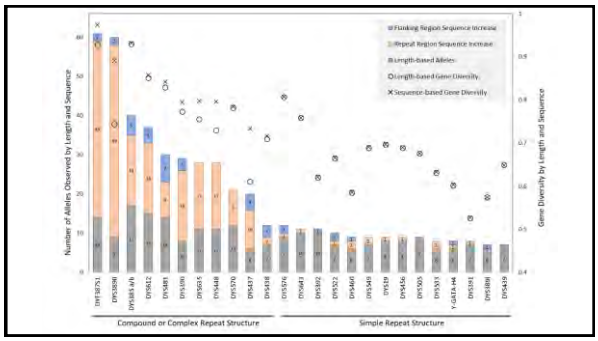
24



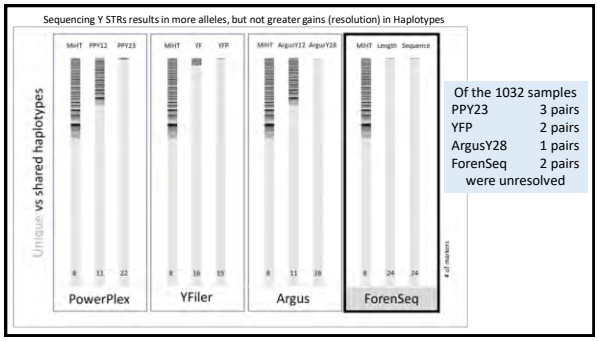
Sequencing of 1032 NIST Population Samples
Recent paper

- Verogen ForenSeq Kit (Multiplex B) + CE-Y STR typing
- Reporting allele frequencies and haplotypes for Y STR marker sets
- Y STR haplotypes submitted to YHRD (length)

26

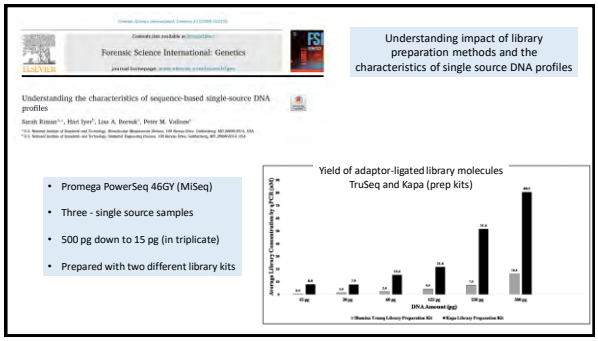


27



Of the 1032 samples
PPY23 3 pairs
YFP 2 pairs
ArgusY28 1 pairs
ForenSeq 2 pairs
were unresolved

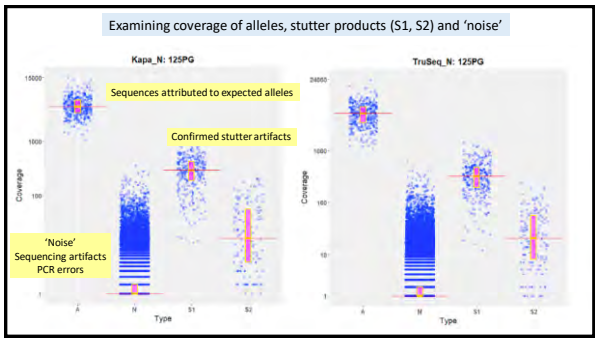
29



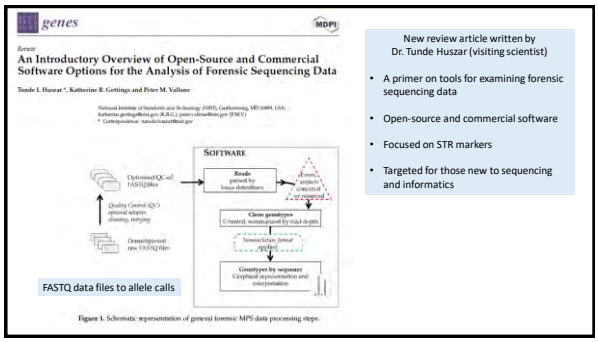
Understanding impact of library preparation methods and the characteristics of single source DNA profiles

- Promega PowerSeq 46GY (MiSeq)
- Three - single source samples
- 500 pg down to 15 pg (in triplicate)
- Prepared with two different library kits

30



31



New review article written by
Dr. Tunde Huszar (visiting scientist)

- A primer on tools for examining forensic sequencing data
- Open-source and commercial software
- Focused on STR markers
- Targeted for those new to sequencing and informatics

36

Table 1. Summary of characteristics of software for the interpretation of MPS data of forensic markers.

Software	Versions	Author/Vendor	Year	Accessibility	Runs on	Locus Definition	Landmarks for Loci
STRait Razor	v1.0	Warshauer et al. [11]	2015	free	Unix/Linux		
	v2.0	Warshauer et al. [12]	2015	free	Unix/Linux		
	v2x	King et al. [14]	2017	free	Unix/Linux	config file	'anchor'
	v3.0	Woerner et al. [14]	2017	free	all platforms		
	Online	King et al. [13]	2021	free	online/ all platforms		
FDSTools	FMSV	Anvar et al. [16]	2014	free	Unix/Linux		
	v1.0	van der Gaag et al. [17]	2016	free	Unix/Linux	library file	
	v1.1.1	Hoogenboom et al. [18]	2017	free	Unix/Linux	STRNaming	'flank'
	v2.0	Hoogenboom et al. [19]	2021	free	all platforms		
STRinNGS	v1.0	Fein et al. [20]	2016	on request	Unix/Linux	configuration file	'flanking sequences'
	v2.0	Jonck et al. [21]	2020	free	Unix/Linux		
MyFLq	v1.1	Van Neste et al. [22,23]	2014	free	online/Unix/Linuxpanels		'recognition elements'
toaSTR	v1.0	Gamchow et al. [24]	2018	free	online	allele database	'primer'
Altius	Cloud	Bailey et al. [25]	2017	on request	online	lookup table	'target regions'
DuacfiD	v2.0	Battelle [26]	2015	commercial	Windows	config file	default
GeneMarker HTS	v1.0	SoftGenetics [27]	2017	commercial	Windows	default	default
MixtureAce	v1.0	NicheVision [28]	2018	commercial	Windows	default	default
CLC Genomics Workbench	AQME	Sturk-Andreaggi et al. [29]	2017	commercial	all platforms	non STR	non STR
Universal Analysis Software	v2.5	Verogen [30]	2021	commercial	Windows	default	default
Converge Forensic Analysis Software	v2.2	Thermo Fisher [31]	2019	commercial	Windows	BED files	default

- Table consisting of software packages, references, platforms
- Each is described in more detail in the paper

37

WASHINGTON, DC
ISFG 2022
INTERNATIONAL SOCIETY FOR FORENSIC GENETICS

(August 29 – September 2, 2022)

NGS WORKFLOWS FOR FORENSIC GENETICS (HALF DAY)

Presenters: Peter Vallone and Kim Andreaggi (AFDIL)
Date and Time: Monday August 29th, 9 AM – 1 PM

This workshop aims to review and explore the details of various NGS/MPS sequencing methods. Common sequencing methods, kits, and platforms that may be applied to forensic genetic analyses will be discussed. The laboratory workflow steps involved in sequencing library preparation and their specific purposes will be presented. These include: targeted PCR, incorporation of sequencing adapters and unique indices, quantification/normalization of final products prior to sequencing. Examples of the process will be illustrated through forensically-relevant workflows for the sequencing of STRs, SNPs and the mitochondrial genome. Brief examinations of the resulting sequence data will be demonstrated by the instructor using open source and/or commercial software tools. This workshop is intended for attendees with some basic familiarity with sequencing methods and interested in the basic and practical aspects of carrying out sequencing experiments in support of adopting these methods of genetic analysis in their laboratory. Questions related to the scope of the workshop can be directed to the instructor. [peter.vallone@nist.gov]

38

Thanks for your attention

Questions?

peter.vallone@nist.gov

NIST APPLIED GENETICS

39

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

STR Sequence Nomenclature Activities

National Institute of Standards and Technology (NIST)

Katherine Gettings
Applied Genetics Group

Module 7

1

Prelude

GENOMIC CHARACTERIZATION OF STRs

2

STR Distribution Across Human Genome

Period Length Thresholds

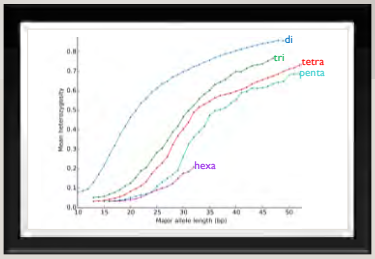
Di-	11bp	≥	[AT]6
Tri-	14bp	≥	[AAT]5
Tetra-	14bp	≥	[AAAT]4
Penta-	16bp	≥	[AAAAT]4
Hexa-	17bp	≥	[AAAAAT]3

Total = 689,512 STRs

adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.

3

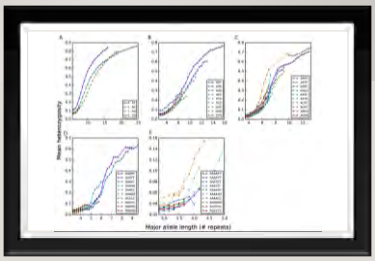
Heterozygosity vs Major Allele Length by Period



adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.

4

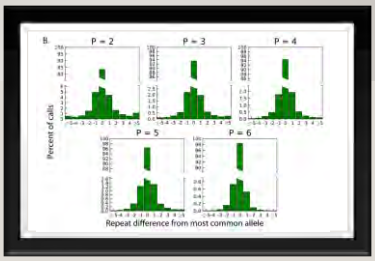
Heterozygosity vs Allele Length by Period by Motif



adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.

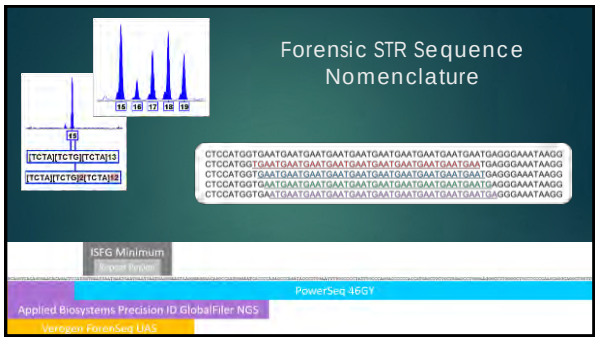
5

Repeat Difference from Most Common Allele by Period

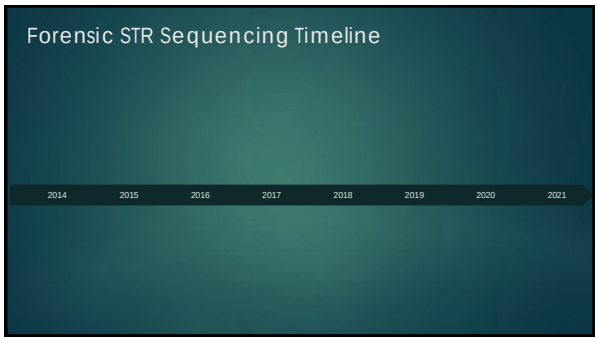


adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.

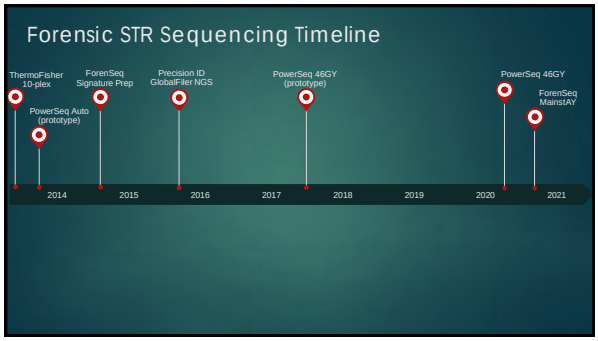
6


Forensic STR Sequence Nomenclature
15
16
17
18
19
[TCTA][TCTG][TCTA]13
[TCTA][TCTG]2[TCTA]12
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG
ISFG Minimum
Repeat Region
PowerSeq 46GY
Applied Biosystems Precision ID GlobalFiler NGS
Verogen ForenSeq UAS

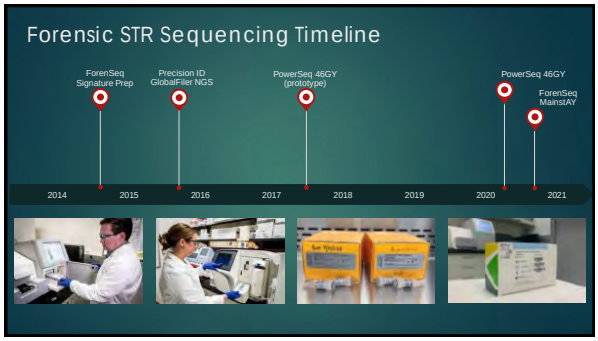
7


Forensic STR Sequencing Timeline
2014
2015
2016
2017
2018
2019
2020
2021

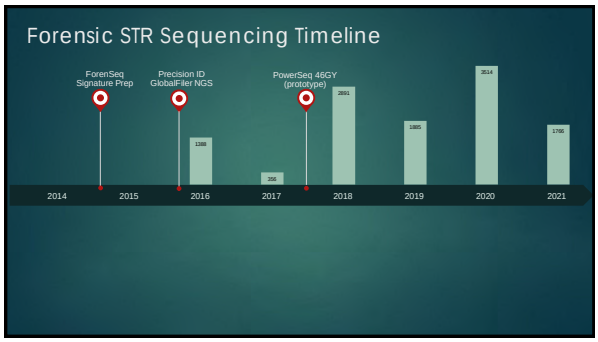
8


Forensic STR Sequencing Timeline
ThermoFisher 10-plex
PowerSeq Auto (prototype)
ForenSeq Signature Prep
Precision ID GlobalFiler NGS
PowerSeq 46GY (prototype)
PowerSeq 46GY
ForenSeq MainstAY
2014
2015
2016
2017
2018
2019
2020
2021

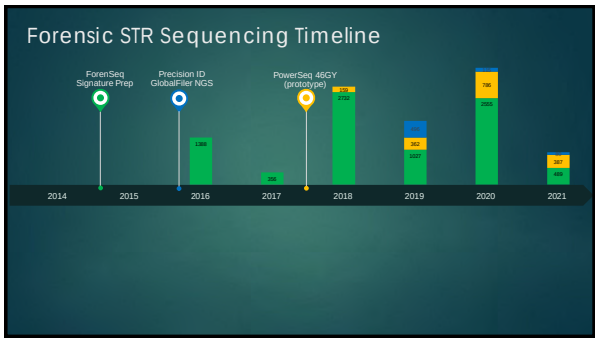
9


Forensic STR Sequencing Timeline
ForenSeq Signature Prep
Precision ID GlobalFiler NGS
PowerSeq 46GY (prototype)
PowerSeq 46GY
ForenSeq MainstAY
2014
2015
2016
2017
2018
2019
2020
2021

10


Forensic STR Sequencing Timeline
ForenSeq Signature Prep
Precision ID GlobalFiler NGS
PowerSeq 46GY (prototype)
1388
356
2891
1885
3514
1786
2014
2015
2016
2017
2018
2019
2020
2021

11


Forensic STR Sequencing Timeline
ForenSeq Signature Prep
Precision ID GlobalFiler NGS
PowerSeq 46GY (prototype)
1388
356
2732
1027
362
2668
786
387
450
2014
2015
2016
2017
2018
2019
2020
2021

12

CODIS Core STR Loci

Locus	Verogen ForenSeq Signature Prep	Verogen ForenSeq MainstrAY	ThermoFisher Precision ID GlobalFiler NGS	Promega PowerSeq 46GY
D13S1656	✓	✓	✓	✓
TPOX	✓	✓	✓	✓
D2S441	✓	✓	✓	✓
D2S1338	✓	✓	✓	✓
D3S1358	✓	✓	✓	✓
FGA	✓	✓	✓	✓
D5S818	✓	✓	✓	✓
CSF1PO	✓	✓	✓	✓
D7S820	✓	✓	✓	✓
D8S1179	✓	✓	✓	✓
D10S1248	✓	✓	✓	✓
TH01	✓	✓	✓	✓
vWA	✓	✓	✓	✓
D12S391	✓	✓	✓	✓
D13S317	✓	✓	✓	✓
D16S539	✓	✓	✓	✓
D18S51	✓	✓	✓	✓
D19S433	✓	✓	✓	✓
D21S11	✓	✓	✓	✓
D22S1045	✓	✓	✓	✓

Additional Autosomal STR Loci

Locus	Verogen ForenSeq Signature Prep	Verogen ForenSeq MainstrAY	ThermoFisher Precision ID GlobalFiler NGS	Promega PowerSeq 46GY
PENTAE	✓	✓		✓
PENTAD	✓	✓		✓
D9S1122	✓	✓		
D17S1301	✓	✓		
D20S482	✓	✓		
D4S2408	✓	✓	✓	
D6S1043	✓	✓	✓	
D1S1677			✓	
D2S1776			✓	
D18S51			✓	
D2S2800			✓	
D6S474			✓	
D12SATA63			✓	
D14S1434			✓	
SE33	(✓)			

(✓) = Sequenced but not reported in software

13

Y-STR Loci

Locus	Verogen ForenSeq Signature Prep	Verogen ForenSeq MainstrAY	ThermoFisher Precision ID GlobalFiler NGS	Promega PowerSeq 46GY
DYS391	✓	✓	✓	✓
DYS19	✓	✓		✓
DYS385a/b	✓	✓		✓
DYS389I	✓	✓		✓
DYS389II	✓	✓		✓
DYS390	✓	✓		✓
DYS392	✓	✓		✓
DYS437	✓	✓		✓
DYS438	✓	✓		✓
DYS439	✓	✓		✓
DYS448	✓	✓		✓
DYS481	✓	✓		✓
DYS533	✓	✓		✓
DYS549	✓	✓		✓
DYS570	✓	✓		✓
DYS576	✓	✓		✓
DYS635	✓	✓		✓
DYS643	✓	✓		✓
GATAH4	✓	✓		✓
DYF387S1	✓	✓		
DYS460	✓	✓		
DYS461	(✓)	(✓)		
DYS505	✓	✓		
DYS522	✓	✓		
DYS612	✓	✓		
DYS393		✓		✓
DYS456	(✓)			✓
DYS458				✓

Bold = YHRD Minimal Haplotype

(✓) = Sequenced but not reported in software

X-STR Loci, Amel

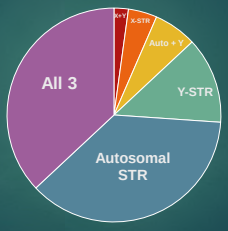
Locus	Verogen ForenSeq Signature Prep	Verogen ForenSeq MainstrAY	ThermoFisher Precision ID GlobalFiler NGS	Promega PowerSeq 46GY
DXS10135	✓			
DXS8378	✓			
DXS7132	✓			
DXS10074	✓			
DXS10103	✓			
HPRTB	✓			
DXS7423	✓			
DXS10148	(✓)			
DXS8377	(✓)			
AMEL	✓	✓	✓	✓

(✓) = Sequenced but not reported in software

14

STR Sequence-Based Population Study Publications

since 2015



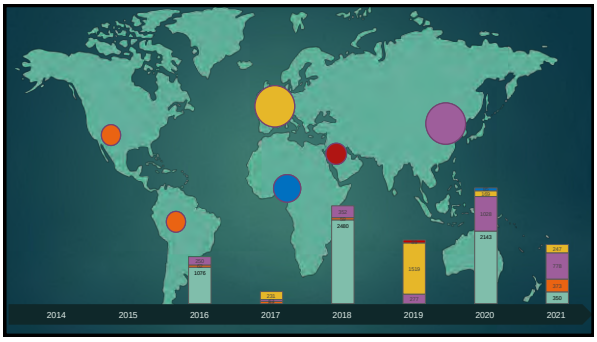
15

Forensic STR Sequencing Timeline

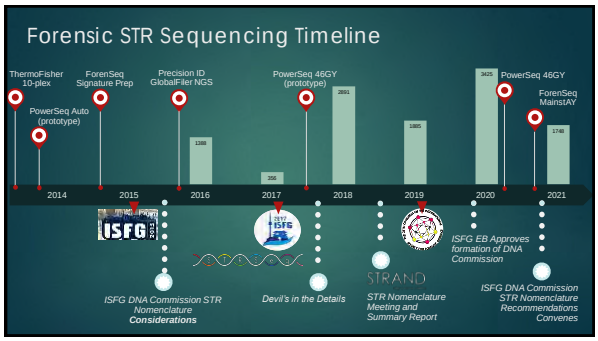
16

Forensic STR Sequencing Timeline

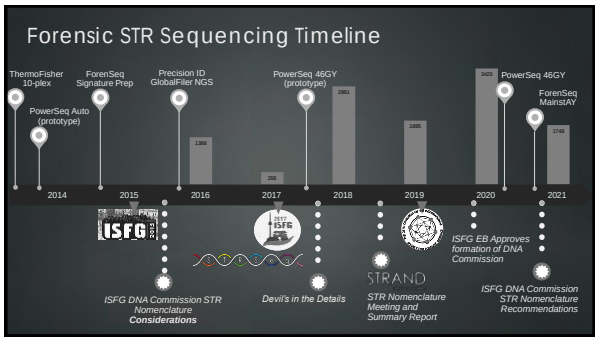
17



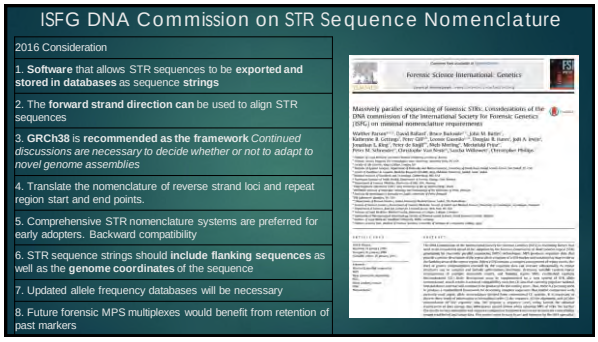
18



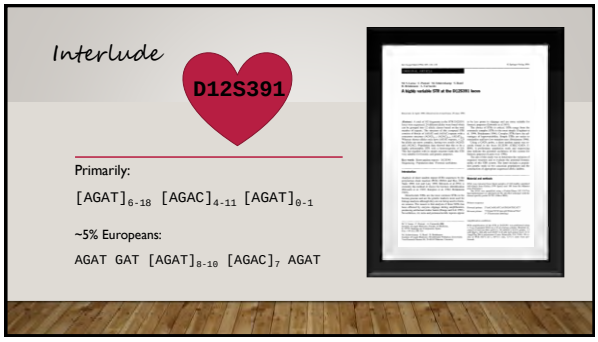
19



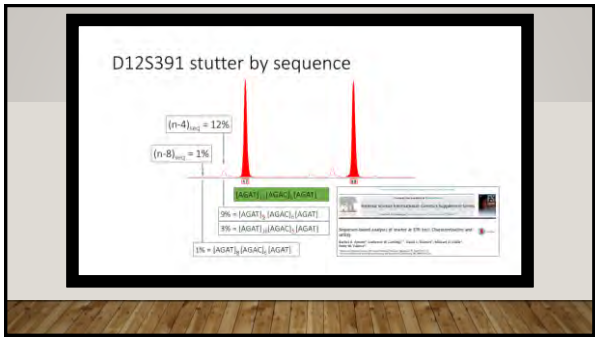
20



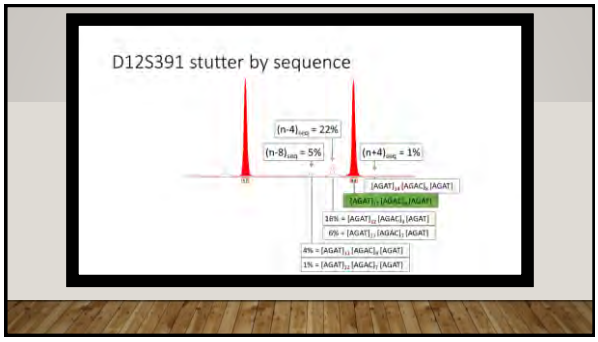
21



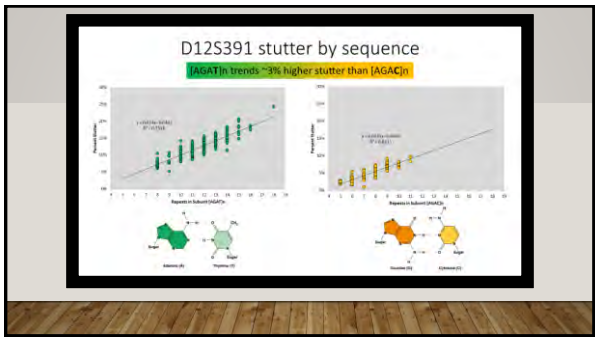
22



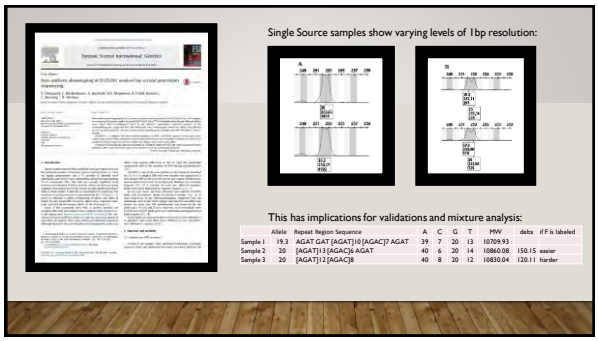
23



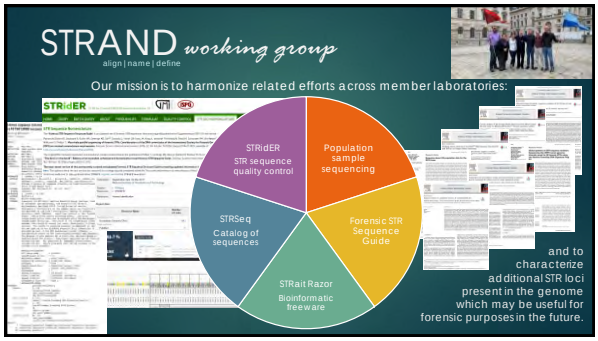
24



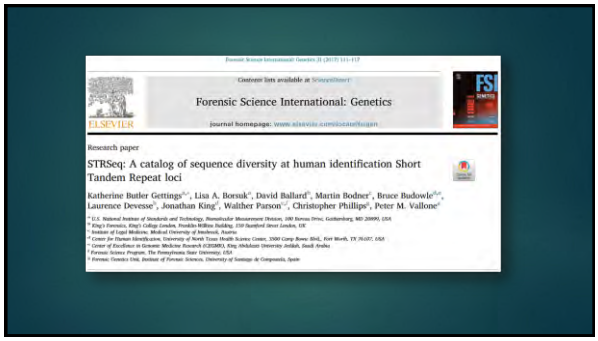
25



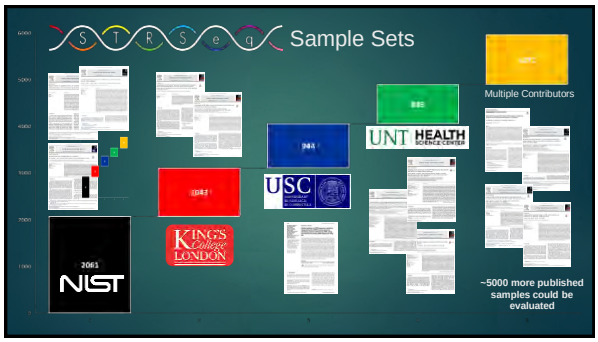
26



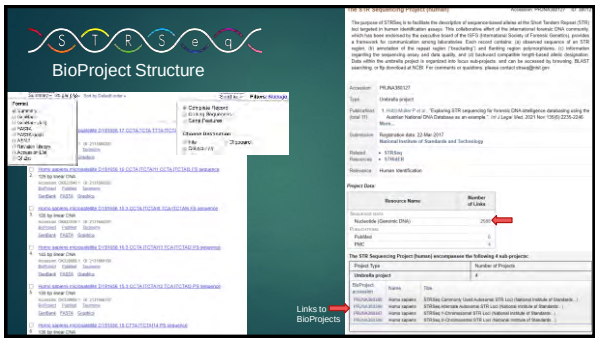
27



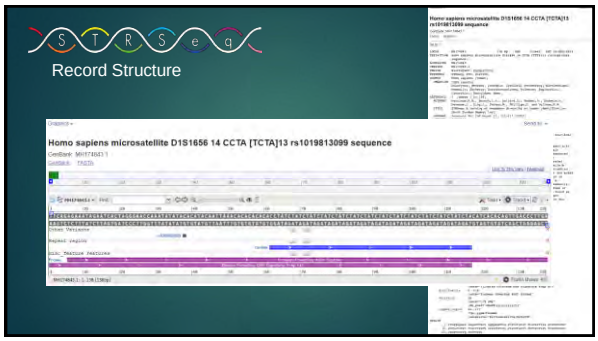
28



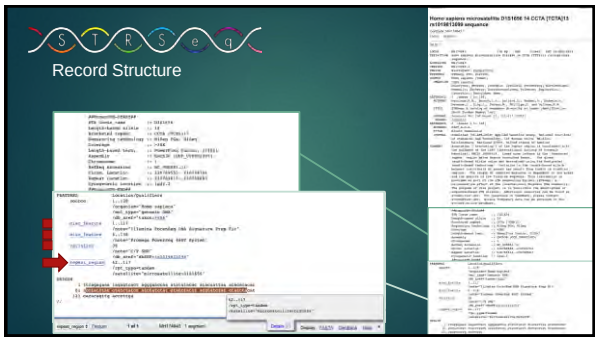
29



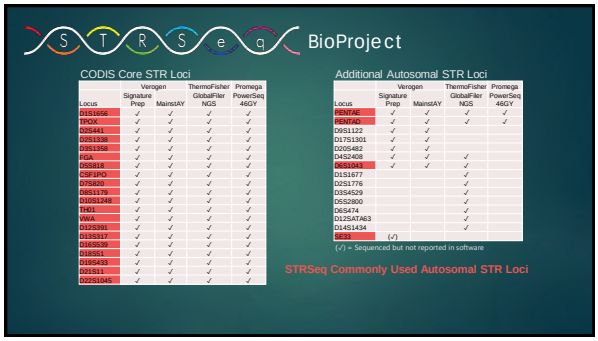
30



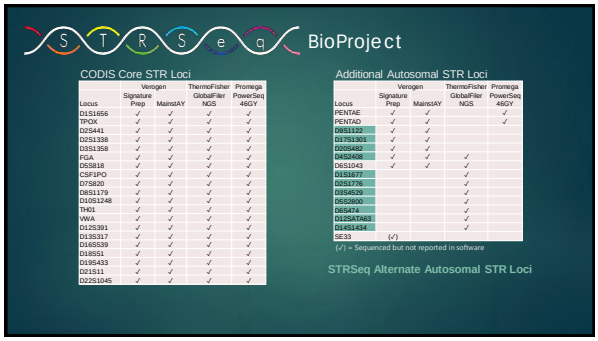
31



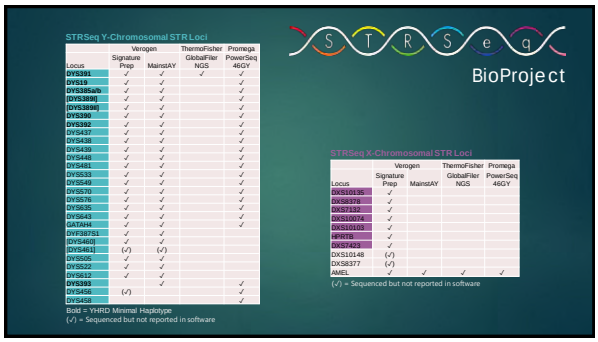
32



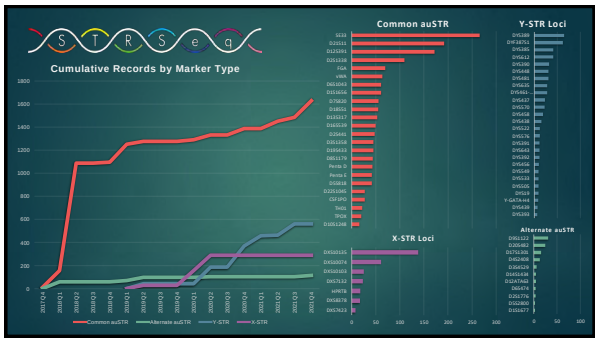
33



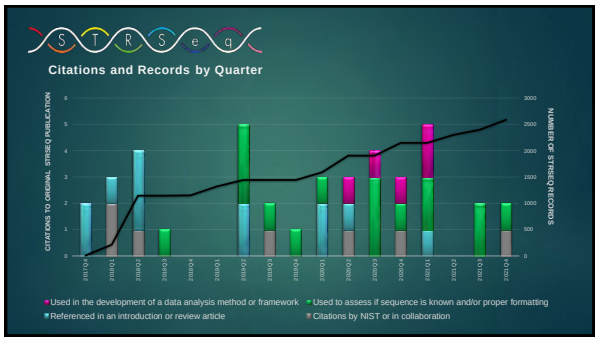
34



35

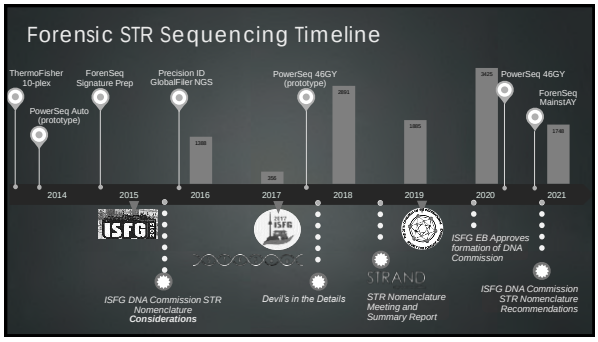


36



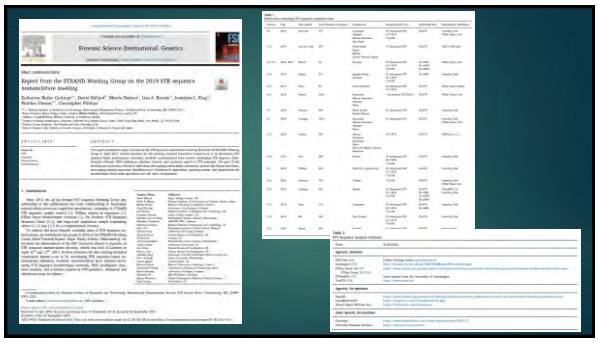
37

38

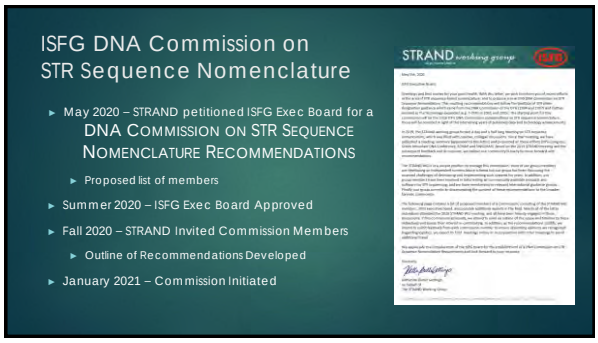


39

40



41



42

ISFG DNA Commission on STR Sequence Nomenclature

2016 Consideration	2021 Recommendation
1. **Software** that allows STR sequences to be **exported and stored in databases** as sequence **strings** 2. The **forward strand direction can** be used to align STR sequences 3. **GRCh38** is **recommended as the framework** *Continued discussions are necessary to decide whether or not to adapt to novel genome assemblies*	**1. STRINGS** Sequenced STR alleles **should** be *maintained* as sequence strings oriented to the forward strand of the current genome assembly. Sequences **should** include the minimum genomic coordinate range described herein, which is designed to provide sufficient flanking region to distinguish the termini of the repeat region.
4. Translate the nomenclature of reverse strand loci and repeat region start and end points. 6. STR sequence strings should **include flanking sequences** as well as the **genome coordinates** of the sequence	Recommended minimum range
5. Comprehensive STR nomenclature systems are preferred for early adopters. Backward compatibility	2. **BRACKETED REPEAT** Connection between length and seq representation Universal parameters for all loci or only new loci
7. Updated allele frequency databases will be necessary 8. Future forensic MPS multiplexes would benefit from retention of past markers	3. **RESOURCES** 4. **NEW LOCI**
	5. **DATABASING** • Length based STR profiles from sequence data can currently be searched/stored • Kit information to alert users when sequence data • Future database software development

43

ISFG DNA Commission on STR Sequence Nomenclature

- January 2021 - Kickoff meeting
  - STRAND Meeting
- April 2021 - 1st Commission Meeting
  - STRAND Meeting
  - Minimum Range Subgroup Meeting & Work
- July 2021 - 2nd Commission Meeting
  - Databasing Meeting
  - Minimum Range Subgroup Meeting & Work
- November 2021 – 3rd Commission Meeting
  - Databasing Meeting
  - STRAND Meeting

44

STR Sequence Nomenclature

Defined Coordinates

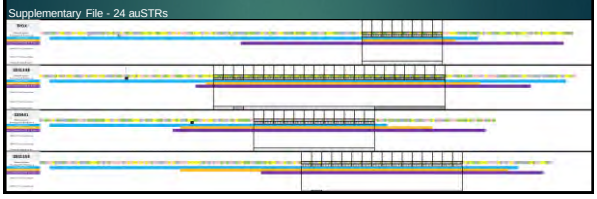
45

STR Sequence Nomenclature

Defined Coordinates

PowerSeq 46GY	GeneMarker NGS Range
ForenSeq DNA Signature Prep Kit	UAS Flanking Region Report Range
Precision ID GlobalFiler NGS v2	Converge .bed file range

Supplementary File - 24 auSTRs

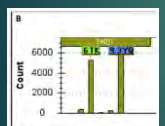


46

STR Sequence Nomenclature

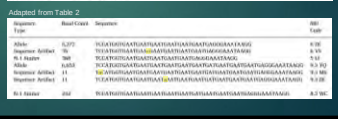
Formats for STR Sequences

Short Designator



Adapted from Figure 1

Adapted from Table 2



47

STR Sequence Nomenclature

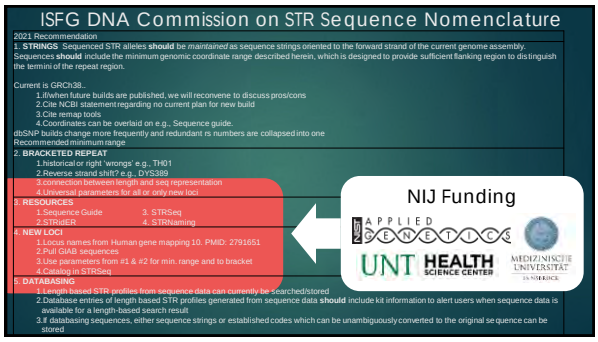
Formats for STR Sequences

Bracketed Repeat

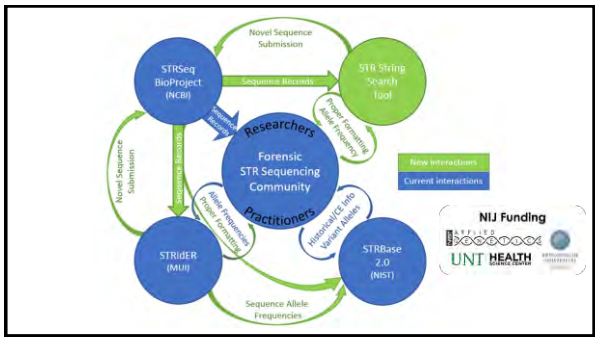
- STRNaming
  - Automated conversion of string into bracketed format
  - Based on defined set of parameters
  - CE length is maintained in the full allele name
  - Bracketing might not indicate CE length, e.g. D13S317
  - Flank polymorphisms oriented relative to repeat


```
CE12_TATC[12]AATC[2]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
```

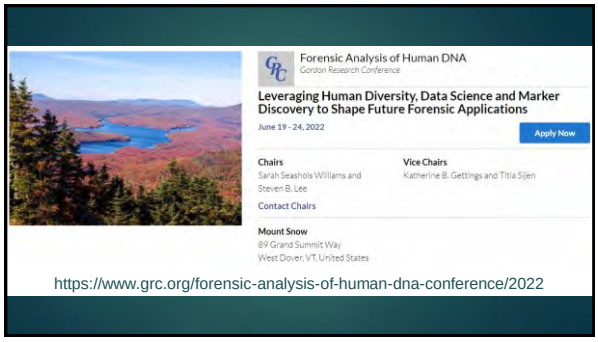
48


ISFG DNA Commission on STR Sequence Nomenclature
2021 Recommendation
1. STRINGS: Sequenced STR alleles should be maintained as sequence strings oriented to the forward strand of the current genome assembly. Sequences should include the minimum genomic coordinate range described herein, which is designed to provide sufficient flanking region to distinguish the termini of the repeat region.
Current is GRCh38..
1.If/when future builds are published, we will reconvene to discuss pros/cons
2.Cite NCBI statement regarding no current plan for new build
3.Cite remap tools
4.Coordinates can be overlaid on e.g., Sequence guide.
dbSNP builds change more frequently and redundant rs numbers are collapsed into one
Recommended minimum range
2. BRACKETED REPEAT
1.historical or right 'wrongs' e.g., TH01
2.Reverse strand shift? e.g., DYS389
3.connection between length and seq representation
4.Universal parameters for all or only new loci
3. RESOURCES
1.Sequence Guide
2.STRidER
3. STRSeq
4. STRNaming
4. NEW LOCI
1.Locus names from Human gene mapping 10. PMID: 2791651
2.Pull GIAB sequences
3.Use parameters from #1 & #2 for min. range and to bracket
4.Catalog in STRSeq
5. DATABASING
1.Length based STR profiles from sequence data can currently be searched/stored
2.Database entries of length based STR profiles generated from sequence data should include kit information to alert users when sequence data is available for a length-based search result
3.If databasing sequences, either sequence strings or established codes which can be unambiguously converted to the original sequence can be stored
NIJ Funding
APPLIED GENETICS
UNT HEALTH SCIENCE CENTER
MEDIZINISCHE UNIVERSITÄT IN INNSBRUCK

49


Novel Sequence Submission
STRSeq BioProject (NCBI)
Sequence Records
STR String Search Tool
Sequence Records
Proposed Formatting Allele Frequency
Novel Sequence Submission
Researchers
Forensic STR Sequencing Community
Practitioners
New Interactions
Current interactions
Allele Frequencies Proposed Formatting
Historical/CE Info Variant Alleles
STRidER (MUI)
STRBase 2.0 (NIST)
Sequence Allele Frequencies
NIJ Funding
UNT HEALTH SCIENCE CENTER

50


Forensic Analysis of Human DNA
Gordon Research Conference
Leveraging Human Diversity, Data Science and Marker Discovery to Shape Future Forensic Applications
June 19 - 24, 2022
Apply Now
Chairs
Sarah Seashols Williams and Steven B. Lee
Vice Chairs
Katherine B. Gettings and Titia Sijen
Contact Chairs
Mount Snow
89 Grand Summit Way
West Dover, VT, United States
https://www.grc.org/forensic-analysis-of-human-dna-conference/2022

51


Thank You!
STRAND Working Group
Jonathan King – UNTHSC
Lisa Borsuk – NIST
Chris Phillips – USC
Martin Bodner – MUI
David Ballard – KCL
Walther Parson – MUI
ISFG Executive Board & DNA Commission on STR Nomenclature Recommendations
NIST Applied Genetics Group
Kevin Kiesler
Becky Steffen
Lisa Borsuk
Tunde Huszar
Peter Vallone
Funding Sources
NIST Special Programs Office
NIJ Interagency Agreement

52

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

NIST DNA Standard Reference Materials

National Institute of Standards and Technology (NIST)
Becky Steffen

Module 8

1

Acknowledgments and Disclaimer

Points of view are the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities 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 National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

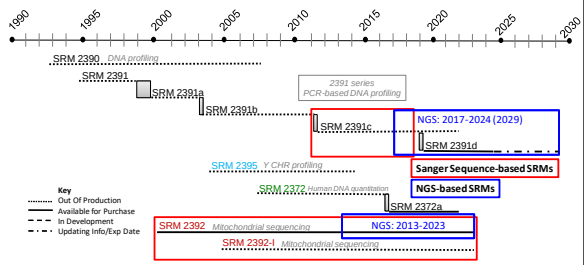
2

Topics for Discussion

- Historical perspective of NIST Standard Reference Materials (SRMs) with a focus on *sequence-based* SRMs
  - SRM 2392 and 2392-I: Mitochondrial DNA Sequencing
  - SRM 2391d: PCR-Based DNA Profiling Standard
- How next generation sequencing (NGS) is used to characterize SRM 2391d
  - Markers, kits, and instruments covered
  - SRM 2391d NGS data
- SRM 2391d update in 2022
  - Recently released NGS kits and panels included
  - Extended expiration date by 5 years (2024→ 2029)

3

NIST Forensic SRM Timeline



4

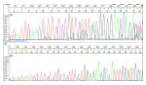
NIST SRM 2392 & 2392-I

- Mitochondrial DNA sequencing Standard Reference Materials
  - Characterized for mtDNA genome sequence composition
  - Reference used to validate measurement techniques
- SRM 2392
  - Contains 3 components (extracted DNA)
    - 2392 A – From cell line CHR
    - 2392 B – From cell line 9947A
    - 2392 C – Cloned region of heteroplasmy
- SRM 2392–I
  - From cell line HL-60

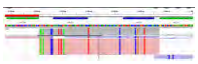
5

NGS vs Sanger Sequencing

- Mitochondrial SRMs were initially characterized with Sanger sequencing
  - Levin et al. NIST Special Publication 260-155 (2003)
  - http://www.nist.gov/srm/upload/sp260-155.pdf



- Introduction of next generation sequencing (NGS) in 2013
  - Whole mitochondrial genome analysis = more information
  - Potential for improved sensitivity
    - Detection of minor SNP variants - heteroplasmy
  - Confirm SRM sequence with an orthogonal technique



- Initial approach for NIST experiments
  - Sequenced on multiple NGS platforms (Ion PGM Torrent and Illumina MiSeq)
  - To understand differences between platforms
  - Gain practical experience in library preparation, sequence data generation, and assembly/variant calling

6

Low Level (<20%) Heteroplasmy Detected with NGS

- SRM 2392: 3 sites in 9947A (Component B)
- SRM 2392-I: 2 sites in HL-60

7

From Table 1, Riman *et al* (2017)

Low Level Heteroplasmy

SRM 2392: Component B (9947A)

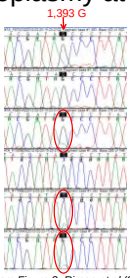
Nucleotide Position	rCRS Reference Sequence	SRM-9947A Sanger Call	SRM-9947A PGM Consensus	SRM-9947A MiSeq Consensus
93	A	G	G	G
195	T	C	C	C
214	A	G	G	G
263	A	G	G	G
309.1		C		
315.2		C		
315.3		C		
750	A	G	G	G
1393	G	G	G/A	G/A
1438	A	G	G	G
3242	G	G	G/A	G/A
4135	T	C	C	C
4769	A	G	G	G
5147	T	C	C	C
7861	T	C	T/C	T/C
8860	A	G	G	G
8860	A	G	G	G
9315	T	C	C	C
11572	T	C	C	C
11759	G	A	A	A
15326	A	G	G	G
16311	T	C	C	C
16519	T	C	C	C

SRM 2392-I: HL-60

Nucleotide Position	rCRS Reference Sequence	SRM-HL-60 Sanger Call	SRM-HL-60 PGM Consensus	SRM-HL-60 MiSeq Consensus
73	A	A	A	A
195	T	T	T	T
263	A	A	A	A
309.1		C	C	C
315.1		C	C	C
750	A	G	G	G
1393	G	G	G/A	G/A
1438	A	G	G	G
3242	G	G	G/A	G/A
4769	A	G	G	G
5147	T	C	C	C
7861	T	T	T	T
8860	A	A	G	G
10172	A	A	A	A
10398	A	A	A	G
11251	A	G	G	G
11719	A	A	A	G
12071	T	T/C	T/C	T/C
12612	A	A	A	G
13708	G	A	A	A
14668	A	A	A	G
14766	T	T	T	T
14857	A	A	A	G
15326	A	G	G	G
15812	A	A	A	G
16093	T	T	T	T
16129	G	A	A	A
16179	T	T	T	T
16362	T	T	T	T

9

Heteroplasmy at 1,393 – Sanger Seq



- 3/6 of Sanger reads show possible low-level heteroplasmy
  - Red circles
- Not reproducible in all reads
  - Not always detected by Sanger sequencing
- NGS PHP detection >15%

MiSeq Sequencing		PGM Sequencing	
1393		1393	
G	A	G	A
82%	18%	84.50%	15.50%

From Figure 2, Riman *et al* (2017)

10

Heteroplasmy and Allele Calling Thresholds

Threshold	Comp A SRM-CHR	Comp B SRM-9947A	Comp C SRM-HL-60
20%	64 C/T	7861 T/C	12071 T/C
10%	64 C/T	1393 G/A 7861 T/C	12071 T/C
5%	64 C/T	1393 G/A 7861 T/C	2445 T/C 5149 C/T 12071 T/C
1%	64 C/T	1393 G/A 3242 G/A 7861 T/C	2445 T/C 5149 C/T 12071 T/C

- PHP detected on both platforms, in replicates, and in different bioinformatic pipelines
- 64 C/T and 12071 T/C were previously detected by Sanger sequencing

Table 2, Riman *et al* (2017)

11

SRM 2392 & 2392-I → SRM 2391d

- SRM 2392-I
  - Discontinued and replaced by SRM 2391d when it was released in 2019
  - Expires on 3/31/2023
- SRM 2392
  - Still available through the Office of Reference Materials (ORM) until 7/31/2023
  - Once it expires, it will be replaced by SRM 2391d moving forward

12

mtDNA Whole Genome Data for SRM 2391d

- mtDNA whole omponents A-C, and E (Compo
- One heteropla

SRM 2391d Component B

Site	rCRS	Component B	Comment
15326	A	G	
16093	T	Y	C/T Heteroplasmy *
16129	G	A	
16187	C	T	
16189	T	C	
16223	C	T	
16263	T	C	
16278	C	T	
16293	A	G	
16294	C	T	
16311	T	C	
16360	C	T	
16368	T	C	
16519	T	C	

*Low-level heteroplasmy approximately 5 % T detected

SRM 2391d Component A

Site	rCRS	Component A	Comment
73	A	G	
195	C	T	
263	A	G	
309.1	:	C	Insertion
315.3	:	C	Insertion
709	G	A	
750	A	G	
930	G	A	
1438	A	G	
1888	G	A	
2706	A	G	
3107	C	:	Deletion
4216	T	C	
4769	A	G	
4917	A	G	
5147	G	A	
7028	C	T	
7418	C	T	
8697	G	A	
8860	A	G	
825	T	A	
1018	G	A	
1438	A	G	
1762	A	G	
1018	G	A	
1738	T	C	
2352	T	C	
2706	A	G	

SRM 2391d Component E

Site	rCRS	Component E	Comment
73	A	G	
195	T	C	
263	A	G	
309.1	:	C	Insertion
315.1	:	C	Insertion
709	G	A	
750	A	G	
1438	A	G	
1888	G	A	
2706	A	G	
3107	C	:	Deletion
4216	T	C	
4464	G	A	
4769	A	G	
4917	A	G	
7028	C	T	
8697	G	A	
8860	A	G	
9180	A	G	
10463	T	C	

Historical Perspective of SRM 2391d

Past

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material 2390
DNA Profiling Standard

RFLP Testing & DNA Probes (1990)

National Institute of Standards and Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

PCR-Based Testing (1995)
- VNTR, Dot Blot
- STR typing (updated 1998)

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

PCR-Based Testing (2000)
- Focus on STR typing
- VNTR, Dot Blot

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

PCR-Based Testing (2003)
- Autosomal STR loci
- More STR loci added (updated 2008)

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

PCR-Based Y-STR Testing (2003)
- Y-STR loci
- More Y-STR loci added (updated 2008)

Present: SRM 2391d

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391d
PCR-Based DNA Profiling Standard

All certification done with NGS
Released July 2019

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

PCR-Based STR Testing (2011)
- Autosomal and Y-STR loci
- More autosomal and Y-STR loci, X-STR loci, and Indels added (updated 2015)
- Identity and Ancestry SNPs, and Y-Indel added via **NGS** (updated 2017)

13

SRM 2391d: PCR-Based DNA Profiling Standard

- Developed as a successor to SRM 2391c
  - SRM 2391c is no longer available for sale at NIST
  - The expiration date was extended by **2 years** to **Feb. 3, 2022**
  - We will not further extend the expiration date after this date
- **Next Generation Sequencing** is used for certification in addition to **Capillary Electrophoresis** testing
  - Length- and sequence-based genotypes are provided
  - Information values are included for all commercially available forensic markers **beyond STR markers**, including whole genome mtDNA

SRM 2391d is the most comprehensive NIST forensic SRM to date

14

Materials – Five Components

- **Components A-D** are genomic DNA extracted from purchased blood:
  - Not from cell lines (challenges in obtaining permission from Coriell/NIGMS)
  - *May* be more commutable (similar to casework)
  - **Different samples from 2391c**
- **Component E** consists of cells spotted onto FTA paper
  - Two 6 mm punches; approximately 75,000 cells per punch
  - Toward the end of SRM 2391c profile degradation was observed for cells stored on 903 paper (cells on 903 paper not included in SRM 2391d)
  - **Same cell line as used in 2391c (CRL-1486)**

Table 1 Description of Components in SRM 2391d

Component	Description	Volume	Concentration[a,b]
A	Amplicons: single-source female genomic DNA in TE^{-4} buffer	55 µL	1 ± 0.4 ng/µL
B	Amplicons: single-source male genomic DNA in TE^{-4} buffer	55 µL	1.7 ± 0.4 ng/µL
C	Amplicons: single-source male genomic DNA in TE^{-4} buffer	55 µL	1.6 ± 0.2 ng/µL
D	Mixture of equal parts A and B and C amplicons in TE^{-4} buffer	55 µL	1.3 ± 0.4 ng/µL
E	Amplicons: single-source female cells spotted on FTA paper[c]	Two 6 mm punches	7.5 × 10^4 cells per punch

[a] DNA concentrations and cell counts are provisional Information Values.
[b] ± 95% proportional error, which is a measure of the uncertainty of the Information Values.
[c] Cells were made using Promega Corporation's Maxwell® 16 Buccal Swab LEV Purification Technical Manual, and were then spotted on FTA® Classic Card (Whatman, Inc.) according to the manufacturer's instructions.

Information Values

Concentrations were assigned using droplet digital PCR (ddPCR)

Components A-D have *different profiles* from SRM 2391c
Component E has *the same profile* as SRM 2391c

15

How are SRM 2391d values assigned?

- **Certified Values** are assigned when there is a high coverage **sequence string** available for a marker

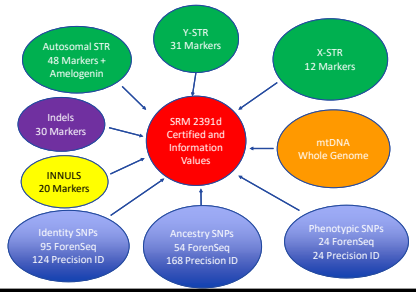
***Highest confidence*; all sources of uncertainty and bias examined**

- **Information Values** are assigned when only one primer set is used from CE testing and there is no sequence string to confirm

***For informational purposes*; no guarantees for uncertainty**

16

Markers included in the Certificate of Analysis



17

Which Autosomal STR Markers have Certified Values?

Autosomal STR Marker List	MiniFiler	Identifiler	Identifiler Plus	Identifiler Direct	NGM	NGM SElect	NGM Detect	Verifiler Plus	Verifiler Express	GlobalFiler	GlobalFiler Express	PP 5S	PP CSF	PP 16	PP 16 HS	PP 18D	PP 21	PP ESX 17	PP ESI 17 Fast	PP ESS 17 Pro	PP ESS 17 Fast	PP Fusion	PP Fusion 6C	PP VersaPlex 27PY	ESSplex SE Plus	HDplex	24plex GO!	24plex QS	ForenSeq	Precision ID GF	PowerSeq 46GY	CODIS 20	Europeam Standard Set	Certified Value	Informative Value
D1S1656																																		X	
D13S317																																		X	
D2S1338																																		X	
D2S441																																		X	
D21S11																																			X
D21S1776																																		X	
D22S1045																																			
D3S1358																																			
D7S820																																			
D8S1179																																			
D10S1248																																			
D12S391																																			
D13S317																																			
D16S539																																			
D18S51																																			
D19S433																																			
D2S1338																																			
D21S11																																			
D22S1045																																			
D3S1358																																			
D5S818																																			
D7S820																																			
D8S1179																																			
D13S317																																			
D16S539																																			
D18S51																																			
D21S11																																			

18

Summary of Values Assigned (2019)

Marker Type	Number of Certified Loci	Number of Information Loci
Autosomal STR	35	13
Y-STR	28	3
X-STR	7	5
Mitochondrial DNA	-	Full mtGenome
Indel/Innuls	-	50
SNPs	-	323

19

Platforms Used for NGS Testing

Next Generation Sequencing (NGS) was performed with two different instruments:

- MiSeq FGx (Verogen)
- Ion S5 XL (ThermoFisher)

MiSeq FGx

Ion S5 XL

20

Commercial NGS Kits that were tested (11 Kits Total)

AFDIL/MiSeq (1)	Verogen/MiSeq (1)	Thermo Fisher/Ion S5 XL (5)	Promega/MiSeq (2)	Qiagen/MiSeq (2)
mtDNA Whole Genome	ForenSeq Signature Prep Kit	Precision ID GlobalFiler NGS STR Panel v2	PowerSeq 46GY (prototype)	QIAseq mtDNA Whole Genome Panel
		Precision ID Ancestry Panel	PowerSeq CRM Nested System (mtDNA control region)	QIAseq SNP Panel
		Precision ID Identity Panel		
		Precision ID Phenotype Panel		
		Precision ID mtDNA Whole Genome Panel		

Ring *et al.* (2017)

21

SRM 2391d NGS Data: Component C

22

mtDNA Whole Genome Sequencing

EMPOP results:
https://empop.online/haplotypes#matches_details

Haplogroup	Ancestry	Match
L1b1a12	African	unique

AFDIL mtDNA Whole Genome Sequencing protocol* (MiSeq)

Ring, J.D., Sturk-Angreaggi, K., Peck, M.A., Marshall, C. (2017) A performance evaluation of Nextera XT and KAPA HyperPlus for rapid Illumina library preparation of long-range mitogenome amplicons. Forensic Sci Int Genet 29:174-180.

23

SNP Phenotype and Ancestry Estimation (ForenSeq DNA Signature Prep Kit)

Hair Color Results

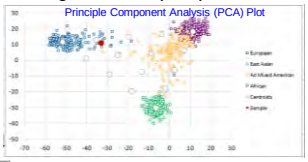
Brown	0.16
Red	0.00
Black	0.84
Blond	0.00

Eye Color Results

Intermediate	0.00
Brown	1.00
Blue	0.00

Biogeographical Ancestry Results

Distance to Nearest Centroid 3.38



Population/Region, samplesize (N)	Probability of Genotype in each Population	Likelihood Ratio
Amindi(Africa,40)	9.1.378E-15	1
African-American/AAW(Africa,123)	9.1.944E-14	7.18
Sandawe(Africa,80)	9.1.824E-14	6.64
Ethiopian Jews(Africa,64)	1.072E-10	5.3.3
African-American(Africa,182)	9.1.14E-13	0.2.1
Mozi(Africa,44)	8.1.7E-14	0.1.0
Chagga(Africa,80)	1.2.80E-8	0.2.9.1.0
Luhya(AWK)(Africa,198)	9.1.97E-20	0.9.9.0.0
Gonja(Africa,30)	3.2.01E-20	0.9.0.0.0
Bantu(Africa,70)	8.4.67E-21	0.1.0.0.0.0

* Indicates the values are within an order of magnitude of the highest likelihood.

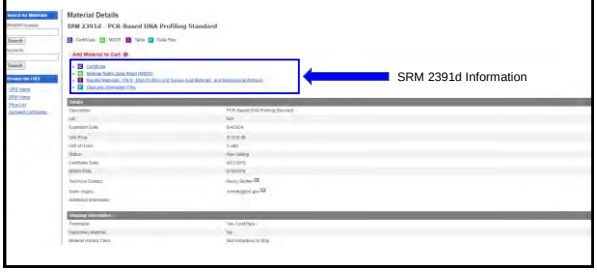
Other Markers Determined:
X-STRs, Indels, INNULS, and other SNP Panels

KiddLab – Set of 55 AISNPs
Population likelihoods based on 55 SNPs and 139 reference populations for this DNA profile
http://frog.med.yale.edu/FrogKB/

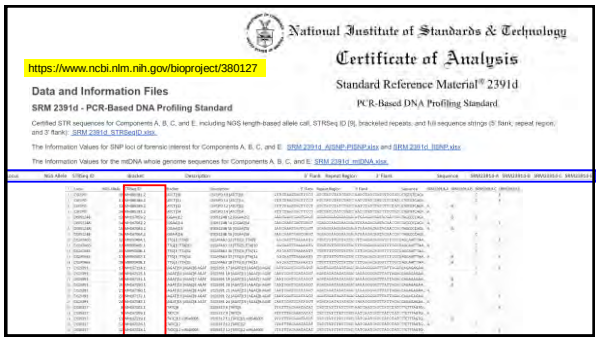
24

NGS Data and Information Files

https://www-s.nist.gov/srmors/view_detail.cfm?srm=2391d



25



26

SRM 2391d: PCR-Based DNA Profiling Standard

2022 Update

27

SRM 2391d: PCR-Based DNA Profiling Standard

2022 Update

- To confirm s omponents A-E
  - Digital PC rm quantitation
  - Quantitati check degradation
  - CE testing
- New CE kits and NGS kits have been released since 2019
  - Ask the vendors which kits to include
    - Thermo Fisher, Promega, QIAGEN, and Verogen
  - 9 total CE kits
  - 9 total NGS panels
- Extend the June 4, 2024 expiration date by 5 years

28

CE Kits to be Added (9)

Thermo Fisher (3)	QIAGEN (Investigator) (6)
GlobalFiler IQ	26plex QS
Y Filer Direct	rgus Y-28 QS
NGM SElect Exp	IDplex Plus
**Y Indel from Glo	IDplex GO!
	ESSplex SE QS
	Argus X-12 QS

*No Promega kits to add for this update

29

NGS Panels to be Added (9)

Promega (1)	Thermo Fisher (Precision ID) (4)	Verogen (ForenSeq) (4)
*PowerSeq 46GY	M	Kintelligence
	I	MainstAY
	lon ex	mtDNA Whole Genome
		mtDNA Control Region

*Prototype PowerSeq 46GY was tested previously; however, PowerSeq 46GY is now available commercially, so it was retested with SRM 2391d Components

QIAGEN (Qia (6)
M e
STR

QIAGEN has 6 prototype QIAseq panels that we are in the process of testing and may or may not be in this update

30

ForenSeq Kintelligence Kit

VEROGEN

ForenSeq Kintelligence Kit

MiSeq FGx System

UAS

GEDmatch PRO

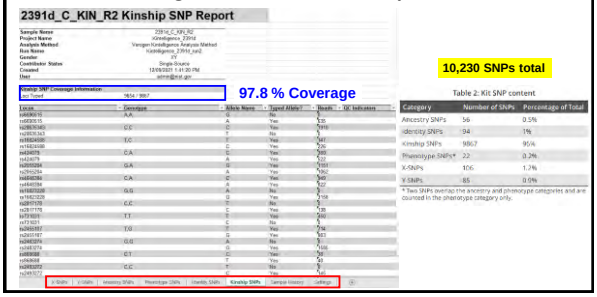
Beyond standard forensic markers: **10,230 SNPs**

- Used in Forensic Genetic Genealogy (FGG)
- Compatible with:
  - ✓ SNPs in most direct-to-consumer (DTC) genetic genealogy tests
  - ✓ SNPs in ForenSeq DNA Signature Prep kit

GED match PRO → The largest database of voluntarily submitted DNA profiles for forensic comparisons

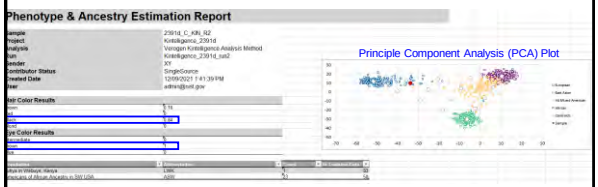
31

Kintelligence Data – Component C



32

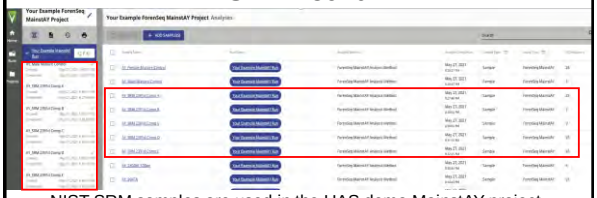
Kintelligence Data – Component C



Same results as ForenSeq Signature DNA Prep Kit

33

Verogen MainstAY UAS Demo Data: SRM 2391d



- NIST SRM samples are used in the UAS demo MainstAY project
- There are 96 samples in the demo project
  - Single source samples, mixtures, sensitivity series, replicates, and the NIST SRM 2391d components A, B, C, D, E

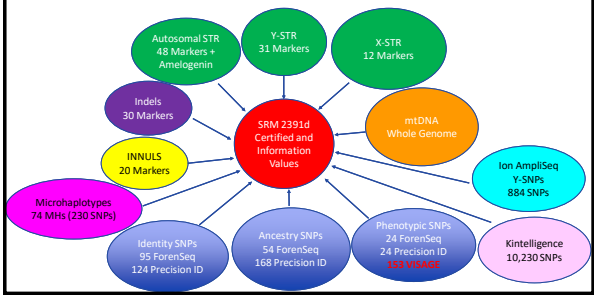
34

Summary of Values Assigned (2022)

Marker Type	Number of Certified Loci	Number of Information Loci
Autosomal STR	35	13
Y-STR	28	3
X-STR	7	5
Mitochondrial DNA	-	Full mtGenome
Indel/Innuls	-	50
SNPs	-	323 → 11,590
Microhaplotypes	-	74

35

Markers included in the current Certificate of Analysis



36

Thoughts on Sequence-Based Standards

- Sanger and NGS methods were used in parallel to characterize all STR alleles for SRM 2391c
  - All results were fully concordant
  - We established NGS as a primary method for certification for SRM 2391d
- We decided to move forward with NGS to add certified values for many reasons
  - NGS provides more information about a DNA sample
  - Multiplexing allows more markers to be sequenced in much less time
  - The process is simplified for STR markers and mtDNA whole genome
  - A sequencing workflow is added for SNPs
  - NGS is high throughput (up to 96 samples can be sequenced with some NGS panels)

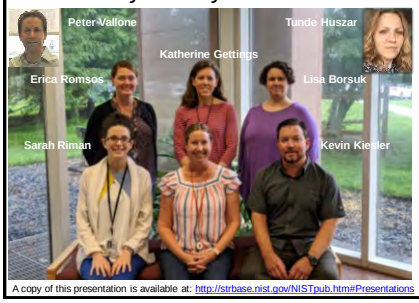
37

Summary and Final Thoughts

- SRM 2391d: PCR-Based DNA Profiling Standard was developed as the most ***comprehensive*** forensic SRM yet
  - Certified Values for STR genotypes and haplotypes
  - Information Values for commercially available forensic markers beyond STRs
  - 270 units sold since the 2019 release
- The "2022 Update" should be complete by August 2022
  - Finish data analysis
  - Complete documentation (ROA, COA, and SP-260)
- We are including Information Values for >11K additional SNPs

38

Thank you for your attention!



Questions?
becky.steffen@nist.gov
1-301-975-4275

Acknowledgements
Margaret Kline
David Duewer
Hari Iyer

A copy of this presentation is available at: http://strbase.nist.gov/NISTpub.htm#Presentations

39

BREAK

15 minutes

40

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

DNA Training Standards on the OSAC Registry and Educational Materials

National Institute of Standards and Technology (NIST)

John Paul Jones
Special Programs Office

Module 9

1

Agenda

1. Training Standards on OSAC Registry to Discuss
   - ASTM E2917-19a (Training, Continuing Ed. & Professional Development)
   - ASB 22 (Forensic DNA Analysts Training Program)
   - ASB 110 (Forensic Serological Methods)
   - ASB 23 (DNA Isolation & Purification Methods)
   - ASB 116 (Forensic DNA Quantification Methods)
   - ASB 115 (STRs using Amplification, Separation & Allele Detection)
2. Educational Materials
3. Staying Connected

2

ASTM E2917-19a Standard Practice for Forensic Science Practitioner Training, Continuing Education, and Professional Development Programs

Scope:
1.1 This practice provides foundational requirements for the training, continuing education, and professional development of forensic science practitioners to include training criteria toward competency, documentation, and implementation of training, and continuous professional development.

1.2 This practice outlines minimum training criteria and provides general information, approaches, and resources for all disciplines. The standard would complement additional specific requirements for each forensic science discipline (for example, relevant degree programs, higher education) if developed by subject matter experts in their respective fields.

3

E2917-19a cont.

Helpful Terminology

2.1.2 **competency,** n—demonstration that a forensic science practitioner has acquired and demonstrated specialized knowledge, skills, and abilities (KSAs) in the standard practices necessary to conduct examinations in a discipline or category of testing prior to performing independent casework (2).
2.1.4 **forensic science practitioner,** n—an individual who (1) applies scientific or technical practices to the recognition, collection, analysis, or interpretation of evidence for criminal and civil law or regulatory issues; and (2) issues test results, provides reports, or provides interpretations, conclusions, or opinions through testimony with respect to such evidence (3).
2.1.5 **forensic science service provider,** n—a forensic science agency or forensic science practitioner providing forensic science services (3).
2.1.7 **knowledge, skills, and abilities (KSAs),** n—the level of information, qualifications, and experience needed to perform assigned tasks.
2.1.7.1 Discussion—**Knowledge** refers to acquired understanding of the principles and practices related to a particular job, **skills** refer to acquired analytical and psychomotor behaviors, and **abilities** refer to the talents, observable behaviors, or acquired dexterity.

2 - National Commission on Forensic Science, Proficiency Testing in Forensic Science, March 2016
3 - National Commission on Forensic Science, Defining Forensic Science and Related Terms, May 2016

4

E2917-19a cont.

Training to Competency Programs: (core & discipline content)

5.3.1 Core specific elements shall include the following:
5.3.1.1 **Standards of conduct** and professional ethics.
5.3.1.2 **Safety**, including biological, chemical, and physical hazards.
5.3.1.3 **Policy**, including administrative, standard operating procedures, quality assurance and control, non-conformance remediation procedures, documentation and record control, accreditation standards and requirements, certification/ licensure standards, regulatory compliance, and security issues.
5.3.1.4 **Legal issues**, including expert testimony, depositions, rules of evidence, criminal and civil law procedures; legal obligations to disclose information and to preserve evidence; and evidence authentication (for example, chain-of custody).
5.3.1.5 **General forensic concepts** including evidence handling, interdisciplinary issues (for example, recognition, collection, and preservation of evidence), and chain of custody.
5.3.1.6 **Communication**, including written, oral, and nonverbal communication skills, report writing and interpretation, exhibit and pretrial preparation, and trial presentation.
5.3.1.7 **Human factor issues**, including factors that affect conclusions and the workplace environment, such as bias (for example, cognitive, contextual, confirmation); the process to determine what information is relevant to a task; fatigue; ergonomics; and response to errors (for example, putative vs. learning opportunity policies).

5

E2917-19a cont.

Evaluation, Time Needed & Note for Management

5.2.2.10 Evaluation of the training program to assess its efficacy and relevance within a four-year period.

6.1.1 An annual average of at least 16 hours of continuing education or professional development shall be obtained by all forensic science practitioners over a three-year period.

8.2 Forensic science practitioners and their supervisors should be allocated time and funding for continuing professional development and mentorship. Management may need to adjust resources and staffing to maintain casework loads. Neglecting to do so will negatively impact organizational effectiveness, service goals, and work product quality.

6

ANSI/ASB 022, Standard for Forensic DNA Analysis Training Programs, First Edition, 2019

1 Scope:
This standard provides the general requirements for a forensic DNA laboratory's training program in DNA analysis including data interpretation.

Note: This standard serves as a foundation for all the downstream forensic DNA training standards.

7

ANSI/ASB 022 cont.

3.1 competency
The demonstration of technical skills and knowledge necessary to perform forensic DNA analysis successfully. (from FBI QAS)

3.13 training program
A written description of activities to be performed for the purpose of establishing and maintaining competency and job-related knowledge, skills or abilities.

4.1.3 Personnel with Previous DNA Experience
Individuals with documented previous experience and training in forensic DNA analysis may be exempted from portions of the training program. The DNA technical leader shall be responsible for assessing an individual's previous training and ensuring that it is adequate and documented.

4.1.4 New DNA Processing, Data Interpretation, and Statistical Analysis Methods
When a new DNA processing, data interpretation, or statistical analysis method is incorporated into the laboratory's protocols, all personnel responsible for performing the method shall successfully complete training and competency testing prior to performing DNA analysis, data interpretation or statistical analysis.

8

ANSI/ASB 022 cont.

4.2 Content
At a minimum, the training program shall include the following topics as they apply to the work conducted by the laboratory and by the individual in training.

a) Expectations for satisfactory progression through the training program and performance on competency test(s).

Training Program covers 18 topic areas including:

1. General operation of laboratory
2. Quality management program
3. Safety
4. Applicable validations
5. Applicable software
6. Evidence handling and chain of custody
7. Theoretical & scientific basis of forensic DNA analysis
8. Technologies, methodologies, and platforms used in the laboratory
9. Practical exercises in the technologies, methodologies, and platforms used in the laboratory on samples representative of the range, type and complexity analyzed by the laboratory.
10. Data interpretation and statistical analysis
11. Report writing
12. Technical and administrative reviews
13. Cognitive bias in decision making processes associated with forensic DNA analysis
14. Applicable laws and regulations
15. Limitations of methods
16. Testimony as an expert witness
17. Ethics
18. How to conduct a validation

9

ANSI/ASB 022 cont.

4.3 Competency Testing
4.3.2 Required Testing
Prior to performing DNA analysis or data interpretation, the trainee shall successfully complete the following knowledge-based and technical competency tests, as they apply to the assigned job responsibilities.

a) Written and/or practical competency test(s) as indicated below covers, at a minimum, the following areas:
1) theoretical and scientific basis of forensic DNA analysis – written test;
2) laboratory's analytical procedures performed on samples representative of the range, type, and complexity typically analyzed by laboratory – practical test;
3) data interpretation – written and practical tests;
4) statistical analysis – written and practical tests;
5) report writing – written and practical tests;
6) technical review – practical test;
7) ethics – written test;
8) cognitive bias – written test.

b) An oral competency test(s) to demonstrate an understanding of ethics and the scientific basis of forensic DNA analysis. The oral competency test shall be designed to demonstrate that the trainee can explain the DNA analysis and data interpretation procedures and statistics used by the laboratory to both a layman and a scientific expert for positions where testimony may be required. The oral assessment shall include a mock trial exercise in addition to any other laboratory-specified requirements.

10

Common Framework for the ASB DNA Training Standards

1. Scope
2. Normative References
3. Terms and Definitions
4. Requirements
   - 4.1 General
   - 4.2 Knowledge-based Training
   - 4.3 Practical Training
   - 4.4 Competency Testing
     - 4.4.1 General
     - 4.4.2 Knowledge-based Competency
     - 4.4.3 Practical Competency
5. Conformance

11

ANSI/ASB 110, Standards for Training in Forensic Serological Methods, First Edition, 2020

1 Scope:
This standard provides the general requirements for a forensic serology training program to evaluate body fluids, stains, or residues related to forensic investigations.

This standard does not address training in forensic DNA analysis procedures.

12

ANSI/ASB 110 cont.

3.1 **confirmatory test**
A test that is specific for the presence of a body fluid, stain, or residue of interest, and reduces or eliminates false positive results.

3.3 **forensic serology**
The detection, characterization, identification, and/or typing of body tissues and fluids, either in native form or as stains or residues left at a crime scene using physical methods (e.g. normal and enhanced lighting), biochemical assays, reactions and/or microscopy.

3.4 **presumptive test**
A screening test that indicates the possible presence of a material of interest. A positive presumptive test result does not constitute the identification of that material. A negative presumptive test indicates that the material of interest was not detected; it is not confirmation of its absence. Presumptive tests are sensitive but not specific and can lead to false positive results.

13

13

ANSI/ASB 110 cont.

4.2.3 At a minimum, the knowledge-based portion of the training program shall cover the following topics:

a) the fundamentals of serological testing and the composition of body fluids;
b) mechanisms of biological fluid examinations to include visual and chemical analyses;
c) information regarding test specificity and limits of detection for presumptive and confirmatory testing;
d) the analytical information involved in establishing which assay to use (e.g. size of stain, age of stain);
e) the proper preservation of biological material to include safety, handling, packaging, storing, and chain of custody procedures to maintain the integrity of the evidence;
f) limitations of the methodology.

14

14

ANSI/ASB 023, Standard for Training in Forensic DNA Isolation and Purification Methods, First Edition, 2020

1 Scope:
This document provides requirements to ensure proper training in the methods of DNA isolation and purification used within the trainee's forensic DNA laboratory.

3.2 contamination - The unintentional introduction of exogenous DNA or other biological material in a DNA sample, PCR reaction, or item of evidence; the exogenous DNA or biological material could be present before the sample is collected or introduced during collection or testing of the sample.

3.3 degradation - The fragmenting, or breakdown, of DNA by chemical, physical, or biological means.

3.9 PCR inhibitor - Any substance that interferes with or prevents the synthesis of DNA during the amplification process.

ANSI/ASB Standard 023, First Edition 2020

Standard for Training in Forensic DNA Isolation and Purification Methods

Added to Registry: August 3, 2021

15

ANSI/ASB 023 cont.

4.2 Knowledge-based Training
4.2.1 The laboratory's training program shall provide the trainee with an understanding of the fundamental principles of the theory behind the various isolation methods, the function of the reagents and other components used in each method, the limitations of each method, and the laboratory's own DNA isolation and purification protocols.
4.2.3 At a minimum, the knowledge-based portion of the training program shall cover the following topics. (11 topics)

1. Composition of DNA within cells…
2. Impact of exposure to heat, humidity, mechanical breakage, and chemicals on DNA stability to include the mechanisms of DNA degradation.
3. Cell lysis and separation of DNA from other materials…
4. Methods for DNA isolation and purification used in the laboratory…
5. Methods based on sample type used in the laboratory…
6. DNA Yield…
7. PCR inhibitors…
8. Contamination…
9. Quality control in the DNA isolation and purification process to include, reagent blank control(s) and any other extraction controls.
10. Storage, preservation, and retention of extracted DNA according to laboratory policy.
11. Troubleshooting…

16

ANSI/ASB 116, Standard for Training in Forensic DNA Quantification Methods, First Edition, 2020

1 Scope:
This standard provides the requirements for a forensic DNA laboratory's training program in DNA quantification.

3.1 **DNA quantification** - A process by which the DNA concentration in a sample is determined.

3.2 **cycle threshold** - Cycle number (in quantitative PCR) at which the fluorescence generated within a reaction exceeds a defined threshold; this value is converted to a DNA concentration for each sample tested using a standard curve developed from DNA samples of known concentrations.

3.5 **Quantitative PCR (qPCR)** - A means for quantifying the amount of nucleic acid present in a sample using PCR.

ANSI/ASB Standard 116, First Edition 2020

Standard for Training in Forensic DNA Quantification Methods

Added to Registry: August 3, 2021

17

17

ANSI/ASB 116 cont.

4.2 Knowledge-based Training
4.2.3 At a minimum, the knowledge-based portion of the training program shall cover the following topics.

a) Principles and limitations of non-PCR based DNA quantification methods…
b) Principles and limitations of quantitative PCR (qPCR) DNA quantification methods…
c) Characteristics, performance, limitations, and information provided by PCR and non-PCR based methods of DNA quantification…
d) Characteristics of results of different methods of DNA quantification…
e) Interpretation of results…
f) Instrumentation and reagents…
g) Troubleshooting…

18

18

ANSI/ASB 115, Standard for Training in Forensic Short Tandem Repeat Typing Methods using Amplification, DNA Separation, and Allele Detection, First Edition, 2020

1 Scope:
This standard provides the requirements of a forensic DNA laboratory's training program in forensic Short Tandem Repeat typing methods using amplification, DNA separation and allele detection.

ANSI/ASB Standard 115, First Edition 2020

Standard for Training in Forensic Short Tandem Repeat Typing Methods using Amplification, DNA Separation, and Allele Detection

OSAC REGISTRY

Added to Registry: August 3, 2021

19

19

ANSI/ASB 115 cont.

3.3 analytical threshold
1)The minimum height requirement at and above which detected peaks on a STR DNA profile electropherogram can be reliably distinguished from background noise; peaks above this threshold are generally not considered noise and are either artifacts or true alleles. 2) A "Relative Fluorescence Units" (RFU) level determined to be appropriate for use in the PCR/STR DNA typing process; a minimum threshold for data comparison is identified by the specific forensic laboratory through independent validation studies.

3.4 artifact
A non-allelic product of the amplification process (e.g., stutter, non-templated nucleotide addition, or other non-specific product), an anomaly of the detection process (e.g., pull-up or spike), or a byproduct of primer synthesis (e.g., "dye blob") that may be observed on an electropherogram; some artifacts may complicate the interpretation of DNA profiles when they cannot be distinguished from the actual allele(s) from a particular sample.

3.13 stochastic
1) Chance, or random variation 2) in DNA testing, refers to random sampling error from extracts containing low levels of DNA and/or random variation in selection of alleles amplified at a particular locus.

20

20

ANSI/ASB 115 cont.

4.2 Knowledge-based Training
4.2.3 At a minimum, the knowledge-based portion of the training program shall cover the following topics.

a) STRs in forensic DNA analysis…
b) Polymerase chain reaction…
c) DNA separation…
d) DNA detection…
e) Instrumentation and reagents…
f) Contamination…
g) Quality control in the amplification, DNA separation and allele detection process to include appropriate controls.
h) Storage, preservation, and retention of amplified DNA product according to laboratory policy.
i) Troubleshooting…

21

21

OSAC Human Forensic Biology Subcommittee: ASB Training Standards on pathway to the Registry

Lets look at 3 ASB DNA Training Standards
that are still in the OSAC Registry Approval Process
(as of January 14, 2022)

22

ANSI/ASB 130, Standard for Training in Forensic DNA Amplification Methods for Subsequent Capillary Electrophoresis Sequencing, First Edition, 2021

1 Scope:
This standard provides the general requirements for a forensic DNA laboratory's training program in forensic **DNA amplification methods** for subsequent capillary electrophoresis (CE) sequencing. This standard applies to forensic human and wildlife mitochondrial DNA amplification, and wildlife nuclear DNA amplification.

ANSI/ASB Standard 130, First Edition 2021

Standard for Training in Forensic DNA Amplification Methods for Subsequent Capillary Electrophoresis Sequencing

Not on OSAC Registry - Yet

23

23

ANSI/ASB 131, Standard for Training in Forensic DNA Sequencing using Capillary Electrophoresis, First Edition, 2021

1 Scope:
This standard provides the general requirements for a forensic DNA laboratory's training program in forensic DNA sequencing using capillary electrophoresis. This standard applies to forensic human and wildlife mitochondrial DNA capillary electrophoresis sequencing, and wildlife nuclear DNA capillary electrophoresis sequencing.

ANSI/ASB Standard 131, First Edition 2021

Standard for Training in Forensic DNA Sequencing using Capillary Electrophoresis

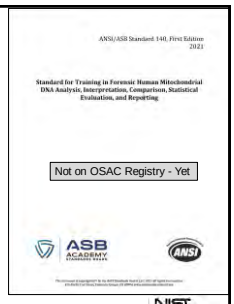
Not on OSAC Registry - Yet

24

24

ANSI/ASB 140, Standard for Training in Forensic Human Mitochondrial DNA Analysis, Interpretation, Comparison, Statistical Evaluation, and Reporting, First Edition 2021

1 Scope:
This document provides the requirements for a forensic DNA laboratory's training program in forensic human mitochondrial DNA (mtDNA) analysis, interpretation, comparison, statistical evaluation, and reporting.



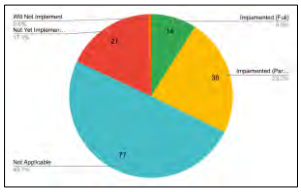
Not on OSAC Registry - Yet

25

25

OSAC Implementation Survey Between June –August 2021: 155 Responses

ANSI/ASB 22 Standard for Forensic DNA Analysis Training Programs



50 Implementors Full & Partial

77 Respondents Indicated Not Applicable

26

26

4 Part Webinar Series: DNA Standards and Best Practices Developed by OSAC & ASB

Part 1: The Process (July 15, 2020) – covers OSAC & ASB

Part 2: Mixture Interpretation Validation, and Protocol Development and Verification (August 5, 2020) – covers ASB 20 & ASB 40

Part 3: Training Standards Overview (September 9, 2020) – coverers ASB 22

Part 4: ASNI/ASB Standard 018, Standard for Validation of Probabilistic Genotyping Systems, First Edition, 2020 (January 20, 2021)

Promega Webinars

https://www.promega.com/resources/webinars/

27

27

Factsheets & Audit Checklists Under Development for Select Standards on the OSAC Registry

- NIST entered into a cooperative agreement with AAFS to develop training, tools, and resources to enhance implementation efforts and broaden awareness of forensic science standards among communities of interest.
- Resources, including auditing checklists for compliance monitoring and gap analysis, as well as Factsheets, understandable to the lay person.

https://www.aafs.org/article/american-academy-forensic-sciences-aafs-awarded-cooperative-agreement-national-institute

28

28

Free Access to ASTM E30 Committee on Forensic Science Standards (approx. 60)

The following **30,000 public criminal justice agencies** receive access to the ASTM Committee E30 on Forensic Science Standards:

- Organization of Scientific Area Committee Members & Affiliates – approximately 750 individuals
- NIST and Federal/State/Local Crime Laboratories – approximately 412 labs
- Public Defenders Offices – approximately 1,000 offices
- Law Enforcement Agencies – approximately 18,000 offices
- Prosecutor Offices – approximately 3,000 offices
- Medical Examiner/Coroners Offices – approximately 3,000 office

ASTM Standards Access

https://www.nist.gov/topics/forensic-science/astm-standards-access

29

29

Stay Informed!

Website: www.nist.gov/osac

OSAC BULLETIN STANDARDS

- Provides monthly updates on forensic science standards moving through development process at SDOs and those moving through OSAC Registry process
- https://www.nist.gov/osac/osac-standards-bulletin

OSAC NEWSLETTER

- Quarterly communication that provides updates on OSAC's program status, activities, accomplishments, and opportunities for public input with internal and external audiences.
- https://www.nist.gov/osac/osac-newsletter

- Follow us! https://www.linkedin.com/showcase/organization-of-scientific-area-committees-osac-for-forensic-science/

30

30

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

STRBase Updates

National Institute of Standards and Technology (NIST)

Peter M. Vallone

Module 10

1

Outline - STRBase Updates

- Current STRBase site
  - https://strbase.nist.gov/
- Migration and Redesign (beta site)
  - https://strbase-b.nist.gov/
- For today: a tour through updated pages

Lisa Borsuk

Primary curator of STRBase

Feedback
strbase@NIST.gov

2

Serving the community since 1997

Over 350k pageviews per year

STR typing Community Resources for:
- STR markers
- Reporting of variant alleles
- Various forensic project pages
- NIST workshops
- NIST team software tools
- Publications
- and more…

3

Drivers for change

- As the site has grown → challenge to keep all materials updated
  - John Butler's role has changed at NIST - less time for manual curation
  - Hyperlinks change (broken links)
  - Certain materials need to be updated
  - Complex design

100s of html files
1000s of documents

- IT technology and support have moved forward
  - More efficient ways to do things (store data, template pages, interactivity)
  - Cannot keep legacy servers running indefinitely
- STRBase → STRBase (slight update) → **STRBase 2.0**

Editing current content
Minor IT upgrades

Redesign interface
Inventory information

4

STRBase 2.0

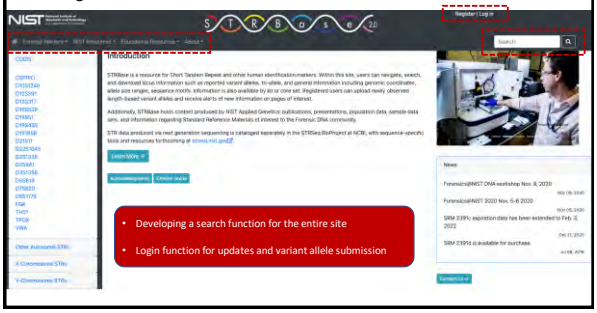
Top 5 (excluding index page)

1. STR Fact sheets
2. FBI core loci
3. Variant allele reports
4. Multiplex STR kits
5. Y-Chromosome STRs

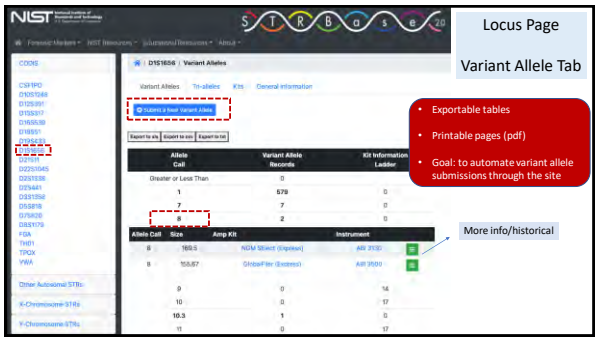
Planning and Triage
- Understand what is highly accessed
- Move away from 'single author' oversight
- Evaluate content to be carried over into a new site (prioritizing)
- Improve and add new pages – make the content more 'stand alone'
- Ideas to improve design interface (searchable, navigation, exportable)
- Make easier to update and maintain
  - **Database concept (central storage)**
  - **MySQL - Server (.NET framework)**
  - Example: edit the name of a kit once versus on all pages

5

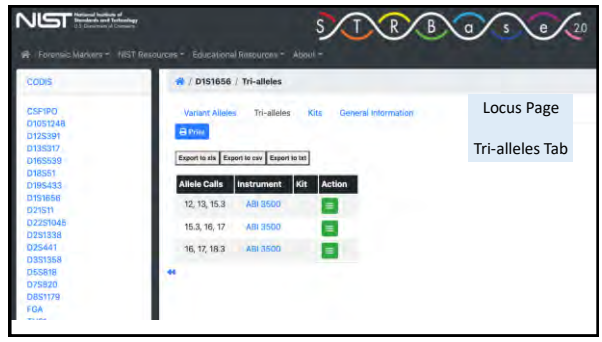
Home Page for STRBase 2.0



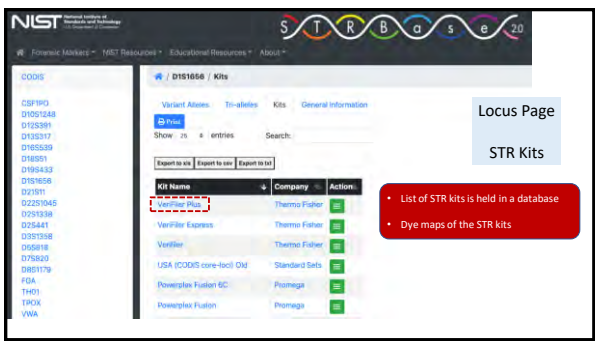
6


Locus Page
Variant Allele Tab
Exportable tables
Printable pages (pdf)
Goal: to automate variant allele submissions through the site
More info/historical

7


Locus Page
Tri-alleles Tab

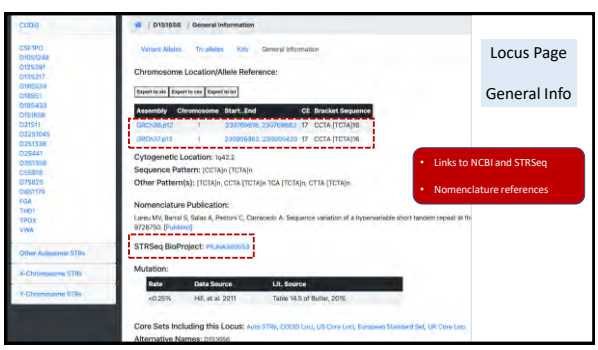
8


Locus Page
STR Kits
List of STR kits is held in a database
Dye maps of the STR kits

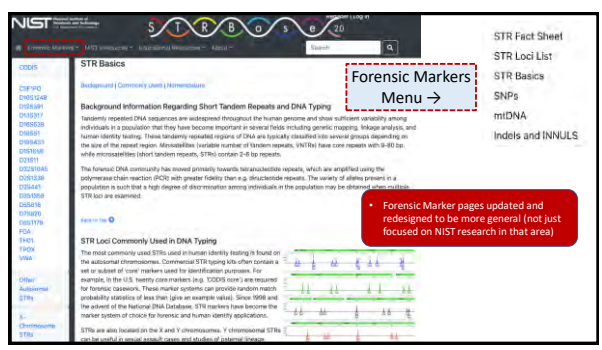
9


Locus Page
STR Kits

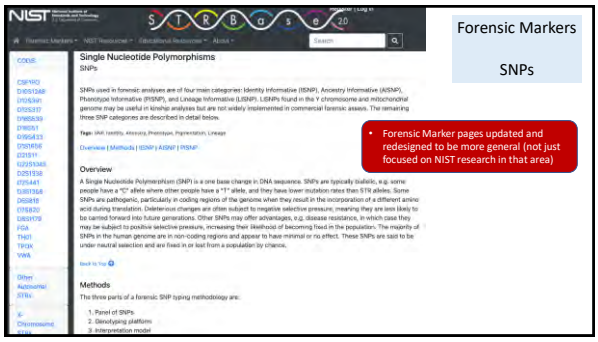
10


Locus Page
General Info
Links to NCBI and STRSeq
Nomenclature references

11

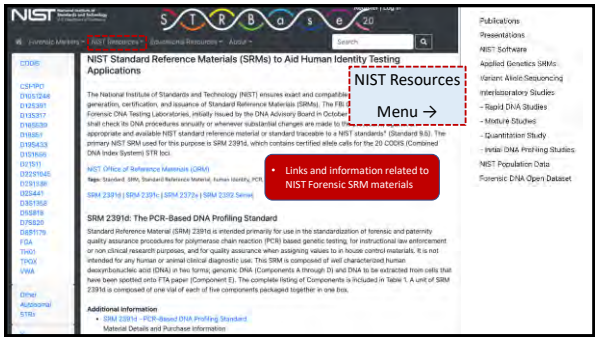

STR Fact Sheet
STR Loci List
STR Basics
SNPs
mtDNA
Indels and INNULS
Forensic Markers Menu →
Forensic Marker pages updated and redesigned to be more general (not just focused on NIST research in that area)

12

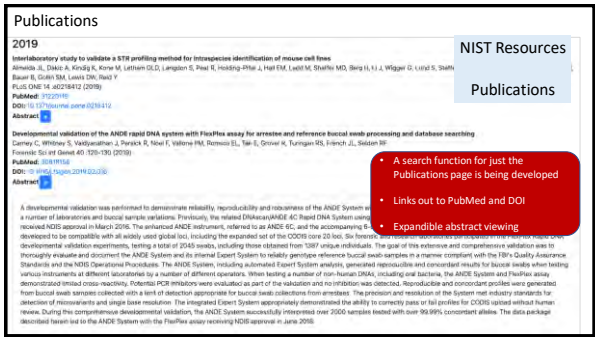


13

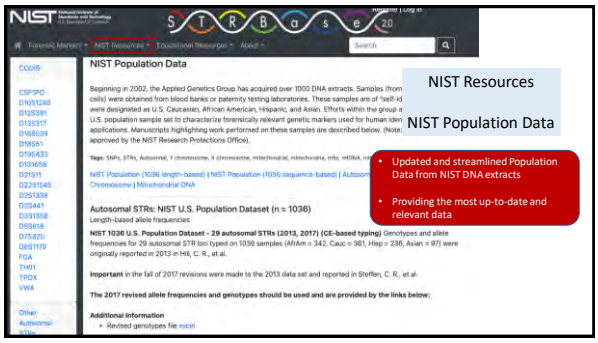
14



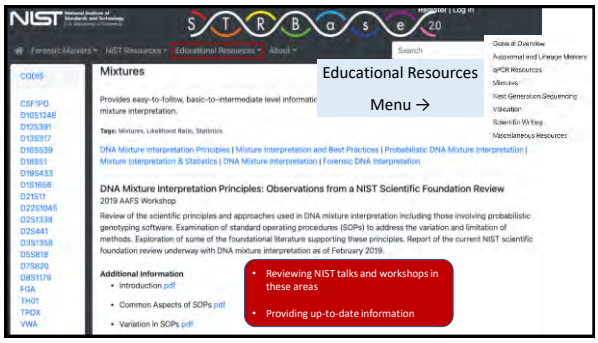
15



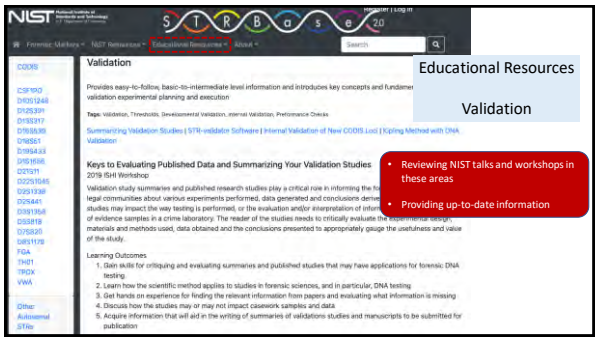
16



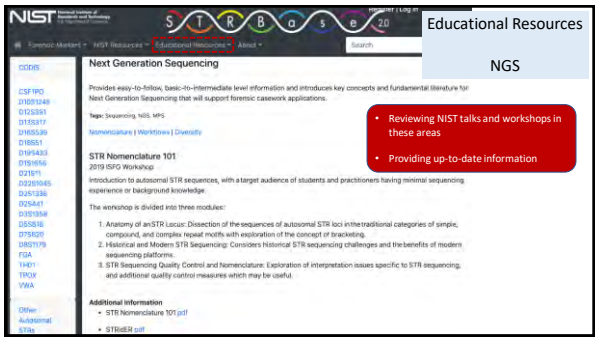
17



18



19



20

Future and Feedback

- Working on a login interface (login.gov account) for update notices and submission of variant alleles
- Update the beta site as stable builds are implemented
- Data has been migrated – fine tuning how it is presented
- **Feedback and suggestions: strbase@NIST.gov**

21

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

DNA Most Valuable Publications List

National Institute of Standards and Technology (NIST)

John M. Butler
Special Programs Office

Module 11

1

Experts Need Up-to-Date Knowledge in Their Field

Dr. Gillian Tully, the UK Forensic Science Regulator at the time, stated in her 2017 annual report:

> "It is a clear expectation of the courts that expert evidence is presented by people who are indeed experts in their field. This necessitates **an up-to-date knowledge of developments in the relevant field**, which in turn necessitates access to scientific literature and sufficient time to ensure that each expert has the current relevant knowledge that they need."

https://www.gov.uk/government/publications/forensic-science-regulator-annual-report-2017
(published January 19, 2018, quote from page 10)

2

American Academy of Forensic Sciences
VIRTUAL WORKSHOP W19 (MVPs of Forensic DNA)
February 16, 2021

2021 AAFS VIRTUAL MEETING

Last Year's Half-Day Workshop

MVPs of Forensic DNA:

Examining the Most Valuable Publications in the Field

Chair **John M. Butler**
Co-Chair **Robin W. Cotton**
Mechthild K. Prinz
Charlotte J. Word

3

Slide from Robin Cotton (AAFS 2021 W19, Module 2): **Value of a Knowledge Base for Educating Students and Practitioners**

Requirement for Reading the Literature from the FBI DNA Quality Assurance Standards (2020)

STANDARD 16.1 The laboratory shall have and follow a program to ensure technical qualifications are maintained through **participation in continuing education.**

16.1.1 …analyst(s)…shall stay abreast of topics relevant to the field of forensic DNA analysis **by attending seminars…in relevant subject areas for a minimum of eight (8) cumulative hours each calendar year.**

16.1.2 The laboratory shall have and follow a program approved by the technical leader for the annual review of scientific literature that documents the analysts' **ongoing reading of scientific literature.**

16.1.2.1 The laboratory shall maintain or have physical or electronic access to a collection of current books, reviewed journals, or other literature applicable to DNA analysis.

Current QAS (2020) – available on FBI website (approved January 11, 2018): https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf/view

4

Slide from Robin Cotton (AAFS 2021 W19, Module 2): **Value of a Knowledge Base for Educating Students and Practitioners**

Challenges the Forensic DNA Community Faces with Continuing Education

- **QAS requirement for continuing education are only a start**
  - Minimum of eight (8) hours per year for seminars and **one** (1) or more articles to read will not cover much ground
  - **How does anyone know if you learned anything since there is no assessment of what was learned?**
  - **For example, which articles are essential for you to understand and will expand your expertise in DNA mixture interpretation?**
- **Rapid and continuous evolution of the field**
  - New STR kits, new CE instruments, new software, new potential approaches for analysis (e.g., NGS) and interpretation (e.g., probabilistic genotyping software)
  - **There are lots of articles to chose from based on interest or need…**
- **Numerous articles are being published each year**
  - **Which articles should you choose to study?**

5

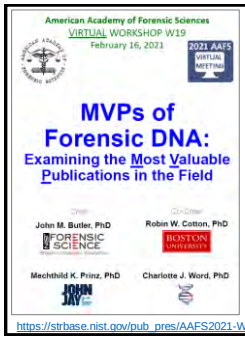
Development of Expert Knowledge

DNA analysts benefit from at least three different levels of expert knowledge:

1. **Education in basic science** covering biochemistry, biology, chemistry, genetics, molecular biology, population genetics, and statistics
2. **Training in forensic science** and specific methods and protocols used in their laboratory to develop competency needed to perform casework
3. **Continued education and professional development** to keep up-to-date as the field evolves and new methods become available

#3 involves knowing the ever-growing scientific literature

6



AAFS 2021 MVPs Workshop

- **4-hour virtual workshop** with pre-recorded videos from each presenter
- **116-page handout** (available on STRBase)
  - **240 slides and a 33-page reference list**
- **480 references in 26 categories + 17 books**
  - Built on lists from SWGDAM July 2020 Training Guidelines and OSAC October 2020 Informative Literature for Forensic Biology and DNA
  - Valuable input from co-presenters is gratefully acknowledged: Robin Cotton, Mecki Prinz, and Charlotte Word
- Discussed principles, MVP trends (article citations, journals used), and **value of the #1 article in each category**
  - **Covered 17 of 26 categories in 10 modules**
  - Also discussed training standards and value of a knowledge base for forensic DNA analysts

https://strbase.nist.gov/pub_pres/AAFS2021-W19-Handouts.pdf

7

The Ultimate Goal

Creation of a defined body of knowledge covering historical and foundational literature ***that qualified DNA analysts should know and understand***

8

Some Improvements That Could Be Beneficial to the Forensic DNA Community

1. **An agreed upon, defined body of knowledge** for DNA mixture interpretation and a means to update and remove outdated information as methods evolve
2. **Access to appropriate relevant literature** for technical leaders and analysts
3. **Dedicated time in the workday to read the literature** so that technical leaders and analysts can keep up-to-date with developments
4. **Uniformly documented knowledge assessment**
5. **A method to acknowledge competence in a specific area** to allow true expertise in testimony (e.g., DNA transfer and activity assessments, see van Oorschot et al. 2019)
6. **Additional training for technical leaders in experimental design and data analysis** to assist with validation studies and protocol development

The workshop last year was intended as a start

From deliberations and discussions of NIST team members and Resource Group in connection with the Scientific Foundation Review on DNA Mixture Interpretation

9

INTERPOL Review of Forensic Biology and DNA

- Conducted every three years by invited author(s)
  - 17th INTERPOL Review on DNA (2010-2013) examined 114 articles (Jolicoeur 2013)
  - 18th INTERPOL Review on DNA (2013-2016) examined 75 articles (Laurent & Pene 2016)
- **INTERPOL Review (2016-2019)**
  - 19th INTERPOL Review on DNA examined 235 references (John Butler & Sheila Willis)
  - Presentation: https://strbase.nist.gov/pub_pres/Butler-INTERPOL-DNAreview-Oct2019.pdf
  - Publication: Butler JM, Willis S (2020) INTERPOL review of forensic biology and forensic DNA typing 2016-2019. *Forensic Sci. Int.: Synergy* 2:352-367
- **INTERPOL Review (2019-2022)**
  - 20th INTERPOL Review on DNA - Underway – completed by July 2022
  - Presented in October 2022 and published in early 2023

10

The 2016-2019 INTERPOL DNA Review

Discussed 235 references from 35 journals

Review of forensic biology and DNA publications from 2016 to 2019

- Category selection and article selection:
  1. Core Loci Expansion
  2. Rapid Analysis of STR Markers
  3. Investigative Genetic Genealogy
  4. Next-Generation Sequencing
  5. DNA Mixture Interpretation and Probabilistic Genotyping Software
  6. DNA Transfer and Activity Level Evaluations
  7. Forensic Biology and Body Fluid Identification
  8. DNA Phenotyping
  9. Privacy and Ethical Issues
  10. Guidance Documents (SWGDAM, OSAC, ASB, ENFSI, UK Regulator)
  11. Contamination Avoidance and DNA Success Rates
  12. Recent Special Issues and Review Articles of Note

https://www.interpol.int/content/download/14458/file/Interpol%20Review%20Papers%202019.pdf

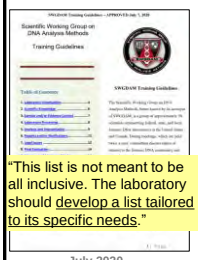
11

Origin of Most Valuable Publication (MVP) List

- AAFS 2016 Half-Day Workshop on Forensic Literature
  - See https://strbase.nist.gov/training/AAFS2016_LiteratureWorkshop.htm
- NIST Scientific Foundation Review on DNA Mixture Interpretation
  - Literature gathered from summer 2017 to 2021; draft report released in June 2021
  - Involved examining >1,000 articles on DNA (528 references were cited in draft report)
- INTERPOL Review of Forensic Biology and DNA (2016-2019)
  - Completed in July 2019 and published in February 2020 (*FSI Synergy* 2:352-367)
  - Reported on 12 topics from 235 articles across 35 journals
  - Invited to complete (2019-2022) review by July 2022
- SWGDAM Training Guidelines (published in July 2020)
  - **129 references** in five categories + 6 websites
- OSAC Efforts on Foundational/Informative Literature (September/October 2020)
  - JMB approached in September 2020 to assist their effort (Phil Danielson and a team of six others)
  - Recommended changing from "Books", "Reviews", "Salient Papers" to **Books + 26 categories**
  - See Informative Literature for Forensic Biology and DNA (10-26-20 has **448 references**)

12

Latest SWGDAM Training Guidelines (July 2020)



July 2020

"This list is not meant to be all inclusive. The laboratory should develop a list tailored to its specific needs."

Recommended References (129 + 6 websites)

The following resources may be helpful to the trainer in defining the breadth and scope of the materials for the trainee's reading. This list is not meant to be all inclusive. The laboratory should develop a list tailored to its specific needs.

1. General Forensic DNA and Autosomal STRs (**42**)
2. Mixture Interpretation/Population Genetics/ Probabilistic Genotyping/Statistics (**40**)
3. Mitochondrial DNA (**37**)
   - General Mitochondrial DNA Information (**6**)
   - Heteroplasmy (**15**)
   - Maternal Inheritance (**1**)
   - Population Studies (**1**)
4. Y STRs (**10**)
5. Informational Websites (**6**)

The previous 2013 version listed **98 references and the same 6 websites** (most of the additions were in mixture interpretation and probabilistic genotyping)

13

Origins of Our Initial MVP Literature List

- On September 10, 2020, Phil Danielson (University of Denver), representing a team of seven OSAC members compiling foundational literature, reached out to me and shared their list for my input (**it had 105 references in three categories** + possible additions) :
  - **5 "foundational" textbooks**,
  - **41 "foundational" reviews** (subtopics: field of forensic sciences in general, serology, collection and storage of biological material, epigenetics, DNA quantification, PCR process, trace/touch type DNA, advanced and emerging DNA profiling technologies, mitochondrial DNA haplotyping, DNA profile interpretation, presenting forensic DNA in the courtroom, and non-human DNA analysis)
  - **59 salient research studies** (subtopics: serology, human factors, DNA extraction/purification, DNA quantification, DNA profiling and validation, mtDNA haplotyping, probabilistic genotyping, presenting DNA in the courtroom, and validation software)

I examined these references along with those in the SWGDAM 2020 Training Guidelines, created a more comprehensive set of categories (from A-to-Z), added many new references, created uniform reference formatting, and changed the titles to "informative textbooks" and "informative forensic DNA reviews and research studies" – this updated information was returned to Phil Danielson ***on September 24, 2020***

14

Additional Input to 2021 MVP Reference List

- **Discussion with fellow presenters as presentations developed**
  - Mecki Prinz, Robin Cotton, Charlotte Word
- Examination of updated OSAC 10-26-2020 list
  - Phil Danielson and six other OSAC members
  - Included additional PGS, DNA transfer, and non-human DNA articles
- Feedback from Other Practitioners and Educators
  - **Amy Brodeur** (Boston University) – serology & body fluid ID, collection & storage
  - **Teresa Cheromcha** (Colorado Bureau of Investigation) – DNA transfer

15

How We Examined MVPs in the 2021 Workshop

1. Discuss important principles involved with the category topic (e.g., DNA extraction or PCR amplification)
2. In each examined category, briefly review the number and types of articles in our reference list and number of times cited in Google Scholar (as of January 2021)
3. Focus on one or a few specific articles and the findings reported
4. Summarize and review key takeaways

16

From AAFS 2021 Workshop | Article titles are available at https://strbase.nist.gov/pub_pres/AAFS2021-W19-Handouts.pdf

#1 MVP(s) on PGS

L1. Coble, M.D. and Bright, J.-A. (2019) Probabilistic genotyping software: An overview. *Forensic Science International: Genetics* 38: 219-224.

Google Scholar
Cited 40 times
(8 Jan 2021)

Google Scholar
Cited 80 times
(10 Jan 2022)

- **Why is this article valuable?**
  - Provides a historical perspective and overview on the movement from binary methods of interpretation to probabilistic methods of interpretation

17

From AAFS 2021 Workshop | Article titles are available at https://strbase.nist.gov/pub_pres/AAFS2021-W19-Handouts.pdf

Interpretation: Probabilistic Genotyping Software - PGS (Discrete, Continuous)

(Category L – 44 articles)

- **Reviews:**
  - DNA Commission on allele drop-out/in (L11)
  - #1 PGS overview and history (L1)
  - Comparison of statistical models (L17)
  - Historical: 20 years of R&D (L26)
  - Paradigm shift (L34)
  - Statistical evaluation of forensic evidence (L24)
- **Validation**:
  - Definitions and recommended tests (L25, L32)
  - EuroForMix studies (L29, L39)
  - TrueAllele studies (L5, L8, L27, L28, L43)
  - STRmix theory and studies (L14, L18, L19, L20, L21, L30, L33, L36, L42, L44)
  - STRmix collaborative exercise (L40)
  - Comparing model performance (L22, L29, L38, L41)
- **Continuous Models:**
  - Early work (L3, L4, L6, L7, L9, L10)
  - Modeling stutter (L13, L15)
  - Low template profiles (L23, L31)
  - Information gain from peak heights (L37)
- **Likelihood Ratios:**
  - Framework for addressing questions (L2)
  - Exploring nondonor distributions (L12, L44)
  - Calibration and method validation (L16, L35)

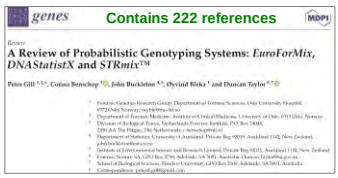
2022 MVP List reduced to 4 articles
(2021 L1, L26, L32) + new 2021 article

18

2022 MVP L4.

Available [open access] at https://www.mdpi.com/2073-4425/12/10/1559

Gill et al. (2021) A review of probabilistic genotyping systems: *EuroForMix, DNAStatistX* and *STRmix*™. *Genes* 12: 1559


genes

Contains 222 references

MDPI

Review

A Review of Probabilistic Genotyping Systems: *EuroForMix, DNAStatistX* and *STRmix*™

Peter Gill [1,2,*], Corina Benschop [3], John Buckleton [4,5], Øyvind Bleka [1] and Duncan Taylor [6,7]

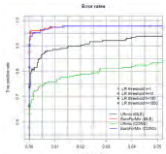


Figure 1: ROC plot comparing PGS methods

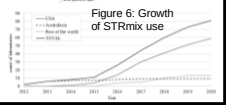


Figure 6: Growth of STRmix use

Describes historical development of PGS and general principles of interpretation as well as the evolution, utility, practice, and adoption of these software programs

19

Thoughts and Observations on the Literature

1. New articles and advances are regularly being published
   - Keep an open mind and remember that science is open-ended
2. Limitations of some publications
   - Claims made do not always correspond to available data
   - We need to encourage more data sharing in publications (as supplemental files)
3. The community seems to make more use of articles on methodology as compared to interpretation
   - For example, Goggle Scholar found fewer citations to PGS articles than to PCR articles (in part because PGS efforts are more recent)
4. Training is challenging as there is simply too much to know in a constantly evolving field
   - Suggestion by Robin Cotton in AAFS 2021 workshop that an analyst learns to think through what is happening to DNA molecules at each step of the process

20

Different Types of Articles

- **Original research articles**
- **Review articles**
- Short communications (termed "technical notes" in *JFS*)
- Book reviews
- Case studies (termed "case reports" in *JFS*)
- Opinion or commentary
- Letters to the Editor
  - typically correcting or commenting on a previous publication
- With *FSI Genetics*: Forensic population genetics (original paper, short communication, or correspondence)

Different journals can have different categories and/or required structures for manuscript submission

https://www.elsevier.com/journals/forensic-science-international-genetics/1872-4973/guide-for-authors

21

Special Issue (*Manfred Kayser, editor*): Trends and Perspectives in Forensic Genetics

https://www.sciencedirect.com/journal/forensic-science-international-genetics/special-issue/10TSDS4360H

12 articles published (2018-2019), topics covered include:

- Introduction to special issue
- Forensic epigenetics
- Y-chromosome match probabilities
- MtGenome search algorithm
- Microhaplotypes
- **DNA transfer in forensic science** (W1)
- **Activity level propositions** (W4)
- Next- to now-generation sequencing
- Postmortem interview using microbes
- HID microbiome markers
- ICMP experience with large scale HID
- **Probabilistic genotyping software** (L1)

22

GCAT TACG GCAT genes

an Open Access Journal by MDPI

Special Issue (Emiliano Giardina, *editor*): Forensic Genetics and Genomics

https://www.mdpi.com/journal/genes/special_issues/Forensic_Genetic

12 articles published (2020-2021), topics covered include:

- Special issue overview
- Nanopore sequencing of STRs
- Ancestry informative SNPs
- Rapid DNA (ANDE 6C)
- Interpreting mixtures with GlobalFiler
- Human skin microbial profiling
- Y-STR allele frequency differences between populations
- **STRidER 2-year QC report** (Y4)
- Databanking in Malaysia
- Chinese population with joint mtDNA and Y-chromosome
- Chinese population with Y-STRs and SNPs
- Chinese population with microhaplotypes

23

Special Issue (*Manfred Kayser, editor*): Forensic Genetics: Unde venisti et quo vadis?

Where do you come from and where are you going?

https://www.sciencedirect.com/journal/forensic-science-international-genetics/special-issue/10D6PT650B2

9 articles published (2021-2022), topics covered include:

- Environmental trace evidence
- Germlines of monozygotic twins
- Forensic transcriptome analysis
- Capture enrichment and MPS
- **Investigative genetic genealogy** (ADDED in 2022)
- Interpreting NUMTs
- Forensic proteomics
- Forensic bone analysis
- Human microbiome

24

genes

an Open Access Journal by MDPI

Special Issue (Niels Morling, editor): Advances in Forensic Genetics

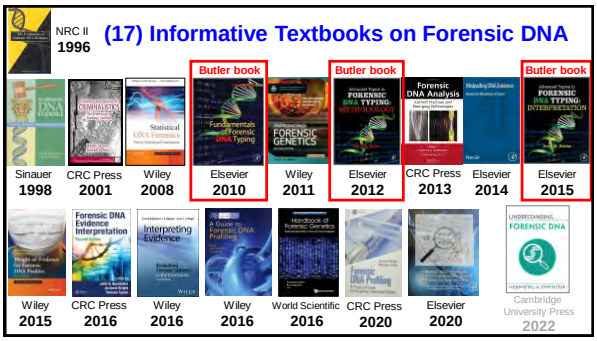
https://www.mdpi.com/journal/genes/special_issues/Advances_Forensic_Genetics

24 articles published (2021-2022), topics covered include:

- **PGS Review: EuroForMix, DNAStatistX, STRmix**
- OpenArray for forensic phenotyping
- Skin pigmentation and genetic ancestry
- Eye color prediction
- Ancestry informative markers (VISAGE)
- Single cell analysis for forensic phenotyping
- Animal forensic genetics
- Predicting visible traits in dogs (CaDNAP)
- Single cell analysis for mixture interpretation
- New STR panel for cross-species bird DNA
- Ancient DNA methods improve Korean/WW2 IDs
- **DNA transfer review and recent progress**
- Bayesian Networks for DNA transfer questions
- SNP markers for investigative genetic genealogy (FORCE panel)
- **DNA sampling in burglary investigations**
- Body fluid ID and tissues
- Microbiome analysis
- **Software options for forensic sequencing**
- ChrY and mtDNA statistics/assessment
- Ethical decision-making as lived practice
- DNA from aged rootless hair shafts in Romanov relics

25

(17) Informative Textbooks on Forensic DNA



26

Reference List Provided with Slide Handouts

480 (2021) → 85 (2022) References

Informative Forensic DNA Reviews & Research Studies (A-to-Z categories)

In our reference list, 26 categories are defined covering topics of interest in forensic DNA analysis and interpretation (listed arbitrarily from A to Z).

Neither the categories nor this reference list are intended to be exhaustive. A much larger list (480 references) was created originally – see https://strbase.nist.gov/pub_pres/AAFS2021-W19-Handouts.pdf. Suggestions for additional, appropriate references and categories are welcome.

A #1 article (in bold font) was subjectively selected in each category and then followed by reference citations defined by date in ascending order with the most recent publications at the end of each category. This letter and number system (e.g., A1, B2, F3) provides a simple method to locate specific articles and enables opportunities for expansion as the literature grows. Although some articles could logically appear under multiple categories, no duplicate listings were used. ***Many recommended references from the SWGDAM 2020 Training Guidelines have been included as well.***

27

Informative Forensic DNA Reviews and Research Studies (A-to-Z)

Category Group	Topic(s) Covered	# Articles 480 (2021)	# Articles 85 (2022)
A	Plain Language Guides to Forensic DNA Analysis	4	2
B	Serology and Body Fluid Identification	24	3
C	Collection and Storage of Biological Material	25	2
D	DNA Extraction/Purification, Differential Extraction	18	2
E	DNA Quantitation, Degraded DNA	10	2
F	PCR Amplification, Inhibition, and Artifacts	13	3
G	Capillary Electrophoresis Separation and Detection	12	2
H	Assessing Sample Suitability & Complexity, Low-Template	7	2
I	Estimating the Number of Contributors	12	4
J	Data Interpretation, Mixture Deconvolution, Interlab Studies	12	4
K	Interpretation: Binary Approaches (CPI, RMP, LR)	11	5
L	Interpretation: Probabilistic Genotyping Software	44	4
M	Report Writing and Technical Review	8	4

28

Informative Forensic DNA Reviews and Research Studies (A-to-Z)

Category Group	Topic(s) Covered	# Articles 480 (2021)	# Articles 85 (2022)
N	Court Testimony, Communication, Juror Comprehension	22	5
O	Autosomal STR Markers and Kits	29	2
P	Mitochondrial DNA Testing	11	3
Q	Y-Chromosome and X-Chromosome Testing	17	4
R	DNA Databases and Investigative Genetic Genealogy	14	3
S	Statistical Analysis	11	2
T	Population Genetics	11	2
U	DNA Phenotyping (Ancestry, Appearance, Age)	24	2
V	New Technologies (Rapid DNA, Massively Parallel Sequencing)	35	5
W	DNA Transfer and Activity Level Reporting	57	8
X	Non-Human DNA Testing	15	2
Y	Method Validation, Quality Control, and Human Factors	23	5
Z	General Forensic Science Topics	11	3

29

Category W:

DNA Transfer and Activity Level Reporting

1. **van Oorschot, R.A.H., Szkuta, B., Meakin, G.E., Kookshoorn, B., Goray, M. (2019) DNA transfer in forensic science: a review.** ***Forensic Science International: Genetics*** **38: 140-166.**
2. Taylor, D., Abarno, D., Rowe, E., Rask-Nielsen, L. (2016) Observations of DNA transfer within an operational Forensic Biology Laboratory. *Forensic Science International: Genetics* 23: 33-49.
3. Kokshoorn, B., Blankers, B.J., de Zoete, J., Berger, C.E.H. (2017) Activity level DNA evidence evaluation: On propositions addressing the actor or the activity. *Forensic Science International* 278: 115-124.
4. Taylor, D., Kokshoorn, B. and Biedermann, A. (2018) Evaluation of forensic genetics findings given activity level propositions: A review. *Forensic Science International: Genetics* 36: 34-49.
5. Burrill, J., Daniel, B., Frascione, N. (2019) A review of trace "touch DNA" deposits: Variability factors and an exploration of cellular composition. *Forensic Science International: Genetics* 39:8-18.
6. Gosch, A. and Courts, C. (2019) On DNA transfer: the lack and difficulty of systematic research and how to do it better. *Forensic Science International: Genetics* 40: 24-36.
7. Gosch, A., Euteneuer, J., Preuss-Wossner, J., Courts, C. (2020) DNA transfer to firearms in alternative realistic handling scenarios. *Forensic Science International: Genetics* 48: 102355.
8. van Oorschot, R.A.H., Meakin, G.E., Kookshoorn, B., Goray, M., Szkuta, B. (2021) DNA transfer in forensic science: recent progress towards meeting challenges. *Genes* 12: 1766. Available [open access] at https://www.mdpi.com/2073-4425/12/11/1766.

30

W8. van Oorschot et al. (2021) DNA transfer in forensic science: recent progress towards meeting challenges. *Genes* 12: 1766.

Abstract: Understanding the factors that may impact the transfer, persistence, prevalence and recovery of DNA (DNA-TPPR), and the availability of data to assign probabilities to DNA quantities and profile types being obtained given particular scenarios and circumstances, is paramount when performing, and giving guidance on, evaluations of DNA findings given activity level propositions (activity level evaluations). In late 2018 and early 2019, three major reviews were published on aspects of DNA-TPPR, with each advocating the need for further research and other actions to support the conduct of DNA-related activity level evaluations. **Here, we look at how challenges are being met, primarily by providing a synopsis of DNA-TPPR-related articles published since the conduct of these reviews and briefly exploring some of the actions taken by industry stakeholders towards addressing identified gaps.** Much has been carried out in recent years, and efforts continue, to meet the challenges to continually improve the capacity of forensic experts to provide the guidance sought by the judiciary with respect to the transfer of DNA.

Available [open access] at https://www.mdpi.com/2073-4425/12/11/1766.

31

Reference Lists Compared

		MVPs Feb 2021	OSAC 10-26-20	SWGDAM July 2020
		2022		
	Informative Textbooks on Forensic DNA	**17**	**17**	**16**	**5 + 2** (6 websites)
A	Plain Language Guides to Forensic DNA Analysis	2	4	3	--
B	Serology and Body Fluid Identification	3	24	15 + 2	--
C	Collection and Storage of Biological Material	2	25	19	--
D	DNA Extraction/Purification, Differential Extraction	2	18	14	1
E	DNA Quantitation, Degraded DNA	2	10	9 + 1	1
F	PCR Amplification, Inhibition, and Artifacts	3	13	10	3
G	Capillary Electrophoresis Separation and Detection	2	12	12	6
H	Assessing Sample Suitability and Complexity, Low-Template DNA	2	7	8	--
I	Estimating the Number of Contributors	4	12	12	--
J	Data Interpretation, Mixture Deconvolution, Interlaboratory Studies	4	12	12	2 + 4
K	Interpretation: Binary Approaches (CPI, RMP, LR)	5	11	9	3
L	Interpretation: Probabilistic Genotyping Software (Discrete, Continuous)	4	44	41	7 + 11
M	Report Writing and Technical Review	4	8	8	--
N	Court Testimony, Communication of Results, Juror Comprehension	5	22	21	3
O	Autosomal STR Markers and Kits	2	29	27	4
P	Mitochondrial DNA Testing	3	11	10 + 1	3 + 32
Q	Y-Chromosome and X-Chromosome Testing	4	17	11	4 + 6
R	DNA Databases and Investigative Genetic Genealogy	3	14	14	--
S	Statistical Analysis	2	11	9	3 + 2
T	Population Genetics	2	11	10	1
U	DNA Phenotyping (Ancestry, Appearance, Age)	2	24	20	--
V	New Technologies (Rapid DNA, Massively Parallel Sequencing)	5	35	31	--
W	DNA Transfer and Activity Level Reporting	8	57	54	--
X	Non-Human DNA Testing	2	15	15	--
Y	Method Validation, Quality Control, and Human Factors	5	23	23	1 + 5
Z	General Forensic Science Topics	3	11	11	1
	Historical				19
	TOTAL	**85**	**497**	**448**	**135**

Different Reference Lists

- **(497)** AAFS 2021 Most Valuable Publications Workshop
- **(448)** OSAC 10-26-20 version
- **(135)** SWGDAM Training Guidelines

- OSAC list shares a common origin with our MVP list
  - some differences exist
- SWGDAM 2020 Training Guidelines Reference List
  - Historical references (19)
  - More mtDNA articles (>30)
  - No coverage of DNA transfer and many other potentially valuable topics

Underlined numbers reflect those found only in that list

3 articles in common with our MVP list + 32 mtDNA articles only in the SWGDAM list

SWGDAM includes 19 articles I have classified as "historical"

32

A1. (Category A) **Plain Language Guides to Forensic DNA Analysis**

***Making Sense of Forensic Genetics* (2017)**

concepts clearly explained in 40 pages

- Developed by European Forensic Genetics Network of Excellence (EuroForGen-NoE) and **published with Sense about Science**
- **Free PDF file** available for download https://senseaboutscience.org/wp-content/uploads/2017/01/making-sense-of-forensic-genetics.pdf
- *Final point made*: "As DNA profiling continues to grow more sensitive, and it is used in more investigations, **the need for accurate communication between scientists and nonscientists only grows** - both **to ensure that their expectations of the technology are realistic, and its limits are properly understood**…"

Translated into Spanish, Polish, Portuguese (so far) with support from the ISFG

33

ISFG DNA Commission Articles

- Several of the #1 MVPs are ISFG DNA Commission articles:

DNA mixture interpretation
- **K1.** Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.R., Morling, N., Prinz, M., Schneider, P.M. and Weir, B.S. (2006) DNA Commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International* 160: 90-101.

mtDNA
- **P1.** Parson, W., Gusmão, L., Hares, D.R., Irwin, J.A., Mayr, W.R., Morling, N., Pokorak, E., Prinz, M., Salas, A., Schneider, P.M., Parsons, T.J. (2014) DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics* 13: 134-142.

Y-STRs
- **Q1.** Roewer, L., Andersen, M.M., Ballantyne, J., Butler, J.M., Caliebe, A., Corach, D., D'Amato, M.E., Gusmão, L., Hou, Y., de Knijff, P., Parson, W., Prinz, M., Schneider, P.M., Taylor, D., Vennemann, M., Willuweit, S. (2020) DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis. *Forensic Science International: Genetics* 48: 102308.

Non-human DNA testing
- **X1.** Linacre, A., Gusmão, L., Hecht, W., Hellmann, A.P., Mayr, W.R., Parson, W., Prinz, M., Schneider, P.M., Morling, N. (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. *Forensic Science International: Genetics* 5(5): 501-505.

- These are freely available on the ISFG website:
  - https://www.isfg.org/Publications/DNA+Commission

ISFG International Society for Forensic Genetics

34

Opportunities to Get Electronic Access to Journals

MVP 2022 List

37 of 85 (44%) from *FSI Genetics*

6 of 85 (7%) from *JFS*

ISFG membership includes free access to the print and online editions of Forensic Science International: Genetics. Please log in to read and download articles via the section reserved for #members. ISFG members have also access to the workshop presentations and lectures of invited speakers at the most recent ISFG congresses.

FSI GENETICS

ISFG International Society for Forensic Genetics

https://www.isfg.org/

60 euros (~$68.20/year)

ELSEVIER

https://www.aafs.org/

Elsevier Forensics Package ($133/year) includes electronic access to AAFS members **($165/year)**: *Journal of Forensic Sciences*
- *Forensic Science International*
- *Forensic Science International: Genetics*
- *Journal of Forensic and Legal Medicine and Legal Medicine*
- *Legal Medicine*
- *Science & Justice*

#1 Journal on Forensic DNA

[open access]

35

Some Final Thoughts

1. No selection criteria or reference list will be perfect or complete
   - continuing research and future review articles add knowledge to our field
   - some references could be removed to focus content in various categories
2. We would love to hear your ideas on how to best maintain an updated list to benefit the community
   - Are there other categories that should be included in MVP lists?
3. How could a national/international MVP list benefit future training?
   - Would it be worth conducting an ASCLD or AAFS survey on this topic?
   - If we understand the need, then we can lay the groundwork for future possibilities in funding
   - Funding would need to be continuing and sustained to be effective (not year-to-year) – would forensic laboratories support a subscription fee of some kind to have access to all the articles?

36

Informative Forensic DNA Reviews and Research Studies (A to Z) (85)

Below, 26 categories cover topics of interest in forensic DNA analysis and interpretation (listed arbitrarily from A to Z). Neither the categories nor this reference list are intended to be exhaustive. A much larger list (480 references) was created originally – see https://strbase.nist.gov/pub_pres/AAFS2021-W19-Handouts.pdf. Suggestions for additional, appropriate references and categories are welcome.

A #1 article (in bold font) was subjectively selected in each category and then followed by reference citations defined by date in ascending order with the most recent publications at the end of each category. This letter and number system (e.g., A1, B2, F3) provides a simple method to locate specific articles and enables opportunities for expansion as the literature grows. Although some articles could logically appear under multiple categories, no duplicate listings were used. Many recommended references from the SWGDAM 2020 Training Guidelines have been included as well.

A. Plain Language Guides to Forensic DNA Analysis

1. **Sense about Science (2017) *Making Sense of Forensic Genetics*. A 40-page plain language guide available at https://senseaboutscience.org/activities/making-sense-of-forensic-genetics/.**

2. The Royal Society (2017) *Forensic DNA Analysis: A Primer for Courts*. A 60-page plain language guide available at https://royalsociety.org/-/media/about-us/programmes/science-and-law/royal-society-forensic-dna-analysis-primer-for-courts.pdf.

B. Serology and Body Fluid Identification

1. **Gaensslen, R.E. (1983) *Sourcebook in Forensic Serology, Immunology, and Biochemistry*. U.S. Department of Justice, National Institute of Justice: Washington, D.C.**

2. Desroches, A.N., Buckle, J.L., Fourney, R.M. (2009) Forensic biology evidence screening: past and present. *Canadian Society of Forensic Science Journal* 42(2): 101-120.

3. Sijen, T. (2015) Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers. *Forensic Science International: Genetics* 18: 21-32.

C. Collection and Storage of Biological Material

1. **Mapes, A.A., Kloosterman, A.D., van Marion, V., de Poot, C.J. (2016) Knowledge on DNA success rates to optimize the DNA analysis process: from crime scene to laboratory. *Journal of Forensic Sciences* 61(4): 1055-1061.**

2. Hedman, J., Jansson, L., Akel, Y., Wallmark, N., Gutierrez Liljestrand, R., Forsberg, C., Ansell, R. (2020) The double-swab technique versus single swabs for human DNA recovery from various surfaces. *Forensic Science International: Genetics* 46: 102253.

D. DNA Extraction/Purification, Differential Extraction

1. **Gill, P., Jeffreys, A.J., Werrett, D.J. (1985) Forensic application of DNA 'fingerprints'. *Nature* 318: 577-579.**

2. Samie, L., Champod, C., Glutz, V., Garcia, M., Castella, V., Taroni F. (2019) The efficiency of DNA extraction kit and the efficiency of recovery techniques to release DNA using flow cytometry. *Science & Justice* 59(4): 405-410.

E. DNA Quantitation, Degraded DNA

1. **Grgicak, C.M., Urban, Z.M., Cotton, R.W. (2010) Investigation of reproducibility and error associated with qPCR methods using Quantifiler® Duo DNA quantification kit. *Journal of Forensic Sciences* 55(5):1331-1339.**

2. Lee, S.B., McCord, B., Buel, E. (2014) Advances in forensic DNA quantification: a review. *Electrophoresis* 35: 3044-3052.

F. PCR Amplification, Inhibition, and Artifacts

1. **Walsh, P.S., Erlich, H.A. and Higuchi, R. (1992) Preferential PCR amplification of alleles: mechanisms and solutions. *PCR Methods & Applications* 1(4): 241-250.**

2. Alaeddini, R. (2012) Forensic implications of PCR inhibition—A review. *Forensic Science International: Genetics* 6(3): 297-305.

3. Cavanaugh, S.E. and Bathrick, A.S. (2018) Direct PCR amplification of forensic touch and other challenging DNA samples: A review. *Forensic Science International: Genetics* 32: 40-49.

G. Capillary Electrophoresis Separation and Detection

1. **Butler, J.M., Buel, E., Crivellente, F., McCord, B.R. (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 Genetic Analyzers for STR analysis. *Electrophoresis* 25: 1397-1412.**

2. Rakay, C.A., Bregu, J. and Grgicak, C.M. (2012) Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out. *Forensic Science International: Genetics* 6(6): 723-728.

H. Assessing Sample Suitability and Complexity, Low-Template DNA

1. **Gill, P., Whitaker, J., Flaxman, C., Brown, N., Buckleton, J. (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International* 112(1): 17-40.**

2. Benschop, C.C., van der Beek, C.P., Meiland, H.C., van Gorp, A.G., Westen, A.A. and Sijen, T. (2011) Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results. *Forensic Science International: Genetics* 5(4): 316-328.

I. Estimating the Number of Contributors

1. **Buckleton, J.S., Curran, J.M. and Gill, P. (2007) Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics* 1(1): 20-28.**

2. Coble, M.D., Bright, J.A., Buckleton, J.S. and Curran, J.M. (2015) Uncertainty in the number of contributors in the proposed new CODIS set. *Forensic Science International: Genetics* 19: 207-211.

3. Norsworthy, S., Lun, D.S., Grgicak, C.M. (2018) Determining the number of contributors to DNA mixtures in the low-template regime: Exploring the impacts of sampling and detection effects. *Legal Medicine* 32: 1-8.

4. Marciano, M.A. and Adelman, J.D. (2019) Developmental validation of PACE™: Automated artifact identification and contributor estimation for use with GlobalFiler™ and PowerPlex® Fusion 6C generated data. *Forensic Science International: Genetics* 43: 102140.

J. Data Interpretation, Mixture Deconvolution, Interlaboratory Studies

1. **Gill, P., Sparkes, R. and Kimpton, C. (1997) Development of guidelines to designate alleles using an STR multiplex system. *Forensic Science International* 89(3): 185-197.**

2. Clayton, T.M., Whitaker, J.P., Sparkes, R. and Gill, P. (1998) Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International* 91(1): 55-70.

3. Butler, J.M., Kline, M.C. and Coble, M.D. (2018) NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): variation observed and lessons learned. *Forensic Science International: Genetics* 37: 81-94.

4. Lynch, P.C. and Cotton, R.W. (2018) Determination of the possible number of genotypes which can contribute to DNA mixtures: non-computer assisted deconvolution should not be attempted for greater than two person mixtures. *Forensic Science International: Genetics* 37: 235-240.

K. Interpretation: Binary Approaches (CPI, RMP, LR)

1. **Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.R., Morling, N., Prinz, M., Schneider, P.M. and Weir, B.S. (2006) DNA Commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International* 160: 90-101.**

2. Buckleton, J. and Curran, J. (2008) A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics* 2(4): 343-348.

3. Schneider, P.M., Fimmers, R., Keil, W., Molsberger, G., Patzelt, D., Pflug, W., Rothämel, T., Schmitter, H., Schneider, H. and Brinkmann, B. (2009) The German Stain Commission: recommendations for the interpretation of mixed stains. *International Journal of Legal Medicine* 123(1): 1-5. [Originally published in German in *Rechtsmedizin* (2006) 16: 401-404].

4. Budowle, B., Onorato, A.J., Callaghan, T.F., Della, M.A., Gross, A.M., Guerrieri, R.A., Luttman, J.C., McClure, D.L. (2009) Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *Journal of Forensic Sciences* 54(4): 810-821.

5. Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., Coble, M.D. (2016) Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genetics* 17(1):125.

L. Interpretation: Probabilistic Genotyping Software (Discrete, Continuous)

1. **Coble, M.D. and Bright, J.-A. (2019) Probabilistic genotyping software: An overview. *Forensic Science International: Genetics* 38: 219-224.**

2. Gill, P., Haned, H., Bleka, O., Hansson, O., Dørum, G. and Egeland, T. (2015) Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-Twenty years of research and development. *Forensic Science International: Genetics* 18: 100-117.

3. Haned, H., Gill, P., Lohmueller, K., Inman, K., Rudin, N. (2016) Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations. *Science & Justice* 56(2): 104-108.

4. Gill, P., Benschop, C., Buckleton, J., Bleka, O., Taylor, D. (2021) A review of probabilistic genotyping systems: *EuroForMix*, *DNAStatistX* and *STRmix™*. *Genes* 12:1559. Available at https://www.mdpi.com/2073-4425/12/10/1559.

M. Report Writing and Technical Review

1. **Association of Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice* 49: 161-164.**

2. Cook, R., Evett, I.W., Jackson, G., Jones, P.J. and Lambert, J.A. (1998) A model for case assessment and interpretation. *Science & Justice* 38(3): 151-156.

3. Cook, R., Evett, I.W., Jackson, G., Jones, P.J. and Lambert, J.A. (1998) A hierarchy of propositions: deciding which level to address in casework. *Science & Justice* 38(4): 231-239.

4. Ballantyne, K.N., Edmond, G. and Found, B. (2017) Peer review in forensic science. *Forensic Science International* 277: 66-76.

N. Court Testimony, Communication of Results, Juror Comprehension Studies

1. **Eldridge, H. (2019) Juror comprehension of forensic expert testimony: a literature review and gap analysis. *Forensic Science International: Synergy* 1: 24-34.**

2. Howes, L.M., Kirkbride, K.P., Kelty, S.F., Julian, R., Kemp, N. (2013) Forensic scientists' conclusions: how readable are they for non-scientist report-users? *Forensic Science International* 231: 102-112.

3. Taroni, F., Biedermann, A., Vuille, J., and Morling, N. (2013). Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Science International: Genetics* 7: 467-470.

4. Gill, P., Hicks, T., Butler, J.M., Connolly, E., Gusmão, L., Kokshoorn, B., Morling, N., van Oorschot, R.A.H., Parson, W., Prinz, M., Schneider, P.M., Sijen, T. and Taylor, D. (2018) DNA Commission of the International Society for Forensic Genetics: Assessing the value of forensic biological evidence – guidelines highlighting the importance of propositions. Part I: Evaluations of DNA profiling comparisons given (sub-) source propositions. *Forensic Science International: Genetics* 36: 189-202.

5. Gill, P., Hicks, T., Butler, J.M., Connolly, E., Gusmão, L., Kokshoorn, B., Morling, N., van Oorschot, R.A.H., Parson, W., Prinz, M., Schneider, P.M., Sijen, T. and Taylor, D. (2020) DNA Commission of the International Society for Forensic Genetics: Assessing the value of forensic biological evidence - guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositions. *Forensic Science International: Genetics* 44: 102186.

O. Autosomal STR Markers and Kits

1. **Butler, J.M. (2006) Genetics and genomics of core STR loci used in human identity testing. *Journal of Forensic Sciences* 51(2): 253-265.**

2. Gettings, K.B., Aponte, R.A., Vallone, P.M., Butler, J.M. (2015) STR allele sequence variation: current knowledge and future issues. *Forensic Science International: Genetics* 18: 118-130.

P. Mitochondrial DNA Testing

1. **Parson, W., Gusmão, L., Hares, D.R., Irwin, J.A., Mayr, W.R., Morling, N., Pokorak, E., Prinz, M., Salas, A., Schneider, P.M., Parsons, T.J. (2014) DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics* 13: 134-142.**

2. Budowle, B., Allard, M.W., Wilson, M.R., Chakraborty, R. (2003) Forensics and mitochondrial DNA: Applications, debates, and foundations. *Annual Review of Genomics and Human Genetics* 4: 119-141.

3. Melton, T. (2004) Mitochondrial DNA heteroplasmy. *Forensic Science Review* 16: 2-19.

Q. Y-Chromosome and X-Chromosome Testing

1. **Roewer, L., Andersen, M.M., Ballantyne, J., Butler, J.M., Caliebe, A., Corach, D., D'Amato, M.E., Gusmão, L., Hou, Y., de Knijff, P., Parson, W., Prinz, M., Schneider, P.M., Taylor, D., Vennemann, M., Willuweit, S. (2020) DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis. *Forensic Science International: Genetics* 48: 102308.**

2. Tillmar, A.O., Kling, D., Butler, J.M., Parson, W., Prinz, M., Schneider, P.M., Egeland, T., Gusmão, L. (2017) DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Science International: Genetics* 29: 269-275.

3. Caliebe, A. and Krawczak, M. (2018) Match probabilities for Y-chromosomal profiles: a paradigm shift. *Forensic Science International: Genetics* 37: 200-203.

4. Gomes, I., Pinto, N., Antão-Sousa, S., Gomes, V., Gusmão, L., Amorim, A. (2020). Twenty years later: A comprehensive review of the X chromosome use in forensic genetics. *Frontiers in Genetics* 11: 926.

R. DNA Databases and Investigative Genetic Genealogy

1. **Struyf, P., De Moor, S., Vandeviver, C., Renard, B., van der Beken, T. (2019) The effectiveness of DNA databases in relation to their purpose and content: A systematic review. *Forensic Science International* 301: 371-381.**

2. Greytak, E.M., Moore, C., Armentrout, S.L. (2019) Genetic genealogy for cold case and active investigations. *Forensic Science International* 299: 103-113.

3. Kling, D., Phillips, C., Kennett, D., Tillmar, A. (2021) Investigative genetic genealogy: current methods, knowledge and practice. *Forensic Science International: Genetics* 52: 102474.

S. Statistical Analysis

1. **Curran, J.M. (2013) Is forensic science the last bastion of resistance against statistics? *Science & Justice* 53: 251-252.**

2. Puch-Solis, R., Roberts, P., Pope, S. and Aitken, C. (2012) Practitioner Guide No. 2. *Assessing the Probative Value of DNA Evidence: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. Royal Statistical Society's Working Group on Statistics and the Law. Available at https://www.maths.ed.ac.uk/~cgga/Guide-2-WEB.pdf.

T. Population Genetics

1. **Balding, D.J. and Nichols, R.A. (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* 64: 125-140.**

2. Buckleton, J., Curran, J., Goudet, J., Taylor, D., Thiery, A., Weir, B.S. (2016) Population-specific FST values for forensic STR markers: A worldwide survey. *Forensic Science International: Genetics* 23: 91-100.

U. DNA Phenotyping (Ancestry, Appearance, Age)

1. **Kayser, M. (2015) Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Science International: Genetics* 18: 33-48.** See also https://www.visage-h2020.eu/#publications.

2. Phillips, C. (2015) Forensic genetic analysis of bio-geographical ancestry. *Forensic Science International: Genetics* 18: 49-65.

V. New Technologies (Rapid DNA, Massively Parallel Sequencing)

1. **Butler, J.M. (2015) The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* 370: 20140252.**

2. Mapes, A.A., Kloosterman, A.D., de Poot, C.J., van Marion, V. (2016) Objective data on DNA success rates can aid the selection process of crime samples for analysis by rapid mobile DNA technologies. *Forensic Science International* 264: 28-33.

3. Phillips, C., Gettings, K.B., King, J.L., Ballard, D., Bodner, M., Borsuk, L., Parson, W. (2018) "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR sequence guide. *Forensic Science International: Genetics* 34: 162-169.

4. Romsos, E.L., French, J.L., Smith, M., Figarelli, V., Harran, F., Vandegrift, G., Moreno, L.I., Callaghan, T.F., Brocato, J., Vaidyanathan, J., Pedroso, J.C., Amy, A., Stoiloff, S., Morillo, V.H., Czetyrko, K., Johnson, E.D., de Tagyos, J., Murray, A., Vallone, P.M. (2020) Results of the 2018 Rapid DNA Maturity Assessment. *Journal of Forensic Sciences* 65(3): 953-959.

5. Ballard, D., Winkler-Galicki, J., Wesoly, J. (2020) Massive parallel sequencing in forensics : advantages, issues, technicalities, and prospects. *International Journal of Legal Medicine* 134: 1292-1303.

W. DNA Transfer and Activity Level Reporting

1. **van Oorschot, R.A.H., Szkuta, B., Meakin, G.E., Kookshoorn, B., Goray, M. (2019) DNA transfer in forensic science: a review. *Forensic Science International: Genetics* 38: 140-166.**

2. Taylor, D., Abarno, D., Rowe, E., Rask-Nielsen, L. (2016) Observations of DNA transfer within an operational Forensic Biology Laboratory. *Forensic Science International: Genetics* 23: 33-49.

3. Kokshoorn, B., Blankers, B.J., de Zoete, J., Berger, C.E.H. (2017) Activity level DNA evidence evaluation: On propositions addressing the actor or the activity. *Forensic Science International* 278: 115-124.

4. Taylor, D., Kokshoorn, B. and Biedermann, A. (2018) Evaluation of forensic genetics findings given activity level propositions: A review. *Forensic Science International: Genetics* 36: 34-49.

5. Burrill, J., Daniel, B., Frascione, N. (2019) A review of trace "touch DNA" deposits: Variability factors and an exploration of cellular composition. *Forensic Science International: Genetics* 39:8-18.

6. Gosch, A. and Courts, C. (2019) On DNA transfer: the lack and difficulty of systematic research and how to do it better. *Forensic Science International: Genetics* 40: 24-36.

7. Gosch, A., Euteneuer, J., Preuss-Wossner, J., Courts, C. (2020) DNA transfer to firearms in alternative realistic handling scenarios. *Forensic Science International: Genetics* 48: 102355.

8. van Oorschot, R.A.H., Meakin, G.E., Kookshoorn, B., Goray, M., Szkuta, B. (2021) DNA transfer in forensic science: recent progress towards meeting challenges. *Genes* 12: 1766. Available at https://www.mdpi.com/2073-4425/12/11/1766.

X. Non-Human DNA Testing

1. **Linacre, A., Gusmão, L., Hecht, W., Hellmann, A.P., Mayr, W.R., Parson, W., Prinz, M., Schneider, P.M., Morling, N. (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. *Forensic Science International: Genetics* 5(5): 501-505.**

2. Ogden, R. and Linacre, A. (2015) Wildlife forensic science: A review of genetic geographic origin assignment. *Forensic Science International: Genetics* 18: 152-159.

Y. Method Validation, Quality Control, and Human Factors

1. **Kloosterman, A., Sjerps, M., & Quak, A. (2014) Error rates in forensic DNA analysis: Definition, numbers, impact and communication. *Forensic Science International: Genetics* 12: 77-85.**

2. Budowle, B., Bottrell, M.C., Bunch, S.G., Fram, R., Harrison, D., Meagher, S., Oien, C.T., Peterson, P.E., Seiger, D.P., Smith, M.B., Smrz, M.A., Soltis, G.L., Stacey, R.B. (2009) A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *Journal of Forensic Sciences* 54(4): 798-809.

3. Basset, P. and Castella, V. (2018) Lessons from a study of DNA contaminations from police services and forensic laboratories in Switzerland. *Forensic Science International: Genetics* 33: 147-154.

4. Bodner, M. and Parson, W. (2020) The STRidER report on two years of quality control of autosomal STR population datasets. *Genes (Basel)* 11(8): 901.

5. Dror, I.E. (2020) Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. *Analytical Chemistry* 92(12): 7998-8004.

Z. General Forensic Science Topics

1. **National Academy of Sciences, Committee on the Conduct of Science (1989) On being a scientist. *Proceedings of the National Academy of Sciences of the United States of America* 86(23): 9053-9074.**

2. Mnookin, J.L., Cole, S.A., Dror, I.E., Fisher, B.A.J., Houck, M.M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N., Siegel, J., Stoney, D.A. (2010). The need for a research culture in the forensic sciences. *UCLA Law Review* 58: 725-779.

3. National Commission on Forensic Science (2017) *Reflecting Back – Looking Toward the Future.* Available at https://www.justice.gov/archives/ncfs/page/file/959356/download.

Informative Textbooks on Forensic DNA (17)

The following informative textbooks are listed by publication date in ascending order with the most recent ones listed last. This list is not comprehensive (e.g., earlier editions of some of these textbooks not included).

1. National Research Council (1996) *The Evaluation of Forensic DNA Evidence*. National Academy Press: Washington, D.C.

2. Evett, I.W. and Weir, B.S. (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer Associates: Sunderland MA.

3. Inman, K. and Rudin, N. (2001) *Principles and Practice of Criminalistics: The Profession of Forensic Science*. CRC Press: Boca Raton.

4. Fung, W.K. and Hu, Y.-Q. (2008) *Statistical DNA Forensics: Theory, Methods and Computation*. Wiley: Chichester, UK.

5. Butler, J.M. (2010) *Fundamentals of Forensic DNA Typing*. Elsevier Academic Press: San Diego.

6. Goodwin, W., Linacre, A., Hadi, S. (2011) *An Introduction to Forensic Genetics Second Edition*. Wiley: Chichester, UK.

7. Butler, J.M. (2012) *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Academic Press: San Diego.

8. Shewale, J.G. and Liu, R.H. (Editors) (2013) *Forensic DNA Analysis: Current Practices and Emerging Technologies.* CRC Press: Boca Raton.

9. Gill, P. (2014) *Misleading DNA Evidence: Reasons for Miscarriages of Justice*. Elsevier Academic Press: San Diego.

10. Butler, J.M. (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*. Elsevier Academic Press: San Diego.

11. Balding, D. J. and Steele, C. D. (2015). *Weight-of-evidence for Forensic DNA Profiles Second Edition*. Wiley: Chichester, UK.

12. Buckleton, J.S., Bright, J.-A., Taylor, D. (Editors) (2016) *Forensic DNA Evidence Interpretation Second Edition*. CRC Press: Boca Raton.

13. Robertson, B., Vignaux, G.A., Berger, C.E.H. (2016) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom Second Edition*. Wiley: Chichester, UK.

14. Jamieson, A. and Bader, S. (Editors) (2016) *A Guide to Forensic DNA Profiling*. Wiley: Chichester, UK.

15. Amorim, A. and Budowle, B. (Editors) (2017) *Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation*. World Scientific Publishing: London.

16. Bright, J.-A. and Coble, M. (2020) *Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios*. CRC Press: Boca Raton.

17. Gill, P., Bleka, Ø., Hansson, O., Benschop, C., Haned, H. (2020) *Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles*. Elsevier Academic Press: San Diego.

American Academy of Forensic Sciences
HYBRID WORKSHOP W2 (NIST Forensic DNA)
February 21, 2022

Wrap-Up and Workshop Conclusions

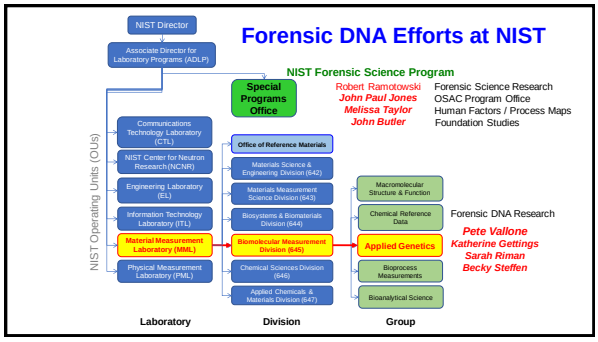
National Institute of Standards and Technology (NIST)

John M. Butler
Special Programs Office

Module 13

1

2



3

Clarification on What NIST Is and Is Not

- NIST is a Federal government **science agency** and does not comment on legal admissibility
- NIST is **not a regulatory agency**, which is why key takeaways are provided in our draft report rather than formal recommendations
- NIST **focuses on research and assisting with developing standards** (e.g., OSAC or SRMs); NIST does not conduct forensic science casework

4

NIST FORENSIC SCIENCE
RESEARCH. STANDARDS. FOUNDATIONS.

NIST Research Efforts to Aid Forensic Science

SPO **SEVEN CURRENT FOCUS AREAS**

PML ***1. Ballistics and Associated Tool Marks***
ITL ***2. Digital and Identification Forensics***
MML ***3. Forensic Genetics (DNA)***
MML ***4. Drugs & Toxins***
MML ***5. Trace Evidence***
ITL ***6. Statistics***
ITL ***7. Biometrics***

Beyond these internal NIST research program efforts, NIST funds a Forensic Science Center of Excellence (CSAFE)

https://www.nist.gov/forensic-science

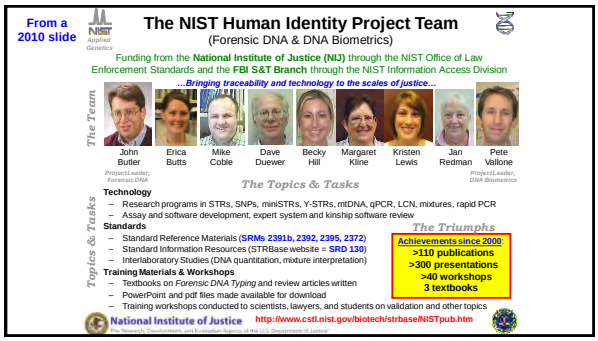
5

Scope of Our Work: A Mission Statement from 2010

Stitching on a custom polo shirt

- The NIST Human Identity Project Team is trying **to lead the way in forensic DNA**… through research that helps bring traceability and technology to the scales of justice.

6



7

8

Margaret Kline (1954-2021)

From https://inet.nist.gov/nist-connections/obituary-margaret-kline

For many years, Margaret was the face, the heart, and the soul of forensic DNA efforts at NIST.

Margaret Kline, a groundbreaking DNA researcher, **passed away on Oct. 4, 2021**, after a long battle with cancer. **Having retired from NIST in November 2020, she was posthumously inducted into the Gallery of Distinguished Scientists, Engineers, and Administrators — NIST's hall of fame — on Oct. 22, 2021.** Margaret was 66 years old when she died, and she had worked at NIST for 35 years.

Her career at NIST was a model of integrity, hard work, and dedication to advancing the science of forensic DNA storage and testing. Her research greatly improved DNA forensic standards, influenced many other fields, and saved the U.S. government many millions of dollars. "**She brought metrology to DNA forensics**," says Peter Vallone, leader of the Applied Genetics Group in MML's Biomolecular Measurement Division.

9

Members of the NIST Biotechnology Group

Summer 1990

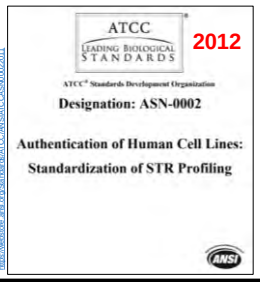
Photo courtesy of Dennis Reeder

10

A Tribute Celebration for Margaret When She Retired in 2020

11

Impactful Documents that Margaret Influenced



12

Some of Margaret's Articles That Have Directly Influenced DNA Measurements Worldwide

TECHNICAL NOTE — *J. Forensic Sci.* 42: 897-906

Margaret C. Kline,[1] M.S.; David L. Duewer,[2] Ph.D.; Pamela Newall,[3] M.A.; Janette W. Redman[1]; Dennis J. Reeder,[1] Ph.D.; and Melanie Richard,[1] M.Sc.

Interlaboratory Evaluation of Short Tandem Repeat Triplex CTT*

These results directly led to the use of STR allelic ladders in all commercial STR kits – a practice that continues today almost 25 years later!

"The most reliable current mechanism for interlaboratory exchange of STR results is the qualitative allelic name."

Anal. Chem. **2003**, *75*, 2463–2469

NIST Mixed Stain Study 3: DNA Quantitation Accuracy and Its Influence on Short Tandem Repeat Multiplex Signal Intensity

Margaret C. Kline, David L. Duewer,* Janette W. Redman, and John M. Butler

Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

TECHNICAL NOTE — *J Forensic Sci*, May 2005, Vol. 50, No. 3 Paper ID JFS2004357 Available online at: www.astm.org

Margaret C. Kline,[1] M.S.; David L. Duewer,[2] Ph.D.; Janette W. Redman[1]; and John M. Butler,[1] Ph.D.

Results from the NIST 2004 DNA Quantitation Study*

"Information from this interlaboratory study is guiding development of a future NIST Standard Reference Material for Human DNA Quantitation, SRM 2372…"

13

Over Her Impactful Scientific Career at NIST, Margaret has Influenced Many Communities

- Biological Evidence Management and Preservation
- Clinical Genetics
- Cell Line Authentication
- Digital PCR
- Documentary Standards Development
- Forensic DNA
- Forensic Science
- Genetic Genealogy
- NIST Standard Reference Materials
- Quality Assurance and Proficiency Testing

Margaret was inducted into the NIST Hall of Fame Portrait Gallery in November 2021

14

Summary

- NIST efforts in forensic DNA research and standards development are significant today and have been for multiple decades – ***thanks to excellent staff, visiting scientists, and many collaborators***
  1. Ongoing, impactful research and physical standards development by the Applied Genetics Group for >30 years
  2. Administration of OSAC to facilitate development and implementation of documentary standards in forensic DNA laboratories
  3. Identification of valuable principles and areas for improvement with recent scientific foundation review of DNA mixture interpretation (draft report is being finalized with input from extensive public comment)

15

Thank you for your attention!

John M. Butler john.butler@nist.gov

Peter M. Vallone peter.vallone@nist.gov

John Paul Jones john.jones@nist.gov

Katherine B. Gettings katherine.gettings@nist.gov

Melissa K. Taylor melissa.taylor@nist.gov

Carolyn R. (Becky) Steffen becky.steffen@nist.gov

Sarah Riman sarah.riman@nist.gov

16

PLOS ONE

RESEARCH ARTICLE

Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset

Sarah Riman[1]*, **Hari Iyer**[2], **Peter M. Vallone**[1]

1 Applied Genetics Group, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America, **2** Statistical Design, Analysis, Modeling Group, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America

* sarah.riman@nist.gov

OPEN ACCESS

Citation: Riman S, Iyer H, Vallone PM (2021) Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset. PLoS ONE 16(9): e0256714. https://doi.org/10.1371/journal.pone.0256714

Editor: Usman Qamar, National University of Sciences and Technology (NUST), PAKISTAN

Received: June 16, 2021

Accepted: August 7, 2021

Published: September 17, 2021

Data Availability Statement: The data are held in a public repository: https://doi.org/10.1101/2021.05.26.445891. All other relevant data are within the paper and its Supporting Information files.

Funding: Yes-This work was funded by NIST Special Programs Office: Forensic Genetics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no conflict of interest.

Abstract

A likelihood ratio (LR) system is defined as the entire pipeline of the measurement and interpretation processes where probabilistic genotyping software (PGS) is a piece of the whole LR system. To gain understanding on how two LR systems perform, a total of 154 two-person, 147 three-person, and 127 four-person mixture profiles of varying DNA quality, DNA quantity, and mixture ratios were obtained from the filtered (.CSV) files of the GlobalFiler 29 cycles 15s PROVEDIt dataset and deconvolved in two independently developed fully continuous programs, STRmix v2.6 and EuroForMix v2.1.0. Various parameters were set in each software and LR computations obtained from the two software were based on same/fixed EPG features, same pair of propositions, number of contributors, theta, and population allele frequencies. The ability of each LR system to discriminate between contributor (H1-true) and non-contributor (H2-true) scenarios was evaluated qualitatively and quantitatively. Differences in the numeric LR values and their corresponding verbal classifications between the two LR systems were compared. The magnitude of the differences in the assigned LRs and the potential explanations for the observed differences greater than or equal to 3 on the $\log_{10}$ scale were described. Cases of LR < 1 for H1-true tests and LR > 1 for H2-true tests were also discussed. Our intent is to demonstrate the value of using a publicly available ground truth known mixture dataset to assess discrimination performance of any LR system and show the steps used to understand similarities and differences between different LR systems. We share our observations with the forensic community and describe how examining more than one PGS with similar discrimination power can be beneficial, help analysts compare interpretation especially with low-template profiles or minor contributor cases, and be a potential additional diagnostic check even if software in use does contain certain diagnostic statistics as part of the output.

1. Introduction

Fully continuous probabilistic genotyping software (PGS) uses computer algorithms and complex calculations to apply biological, statistical, and mathematical models to resolve genotypes of contributors or assign evidential weight for the DNA typing results [1–4]. These models, unlike binary and semi-continuous models, use quantitative information contained within a profile (e.g. allelic designations, peak heights, molecular weights/fragment length), take into account stochastic effects, model peak height variability, and allow interpretation of low-level and complex DNA mixtures, therein reducing the need to infer using subjective reasoning [3, 5–10].

Numerous commercial [11–16] and open-source [17–21] software and freeware [22, 23] packages implementing fully continuous models have been developed. Differences exist among the programs in the way they model the distribution of allelic peak heights, stutter artifacts, mixture ratios, degradation, and stochastic events [8, 9, 24–28], though all use the same underlying genetic, physical, and chemical principles.

Most PGS require the assignment of two propositions, the prosecution proposition (H1) and defense proposition (H2) that include the specification of the number of contributors. Other parameters specific for each PGS are also required to deliver a key output, a Bayes factor, commonly referred to as the likelihood ratio (LR) [29, 30]. LR is the strength of the evidence in favor of H1 relative to H2. It is expressed as the ratio of two probabilities:

$$\boldsymbol{LR}=\frac{\boldsymbol{Pr}(\boldsymbol{E}|\boldsymbol{H1},\boldsymbol{I})}{\boldsymbol{Pr}(\boldsymbol{E}|\boldsymbol{H2},\boldsymbol{I})}$$

where E is the findings or evidence and I is the relevant background information. The numerator is the probability of the findings given that H1 and background information are true and the denominator is the probability of the findings given that H2 and background information are true [31, 32].

So far there is no consensus within the forensic DNA community on implementing a standardized fully continuous PGS [33, 34]. As a result, depending on the software being used, the interpretation of the same DNA profile could yield different numeric LR values and, if used, different verbal characterizations [34, 35]. Even if the same PGS is used, the overall LR system could be different and hence will lead to different LRs [36–38]. Few studies explored the question of the degree of variability in LR values across various fully continuous PGS [39–42]. These studies were based on limited number of samples, did not quantify the differences in LRs, and concluded that the models yielded similar LRs despite the differences among the PG modeling assumptions. Only in [34], the authors demonstrated the impact of inter-model variability on numerical values and verbal expression of the LRs of H1 true cases when four variant models of the same continuous software, CEESIt, were compared.

To further understand the amount of variability expected when mixtures are interpreted using different systems, we here performed large-scale comparison and assessed the LR values produced by two reputable and well-cited fully continuous PG models. For this illustration, we chose STRmix v2.6 (a commercial software that uses the Bayesian approach) and EuroForMix v2.1.0 (an open-source software that uses the maximum likelihood estimation (MLE) method) [11, 17]. A large dataset of ground truth known 2-person, 3-person, and 4-person mixture profiles was selected from the publicly available PROVEDIt database [43, 44]. We first investigated the discriminating power of the two LR systems using Receiver Operating Characteristic (ROC) plots to ensure that we are not comparing two PG models with substantially different discriminating performance. We then quantified the differences in the $\log_{10}$(LRs) assessed by the two systems in H1 true cases as well as in H2 true cases and evaluated the possible reasons

behind these discrepancies. Various decisions made as to the choice of thresholds and software parameter settings are outlined in detail in the methods section.

We believe that this is the first study that is large-scale, uses publicly available data, and evaluates the extent to which different models disagree (e.g. by a factor of 10, factor of 100, more than a factor of 1000). We outline the steps that may be used by other laboratories to assess the performance of different LR systems and analyze the resulting data. We also share the generated LR values in the interest of transparency and literature-to-literature comparisons by other researchers. Notably, the results are expected to vary if other parties conduct a similar analysis but use different software versions and protocols. Nevertheless, the process of comparison will essentially consist of the same steps outlined herein.

2. Methods

2.1. PROVEDIt dataset description

In this study, the Short Tandem Repeat (STR) profiles used to set the PGS parameters and calculate the LRs were selected from the PROVEDIt (Project Research Openness for Validation with Empirical Data) dataset that was amplified with GlobalFiler (GF) kit (29 cycles) and analyzed on 3500 Genetic Analyzer with an injection time of 15 seconds (s) [43, 44]. Both raw (.hid) and filtered (.CSV) files were used in the analysis. The filtered files present in the PROVEDIt database consist of the exported genotype tables containing allele designation, base pair (bp) size, and peak heights information for each sample profile analyzed in GeneMapper ID-X at an analytical threshold (AT) of one Relative Fluorescent Unit (RFU). Also, these filtered files did not contain artefacts such as pull-up, minus A, and– 2 bp in the SE33 locus as they were removed according to a defined criteria set by Alfonse et al. [43].

A total of 154 two-person (2P), 147 three-person (3P), and 127 four-person (4P) mixture profiles were obtained from the filtered (.CSV) files and used for LR calculations. The 2P, 3P, and 4P testing sets were prepared using DNA from 22 individuals for whom reference profiles were also available. The profiles used had varying: (1) minor contributor template amounts, (2) total input template amounts, (3) contributor ratios, and (4) DNA quality. A detailed description of the 2P, 3P, and 4P profiles that were used in the study is shown in S4 Table.

The mixture input files were analyzed using the per dye specific ATs discussed in Section 2.3 (shown in Table 1) and converted along with person of interest (POI) files into a format specific to each software [45, 46]. Non-numeric values, Off-Ladder "OL" peaks, were eliminated from all the analysis [46].

2.2. The LR system

The conventional CE genotyping workflow used in forensic DNA laboratories is composed of several steps that can be grouped into two processes: measurement and interpretation (Fig 1A) [47]. The measurement process involves genomic DNA extraction, quantification, amplification using commercial multiplex STR kits (herein GF 29 cycles), and electrophoretic separation (herein 3500 at 15s injection time). The outcome of the measurement process is an electropherogram (EPG) composed of the length variants, heights, and sizes of the allelic and non-allelic peaks. The interpretation process involves data analysis. The outcome of the interpretation process is a strength of evidence statement often reported in the form of a LR and typically requires PGS.

Our definition of the LR system is the entire pipeline starting from sample acquisition all the way to LR calculation. The PGS is a piece of the whole LR system. Therefore, performance assessment of the LR system is not only an assessment of the software but an assessment of the entire process.

Table 1. Summary of STRmix v2.6 and EFM v2.1.0 interpretation parameters and reported LR values.

Software	Interpretation summary
STRmix v2.6 https://www.strmix.com/	• Per dye ATs were set in STRmix kit settings (Blue = 35; Green = 65; Yellow = 45; Red = 50; Purple = 60)
	• Drop-in frequency = 0.0015 and drop-in cap = 180 RFU
	• MCMC settings: 8 chains of 100,000 burn-in accepts, 50,000 post burn-in accepts per chain
	• N-1, N-2 and N+1 stutter peaks modeled
	• 333 single source profiles used for Model Maker
	• Allelic variance (α, β) 5.653, 2.961; back stutter variance (α, β) 1.501, 27.227; forward stutter variance (α, β) 1.501, 31.710; double back stutter (α, β) 1.771, 21.655; LSAE variance 0.031
	• Sub-source LR values labeled as sub-source LRs in STRmix report were considered
EuroForMix v2.1.0 http://www.euroformix.com	• Overall lowest AT value was set (35 RFU)
	• Drop-in probability = 0.0015 and Drop-in hyper-parameter (λ) = 0.032
	• N-1 stutter peaks modeled
	• Degradation and stutter models jointly selected
	• Sub-source LR values labeled as MLE based LRs in EFM reports were considered
Both software	• Same/fixed mixture EPG features
	• Input mixture profiles were analyzed using the per dye ATs (Blue = 35; Green = 65; Yellow = 45; Red = 50; Purple = 60)
	• Same defined pair of propositions
	• Same combination of comparisons (mixture vs POI) per each analysis
	• True NOC
	• NIST 1036-Caucasian allele frequencies [57]
	• F_{ST} (θ) = 0.01 [28, 58]

https://doi.org/10.1371/journal.pone.0256714.t001

Herein the measurement process was established by Alfonse et al. [43] as mentioned previously in Section 2.1 and therefore was fixed for both LR systems. Thus, the performance assessment in this study encompasses the interpretation process as shown in Fig 1B that includes:

- our decision of using and processing the filtered PROVEDIt files
- parameter values determined according to the chosen software (discussed below in detail)
- the choice of PGS (herein STRmix v2.6 and EFM v2.1.0)
- the initial assessment of the LR values
- the check of diagnostics (review of per locus LR, deconvolution, genotypic weights, Gelman-Rubin statistics, log likelihood, and model selection)
- the reporting of the LRs

2.3. Analytical Thresholds (ATs)

To determine the AT, 41 pristine single source DNA profiles with varying amounts of DNA template 0–0.5 ng were obtained from the filtered version (.CSV) of PROVEDIt files [44]. The list of the 41 samples selected for AT determination are detailed in S1 Table. Allelic, stutter, and other artifactual peaks were discarded from these profiles. The mean (μ) and standard deviation (σ) of the remaining peaks (noise observations) were estimated per dye-color

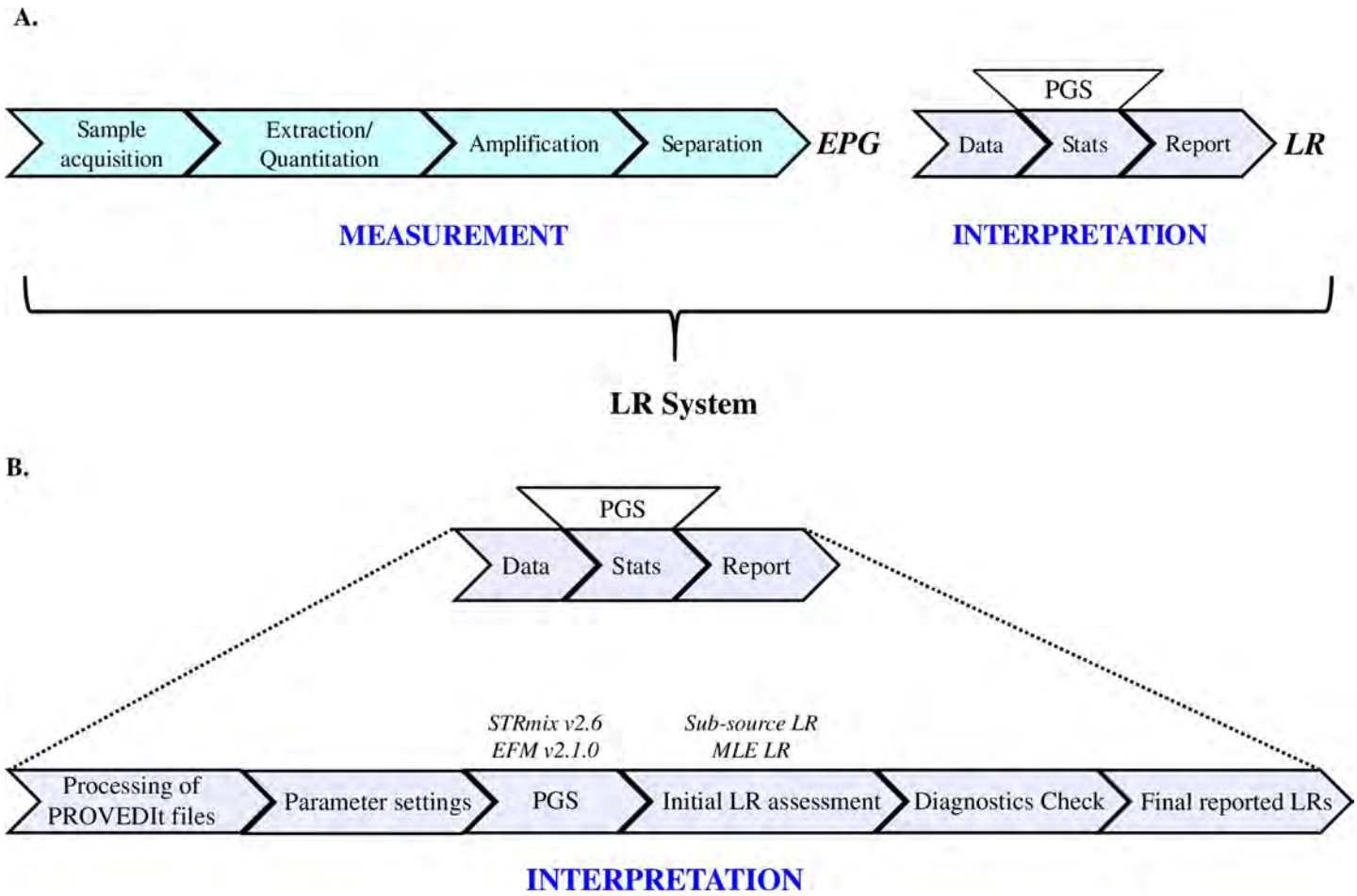


Fig 1. Schematic overview of the Likelihood Ratio (LR) system, adapted from [47, 48].

https://doi.org/10.1371/journal.pone.0256714.g001

channel. Then, AT was determined by substituting the values of μ and σ in the following equation: $AT = \mu + k^{*}\ \sigma$, where k was set to 10 [36, 49–52]. The AT values were rounded up to the nearest multiple of 5 (Table 1) [36]. All peaks in the input profiles explored in this study with peak heights below the determined dye-specific AT values (shown in Table 1) were filtered out before importing the data into STRmix and EFM.

Dye-specific ATs were set in STRmix as determined empirically and shown in Table 1. EFM v2.1.0 allows the user to set an overall single AT value [45]. The lowest RFU value (35 RFU) was used as the AT parameter in the EFM software.

2.4. Drop-in

Raw data (.hid files) of a set of 189 negative control profiles (listed in S2 Table) from PROVEDIt database were analyzed in GeneMapper ID-X v1.5 with a 35 RFU for all the dye channels. The selected profiles that were amplified with no DNA (0 ng) resulted in 7 drop-in events.

For STRmix, the drop-in frequency and drop-in cap parameters were determined and entered in the software. The frequency of the observed drop-in events was determined by using the instructions contained in the drop-in worksheet available on STRmix support website [53]. The highest drop-in peak observed in this study was 101 RFU. To set the drop-in cap at a value that is greater than the 101 RFU as recommended in [36, 53], we calculated the μ

and σ of the peak heights of the observed drop-in events and substituted these values in the following equation: Drop-in cap = $\mu + k^* \sigma$, where k was set to ≈ 5. The drop-in cap was then rounded up to the nearest multiple of 5. The determined values of drop-in frequency and drop-in cap are shown in Table 1. Due to the few drop-in events (only 7) observed within our analysis a uniform distribution was selected in the software [46, 53].

EFM requires the setting of the drop-in parameters, the drop-in probability (C), and the hyper-parameter (λ). Here, C was determined using $C = n/N^*L$, where C is the drop-in probability per marker, n is the number of drop-in events, N is the number of samples used to count the number of drop-ins, and L is the number of markers in each sample used to count number of drop-ins. The estimated λ was determined using $\lambda = n/\Sigma_i(x_i - T)$, where T is the analytical threshold used for analyzing drop-in, x_i is the peak height of each drop-in observed, and n is the number of drop-in events [17, 45, 54]. The determined values of C and λ are shown in Table 1.

2.5. Stutter

In this study, only double-back/N-2 (B2), back/N-1 (B1), and forward/N+1 (F1) stutter models in STRmix were applied when assessing LRs. All the mixture profiles analyzed herein did not contain stutter peaks at the– 2bp position at SE33 and D1S1656. Stutter files that already exist within the software from a previously validated GF 29 cycle kit were used [36, 55]. EFM v2.1.0 models only back stutter [45, 54, 56]. Stutter types chosen to be modelled in STRmix (B1, F1, and B2) were retained in the input files after applying the AT values, and imported in both software even though F1 and B2 were not modelled in EFMv2.1.0. Any unmodelled stutter can also be explained as drop-in allelic events [39].

2.6. Variance parameters

Single source profiles (n = 333) obtained from the PROVEDIt database (filtered CSV files) were analyzed at an AT = 10 RFU at all the dye channels to maximize stutter observations of all the stutter types being modelled. A detailed description of the quality and quantity of the samples used in the calibration set is listed in S3 Table.

The α, β parameters describing the gamma distribution (Gamma (α, β)) of the allele peak height variance (c^2) and stutter peak height variances (k^2), and the mean of the locus-specific amplification efficiency variance (LSAE) derived from the Model Maker analysis (shown in Table 1) were set into the software prior to the interpretation of the DNA mixture profiles.

2.7. LR calculations and data analysis

The strength of evidence was assessed after setting parameters specific to each software as summarized in Table 1. STRmix interpretations were undertaken using the recommended MCMC parameters (shown in Table 1) [46]. In follow up analyses two interpretations were repeated with an increase in the number of accepts (1,000,000 burn-in and 500,000 post burn-in accepts per chain) to allow each of the chains to explore more possibilities in the probability space [59]. The reported sub-source LRs within the STRmix reports were considered for the analysis in this study.

LR calculations in EFM were performed using the maximum likelihood estimate (MLE) method with both the degradation and stutter statistical models jointly turned on and included in all the EFM analysis. The reported sub-source LRs within the EFM labeled as MLE based LRs were used in the data analysis.

The true NOC (ground truth) was specified in the settings of the software for each mixture profile that was interpreted. Each of the PROVEDIt mixture profile was compared to the

Table 2. Summary of the total number of PROVEDIt mixture profiles and H1-true and H2-true propositions analyzed in both STRmix and EFM for 2P, 3P, and 4P mixtures.

Number of contributors	Number of mixtures	Propositions	Number of H1-true tests	Number of H2-true tests
2P	154	H1: POI + U1	308	308
		H2: U1 + U2		
3P	147	H1: POI + U1 + U2	441	441
		H2: U1 + U2 + U3		
4P	127	H1: POI + U1 + U2 + U3	508	508
		H2: U1 + U2 + U3 + U4		

POI indicates the person of interest that can be either known contributor or known non-contributor. U1, U2, U3, and U4 indicate one, two, three, or four unknown, unrelated individual(s) to the mixtures. For each mixture, we performed as many known contributor LR analysis (H1-true tests) as there are contributors to each mixture. For each contributor analysis, a non-contributor LR analysis (i.e. single H2-true test) was also performed using real (true-genotype) profiles randomly chosen from NIST 1036 US population dataset [57].

https://doi.org/10.1371/journal.pone.0256714.t002

appropriate known contributors (S4 Table) and known non-contributors (S5 Table). The known non-contributors were real (true-genotype) profiles randomly selected from the NIST 1036 US population dataset [57].

The allele frequencies and coancestry coefficient (F_{ST} or θ) set in both software for LR calculations are shown in Table 1. The propositions considered and the total number of propositions generated from each software are outlined in Table 2.

All the LR values yielded from both software are reported in $\log_{10}$ scale in S4 Table ($\log_{10}$(LRs) for H1-true tests) and S5 Table ($\log_{10}$(LRs) for H2-true tests) with the corresponding combination of comparisons (mixture vs POI). The profile LRs and the per-locus LRs assigned by STRmix and EFM were for the 21 autosomal STR markers only. LR assessment for the gender and Y-STR markers, Amelogenin, Y-indel, and DYS391, were not considered by either software.

All data analysis and visualization discussed were conducted using the open source software **R** [60].

3. Results and discussion

3.1. Empirical assessment of LR systems using discrimination performance of H1-true and H2-true LR distributions

We first examined the overall performance of the two systems to ensure that we are not comparing two PG models with substantially different discriminating performance. The distributions of the assigned $\log_{10}$(LR) values were plotted as function of NOC (2P, 3P, and 4P), propositions (H1 and H2), and software (STRmix and EFM) (Fig 2). The overall distribution plot shown in Fig 2 was further broken down by varying mixture ratios (S1 File) and different DNA treatments used to compromise the DNA quality of the samples (DNA damage, DNA degradation, and PCR inhibition) (S2 File).

The magenta and blue data points are the $\log_{10}$(LRs) of the H1-true tests generated in STRmix and EFM, respectively. Log_{10}(LRs) of the H2-true tests assigned by STRmix and EFM are shown in cyan and green, respectively (Fig 2, and S1 and S2 Files). The distribution of $\log_{10}$(LRs) from the H1-true tests is well separated from the distribution of $\log_{10}$(LRs) from the H2-true tests when the quality and DNA template amount of the contributor or total template amount of the samples are sufficiently high and the NOC in a mixture profile is low. As the quality and template amount per contributor of interest or mixture profile decreases and/or

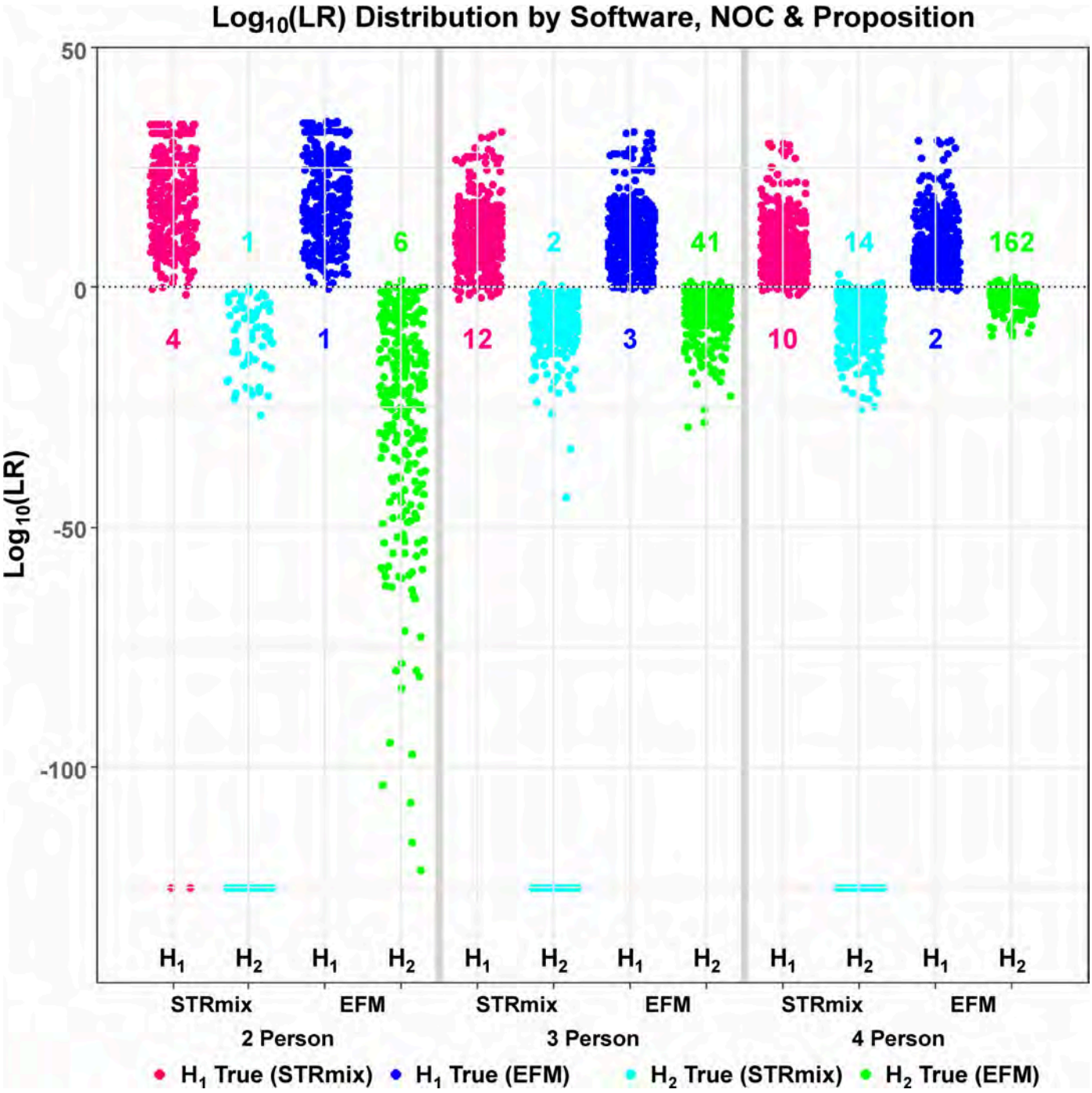


Fig 2. Distribution of $\log_{10}$ (LR) values for H1-true and H2-true tests assessed by STRmix and EFM for two, three, and four person mixtures. The x-axis shows the labels of propositions (H1 and H2), software (STRmix and EFM), and the NOC = 2 Person, 3 Person, and 4 Person. LRs are plotted on the y-axis as $\log_{10}$(LR) values. All samples from different mixture ratios, total DNA template amounts, and DNA treatments are built into this global/overall distribution plot. The plot contains a total of 308 H1-true tests and 308 H2-true tests for the 2P analysis, 441 H1-true and 441 H2-true calculations for the 3P analysis, and 508 H1-true and 508 H2-true tests for the 4P mixtures. STRmix provides an LR value of 0 for excluded loci resulting in profile LR of 0, while EFM gives a non-zero LR value (generally very close to zero). Profiles with LR results of 0 from STRmix are plotted at −125 on the $\log_{10}$ scale. ** Two H1-true test interpretations of 2P mixtures for which STRmix assigned profile LRs of 0 (plotted at H1 true STRmix NOC = 2 Person in magenta at −125 on the $\log_{10}$ scale and discussed in detail in Section 3.6).

https://doi.org/10.1371/journal.pone.0256714.g002

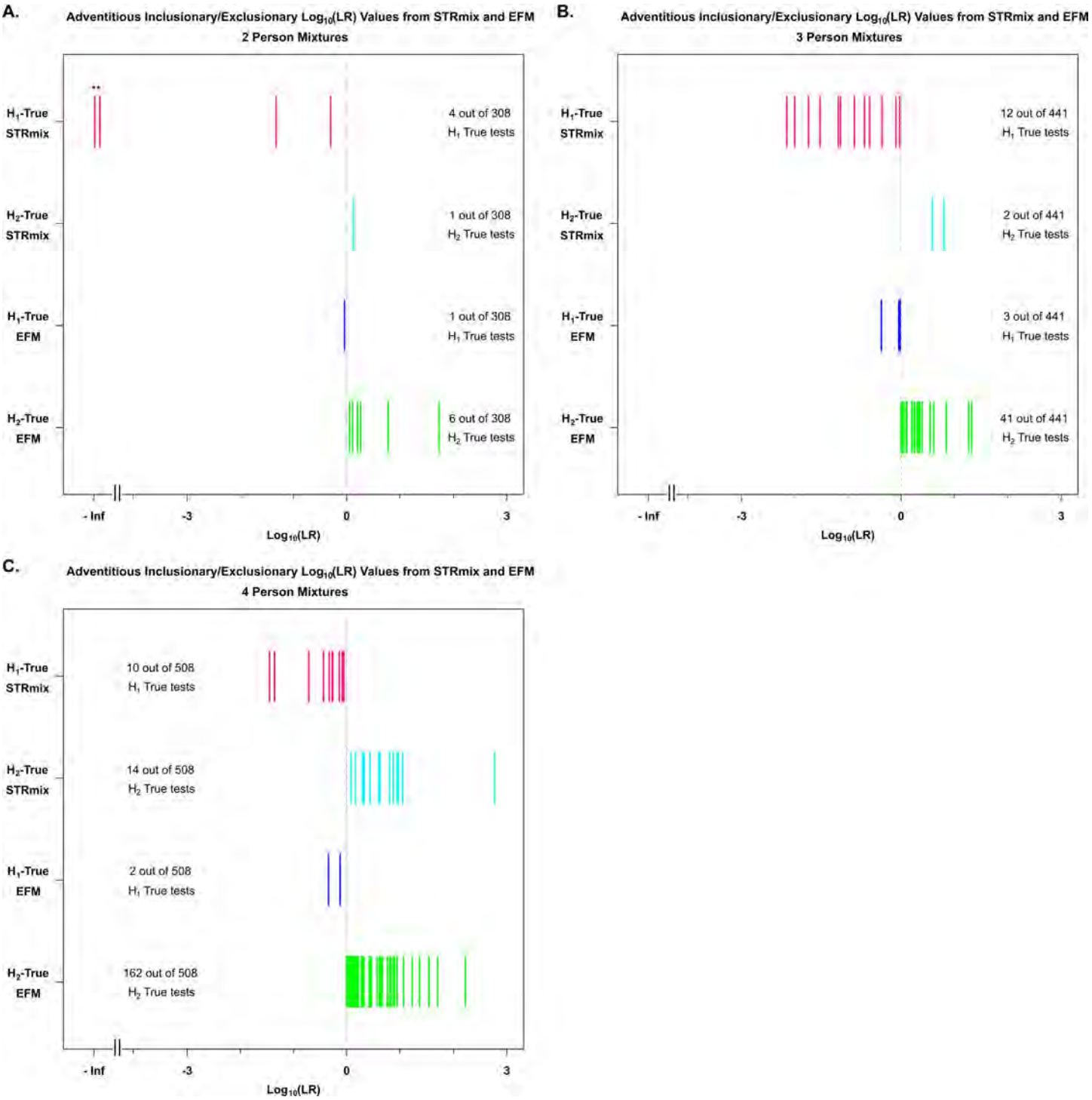


Fig 3. Summary of adventitious exclusionary and inclusionary support from both LR systems with their corresponding $\log_{10}$(LR) values. The x-axis shows the $\log_{10}$(LR) values for these adventitious exclusionary and inclusionary cases. The y-axis shows the labels of the tested propositions (H1 and H2) from each software (STRmix and EFM). ** in (A.) are the two 2P H1-true test interpretations for which STRmix assigned profile LRs of 0 (plotted in magenta at–Infinity (-Inf) on the $\log_{10}$ scale and discussed in detail in Section 3.6).

https://doi.org/10.1371/journal.pone.0256714.g003

the NOC increases, $\log_{10}$(LRs) assigned from H1-true tests and H2-true tests become less discriminatory and trend downwards and upwards towards 0 (horizontal line), respectively, (Fig

2, and S1 and S2 Files). Furthermore, as expected, when the distinction between the major-minor contributions to the same mixture increases so does the LRs of the major contributors as opposed to mixtures with equal contributor proportions (S1 File). As expected, the latter have lower LRs since information content associated with peak heights is limited or has no effect on LR calculations [3, 61–64].

The magenta and blue data points below the central dashed horizontal line plotted at $\log_{10}$(LR) of zero in Fig 2 and S1 and S2 Files, correspond to the analyses of known contributors within STRmix and EFM that yielded $\log_{10}(LRs) < 0$ (adventitious exclusionary LRs). Cyan and green points above the horizontal line at $\log_{10}(LR) = 0$ in Fig 2 and S1 and S2 Files are instances of H2-true tests that yielded $\log_{10}(LRs) > 0$ (adventitious inclusionary LRs). The number of these adventitious inclusionary and exclusionary LR instances are indicated in Fig 2. These profiles are also presented with their corresponding $\log_{10}$(LRs) in Fig 3 and S6 and S7 Tables and are discussed in further details in Section 3.2.

Visual comparisons of the global aggregate of $\log_{10}$(LRs) in the distribution plot of Fig 2 indicate qualitatively that STRmix and EFM seem to have equal ability in discriminating between H1-true and H2-true scenarios. Both LR systems indicate better discrimination performance for lower complexity mixtures than for higher complexity mixtures (mixtures characterized by an increase in NOC and/or decrease in DNA quantity and quality). These qualitative observations are substantiated statistically in Section 3.3.

3.2. Overall specificity and sensitivity of the two LR systems

In this section we discuss overall specificity and sensitivity (Table 3) and instances of adventitious exclusionary LRs of which H1-true tests resulted in LR < 1 and cases of adventitious inclusionary LRs of which H2-true tests yielded LR > 1 across both NOC and LR systems (Fig 3 and S6 and S7 Tables).

Across all the 2P, 3P, and 4P mixtures, 97.93% and 99.52% of H1-true test LRs assigned by STRmix and EFM, respectively, were greater than 1 (or $\log_{10}(LR) > 0$) (Table 3) while 98.65% and 83.37% of H2-true test LRs assigned in STRmix and EFM, respectively, resulted in LRs lower than 1 (or $\log_{10}(LR) < 0$) (Table 3). The number of observations and frequency values are broken down by NOC and LR systems as shown in Table 3.

3.2.1. Adventitious exclusionary support (examples of LR $<$ 1 when H1 is true). There were instances of adventitious exclusionary LRs for true contributor analyses (H1-true tests) within both LR systems that returned $\log_{10}(LRs) < 0$ as illustrated in Fig 3 and S6 Table. Across the 1,257 H1-true tests conducted, there were 26 instances of adventitious exclusionary support with STRmix (4 out of 308 with 2P profiles, 12 out of 441 with 3P profiles, and 10 out of 508 with 4P profiles) and 6 instances with EFM (1 out of 308 with 2P profiles, 3 out of 441 with 3P profiles, and 2 out of 508 with 4P profiles) of which $\log_{10}$(LR) values for the POI were below 0. These are shown with their corresponding $\log_{10}$(LRs) in Fig 3 and S6 Table. As expected from the behavior of the LR [65] and as shown in S6 Table, all the cases of H1-true tests with $\log_{10}(LRs) < 0$ from both LR systems mainly occurred when comparing the minor contributors to DNA mixture profiles that contained limited amount of information due to low minor template amount (e.g. $\leq$ 63 pg), low total template amount, compromised/degraded DNA, loci with allelic dropout, increase in the number of contributors, stochastic variation causing confounding information from the allelic and stutter peaks, and allele sharing between contributors [7].

The number of instances of H1-true tests with $\log_{10}(LRs) < 0$ was greater with STRmix than EFM. However, the $\log_{10}$(LRs) generated in EFM for these STRmix cases were mostly true inclusions of low-level LR range between (1–1,453) (i.e., uninformative or slightly to

Table 3. Summary of the number of observations and frequency (%) of known contributor analyses (H1-true tests) and known non-contributor analyses (H2-true tests) that yielded $\log_{10}$(LR) values > 0 (or LR > 1) and $\log_{10}$(LR) values < 0 (or LR <1), respectively.

H1-True Tests: LR > 1					H2-True Tests: LR < 1				
	STRmix		EFM			STRmix		EFM	
# of contributors	Counts	Frequency %	Counts	Frequency %	# of contributors	Counts	Frequency %	Counts	Frequency %
2 (N = 308)	304	98.70	307	99.68	2 (N = 308)	307	99.68	302	98.05
3 (N = 441)	429	97.28	438	99.32	3 (N = 441)	439	99.55	400	90.70
4 (N = 508)	498	98.03	506	99.61	4 (N = 508)	494	97.24	346	68.11
Total (N = 1,257)	1,231	97.93	1,251	99.52	Total (N = 1,257)	1,240	98.65	1,048	83.37

N represents the total number of either H1-true tests or H2-true tests conducted for the different number of contributors.

https://doi.org/10.1371/journal.pone.0256714.t003

moderately supporting H1 over H2) with the exception of three 2P instances that are discussed in Section 3.4. For example, as seen in S6 Table, when mixture F10_RD14-0003-39_40–1;2-M3c-0.045GF was compared to the minor contributor "39", STRmix gave a $\log_{10}$(LR) of -0.2 while EFM gave a $\log_{10}$(LR) of 0.9.

3.2.2. Adventitious inclusionary support (examples of LR > 1 when H2 is true). There were also instances of adventitious inclusionary LRs for known non-contributor analyses (H2-true tests) that returned $\log_{10}$(LRs) > 0 within both LR systems as illustrated in Fig 3 and S7 Table. Out of the 1,257 total H2-true tests performed for 2P, 3P, and 4P, there were 17 $\log_{10}$(LRs) greater than zero analyzed with STRmix (1 out of 308 with 2P, 2 out of 441 with 3P profiles, and 14 out of 508 with 4P profiles) and 209 $\log_{10}$(LRs) greater than zero with EFM (6 out of 308 with 2P profiles, 41 out of 441 with 3P profiles, and 162 out of 508 with 4P profiles). These cases are presented with their corresponding $\log_{10}$(LRs) in Fig 3 and S7 Table. The largest observed LR for the known non-contributors assigned by STRmix was 587 ($\log_{10}$(LR) = 2.7) and in EFM was 167 ($\log_{10}$(LR) = 2.2), when comparing a known non-contributor with the 4P mixture D02_RD14-0003-40_41_42_43–1;1;1;1-M2e-0.124GF.

As expected, positive $\log_{10}$(LRs) obtained from non-donors in both software were attributed to one or more of the following: increased complexity of mixtures, increase in the number of contributors, mixtures generated from low total template and/or compromised low quality DNA, stochastic effects, and chances of allele sharing between the non-contributor profiles and evidence profiles [2, 65–67].

The number of instances of positive $\log_{10}$(LRs) from non-contributors were greater with EFM than STRmix (Fig 3 and S7 Table). The LR values assigned by EFM were based on the MLE method, an approach that has elevated rates of LR > 1 for the H2-true tests than the conservative method as stated and observed in [54, 62]. However, these adventitious inclusionary LRs were low-level with range of values between 1 to 53 (i.e., uninformative or slightly supporting H1 over H2) (S7 Table) [54, 62, 68].

3.3. Using empirical Receiver Operating Characteristic (ROC) plots to study discrimination performance of the LR systems

We used Empirical Receiver Operating Characteristic (ROC) plots [69] as statistical tools to quantify the discrimination performance between the H1-true scenarios and H2-true scenarios of the two different LR systems. The discrimination performances were quantified using a numerical metric, the Area Under ROC Curve (AUC). AUC is the area between each ROC plot and the horizonal x-axis (Fig 4). Statistical tests (p-values) for AUC comparisons (i.e., differences between the ROC plots) were calculated and listed in Fig 4 [70].

Comparison Group	*P-values*
STRmix 2P *vs* EFM 2P	0.1638
STRmix 3P *vs* EFM 3P	0.1093
STRmix 4P *vs* EFM 4P	0.1859

LR values of the H1-true tests and H2-true tests were combined across each NOC level (2P, 3P, and 4P) generated from each software (STRmix and EFM), thus creating six datasets: STRmix 2P, EFM 2P, STRmix 3P, EFM 3P, STRmix 4P, and EFM 4P. To construct the ROCs shown in Fig 4, a series of various LR thresholds were applied to each of the 6 datasets generating true positive rates (TPR) and the corresponding false positive rates (FPR). TPR represented the counts of the true contributors of which LR values were > a given threshold value divided by the total counts of the known contributors in the considered dataset. FPR represented the counts of the known non-contributors with LR values > a given threshold value divided by the total counts of known non-contributors in the considered dataset. ROC plots were created by plotting the TPR (along vertical axis) versus the FPR (along horizontal axis). The p-values of the comparisons of areas under the ROC plots of: STRmix 2P vs EFM 2P, STRmix 3P vs EFM 3P, and STRmix 4P vs EFM 4P were > 0.05 (Fig 4), indicating that for the considered data the differences between the two software in the ability to discriminate between H1-true and H2-true scenarios were not statistically significant.

The ROC plots shown in Fig 4 statistically support the qualitative observation visualized in the distribution plots of Fig 2. Therefore, the ability for the two LR systems to discriminate between known contributors and known non-contributors are statistically indistinguishable for the data considered. However, that does not imply that STRmix and EFM are producing equal LR values or agreeing when the same profile is being interpreted within both software. Sample to sample comparisons are discussed in Section 3.4. Rather the plots in Figs 3 and 5 are considering the data in aggregate.

3.4. Global overall profile $\log_{10}$(LR) values of H1-true tests and H2-true tests from each LR system

Scatter plots (Fig 5) were produced by plotting the $\log_{10}$(LRs) of the H1-true tests (magenta datapoints) and the H2-true tests (blue datapoints) obtained from STRmix on the x-axis against the corresponding $\log_{10}$(LRs) assigned using EFM on the y-axis for the 2P (Fig 5A), 3P (Fig 5B), and 4P (Fig 5C) mixture profiles. Identical or near identical $\log_{10}$(LR) values assigned by both LR systems fell on the solid black 45˚ degree line, X = Y. Datapoints that did not fall on the diagonal line corresponded to instances with varying degrees of difference in the overall LR profile between the two LR systems. For example, datapoints located within the two black dashed lines, two black dash-dotted lines, and two black dotted lines surrounding the line X = Y, corresponded to cases with LR results differing by a factor as high as 10^2, 10^4, and 10^6, respectively (Fig 5). Datapoints that are outside the pair of black dotted bands represented LRs assigned by the two software that differed by more than a factor of 10^6. These differences represented instances where either the LRs obtained from STRmix exceeded the ones obtained in EFM or vice versa. It is interesting to note that differences in the assigned LR values were greater with the non-contributor testing profiles than with the H1-true testing cases. Instances that differed by factor of $\geq 10^3$ and the potential explanations for the differences will be discussed in Section 3.6. Impacts of the differences in the inter-software numerical LRs on verbal expression will be discussed in Section 3.7.

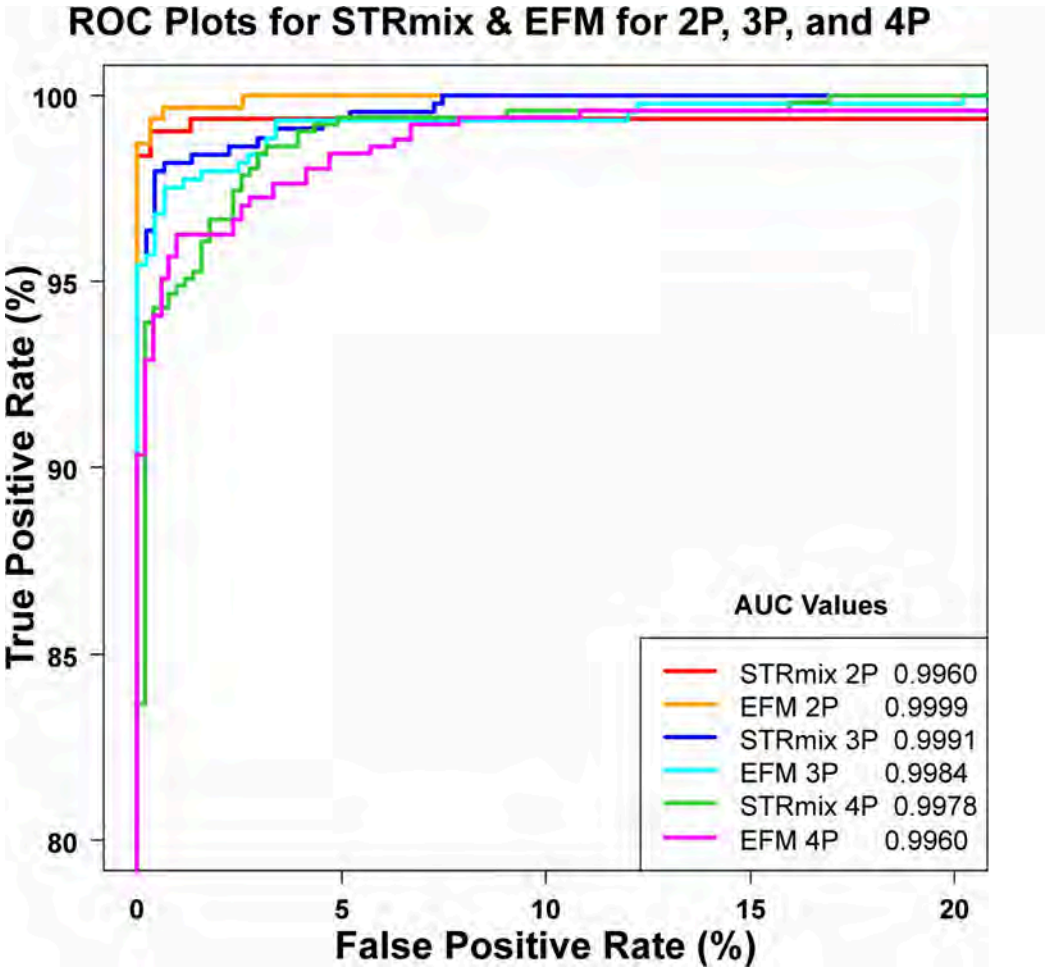


Fig 4. Empirical ROC plots used to study discrimination performance of the LR systems. ROC plots are built per varying NOC and software. Each NOC dataset is composed of profiles of different DNA quality, quantity, and mixture proportions. The red, blue, and green curves are the ROC plots constructed using LR values of known contributors and known non-contributors of 2P, 3P, and 4P mixtures analyzed within STRmix, respectively. ROC plots constructed with LR values assigned by EFM are shown in orange (2P), cyan (3P), and magenta (4P). The plot contains a total of 308 H1-true tests and 308 H2-true tests for the 2P analysis, 441 H1-true and 441 H2-true calculations for the 3P analysis, and 508 H1-true and 508 H2-true tests for the 4P mixtures. The calculated AUCs and p-values are shown. All p-values were > 0.05.

https://doi.org/10.1371/journal.pone.0256714.g004

To conclude this section, although both LR systems show comparable discrimination performance, differences exist in $\log_{10}$(LR) values on a case-by-case basis. Differences in $\log_{10}$(LR) values assigned by STRmix and EFM at the profile level covered a wide range from zero to over a million (discussed in detail in Sections 3.5 and 3.6) for the same input data (i.e., the same EPG). The differences appear to be greater in the H2-true cases than in the H1-true cases.

3.5. Distribution of differences in $\log_{10}$(LR) values between the two LR systems

Here, we describe and plot the degree and distribution of the observed differences between the two LR systems. The actual differences in $\log_{10}$(LRs) were calculated in both directions (i.e., $\log_{10}(LR)_{STRmix}-\log_{10}(LR)_{EFM}$ as well as $\log_{10}(LR)_{EFM}-\log_{10}(LR)_{STRmix}$) for the H1-true tests and H2-true tests (histograms shown in Fig 6). These differences were broken down into factor

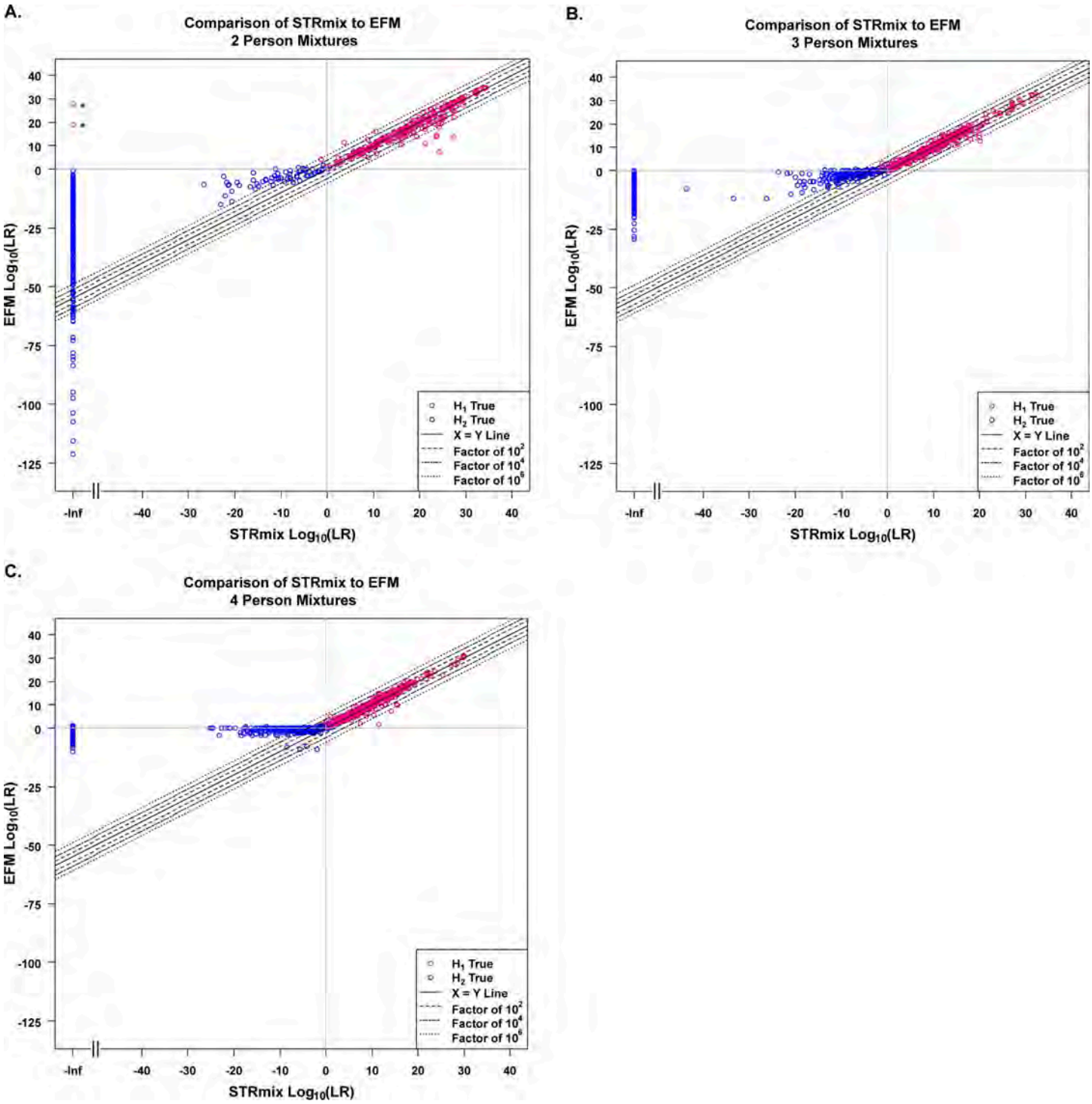


Fig 5. Global overall profile H1-true test and H2-true test $\log_{10}$(LR) values assigned by STRmix and EFM. ** in (A.) are the two 2P H1-true test interpretations for which STRmix assigned profile LRs of 0 (plotted in magenta at–Infinity (-Inf) on the $\log_{10}$ scale and discussed in detail in Section 3.6).

https://doi.org/10.1371/journal.pone.0256714.g005

of 10 bins for the 2P (Fig 6A), 3P (Fig 6B), and 4P (Fig 6C) analysis and the relative frequencies (in %) of these differences are indicated for each bin in Fig 6. For example in Fig 7A, 21.4% and 47.1% of the differences for the 2P H1-true tests were between 0 to 1 on $\log_{10}$ scale for

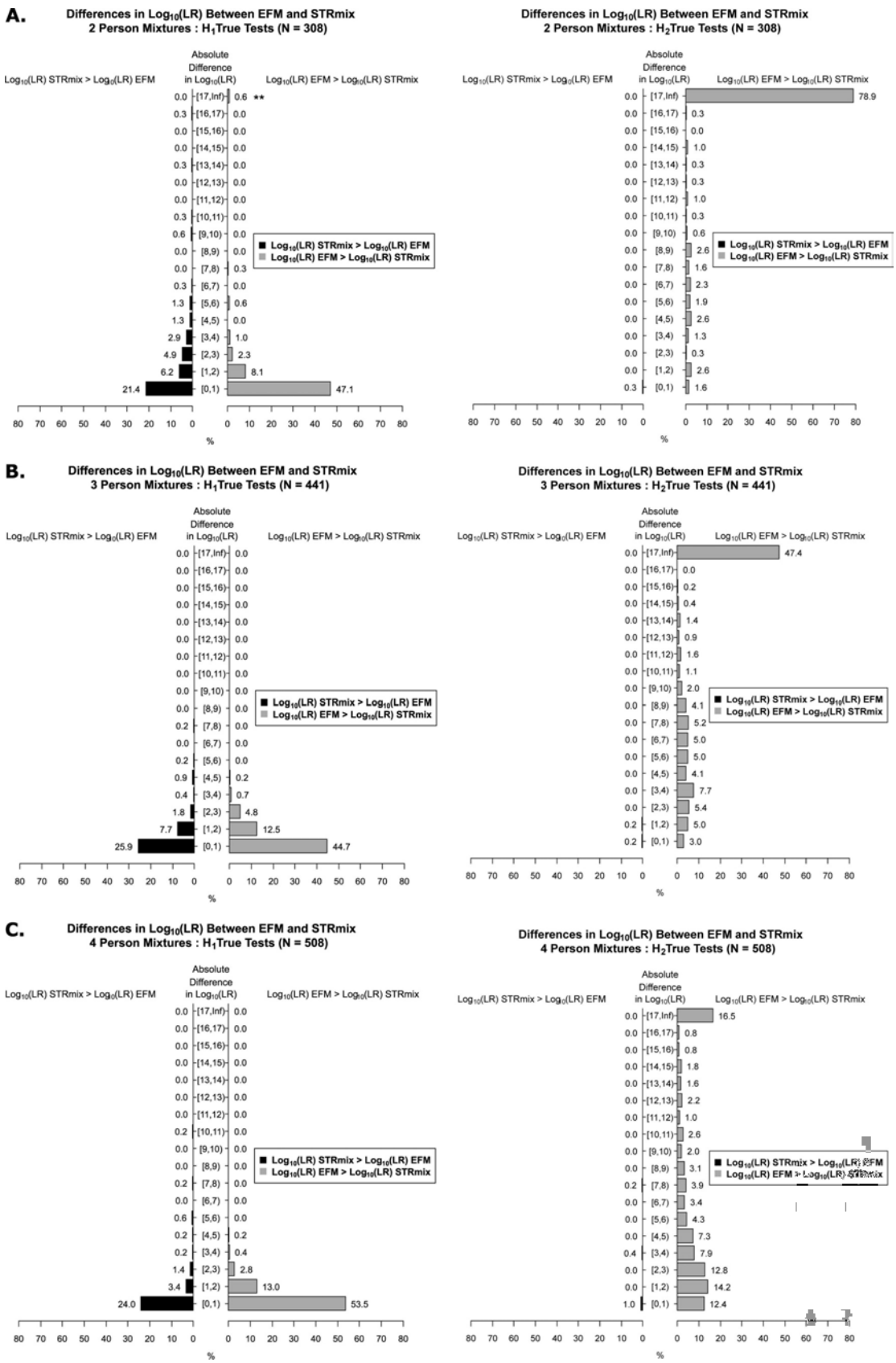

A.
Differences in Log10(LR) Between EFM and STRmix
2 Person Mixtures : H1True Tests (N = 308)
Differences in Log10(LR) Between EFM and STRmix
2 Person Mixtures : H2True Tests (N = 308)
B.
Differences in Log10(LR) Between EFM and STRmix
3 Person Mixtures : H1True Tests (N = 441)
Differences in Log10(LR) Between EFM and STRmix
3 Person Mixtures : H2True Tests (N = 441)
C.
Differences in Log10(LR) Between EFM and STRmix
4 Person Mixtures : H1True Tests (N = 508)
Differences in Log10(LR) Between EFM and STRmix
4 Person Mixtures : H2True Tests (N = 508)
Absolute Difference in Log10(LR)
Log10(LR) STRmix > Log10(LR) EFM
Log10(LR) EFM > Log10(LR) STRmix
%

Fig 6. Relative frequency histograms of the degree of differences in $\log_{10}$(LR) values between the two LR systems. The absolute difference in $\log_{10}$(LR) are shown on the y-axis. The square bracket "[" in the interval notation "[)" indicates that the endpoint is included in the interval and the parenthesis ")" in the interval notation "[)" indicates that the endpoint is not included. For example, [1, 2], is the interval of values between 1 and 2, including 1 and up to but not including 2, i.e., $1 \leq$ values < 2. The x-axis shows the relative frequencies (in %) of the differences in $\log_{10}$(LR) values between the LR systems occurring within each bin. The relative frequencies are also labeled above each bar of the histogram. ** are the two 2P H1-true test interpretations for which STRmix assigned profile LRs of 0 (binned into the [17, Inf) category and discussed in detail in Section 3.6).

https://doi.org/10.1371/journal.pone.0256714.g006

$\log_{10}(LR)_{STRmix}$—$\log_{10}(LR)_{EFM}$ (black histograms) and $\log_{10}(LR)_{EFM}$—$\log_{10}(LR)_{STRmix}$ (grey histograms), respectively. The relative frequency histograms (Fig 6) indicate that (i) the differences between the two LR systems were smaller with the H1-true testing cases than with the non-contributor tests and (ii) EFM tended to give higher LR values than STRmix for both the H1-true tests and H2-true tests.

The actual differences in $\log_{10}$(LRs) for the H1-true tests were further stratified by the type of POI (i.e., major, minor, and equal contributors as defined in S4 Table) constituting the 2P (Fig 7A), 3P (Fig 7B), and 4P (Fig 7C) mixture profiles. As shown from the distribution plots in Fig 7, the magnitude of the differences for the two LR systems were greater for the minor contributors (shown in magenta) than for the major (shown in blue) and for the equal (shown in green) contributors. LRs assigned by STRmix and EFM agreed more when POI(s) constitute the equal contributors of the mixture (Fig 7). This is expected because with balanced profiles, peak height information content has less effect on LR calculations than in cases of major: minor profiles [3, 61–64].

3.6. Evaluation of apparent differences in $\log_{10}$(LR) values between the two LR systems

In this section we discuss the steps performed to further investigate differences in the assigned LR values obtained from the two LR systems on a case-by-case basis, where the differences are observed, and the potential explanations for these differences. We restrict our discussion to instances when LR (STRmix) $\geq 1000^{*}$LR (EFM) that constituted 7.3% of the 2P, 1.7% of the 3P, and 1.4% of the 4P H1-true tests (histograms of Fig 6 and S11 Table) and instances when an LR (EFM) $\geq 1000^{*}$LR (STRmix) that accounted for 2.5% of the 2P, 0.9% of the 3P, and 0.6% of the 4P H1-true tests (histograms of Fig 6 and S12 Table). Only differences in H1-true results (true known contributor samples) are discussed.

LR computations obtained from the two software were based on same/fixed EPG features, same pair of propositions, NOC, theta, and population allele frequency. Therefore, results presented here shows that differences observed in LR values can occur due to one or more of the following reasons:

I. Nonconvergence of the Markov Chain Monte Carlo (MCMC) algorithms and MLE

II. Decision to provide identical EPGs for both LR systems

III. Different modeling assumptions and parameters settings between the two software

We discuss each of the above reasons and provide examples from the data set. The availability of both the mixture and reference profiles was beneficial and helped in the investigation of observed differences of the assigned LR values.

3.6.1. Non-convergence of the MCMC algorithms and MLE. STRmix pdf reports contain summary statistics for each interpretation conducted in the software and can be used by analysts as diagnostics on the performance of the interpretation according to the specified models. These diagnostics have been classified into primary and secondary categories and are

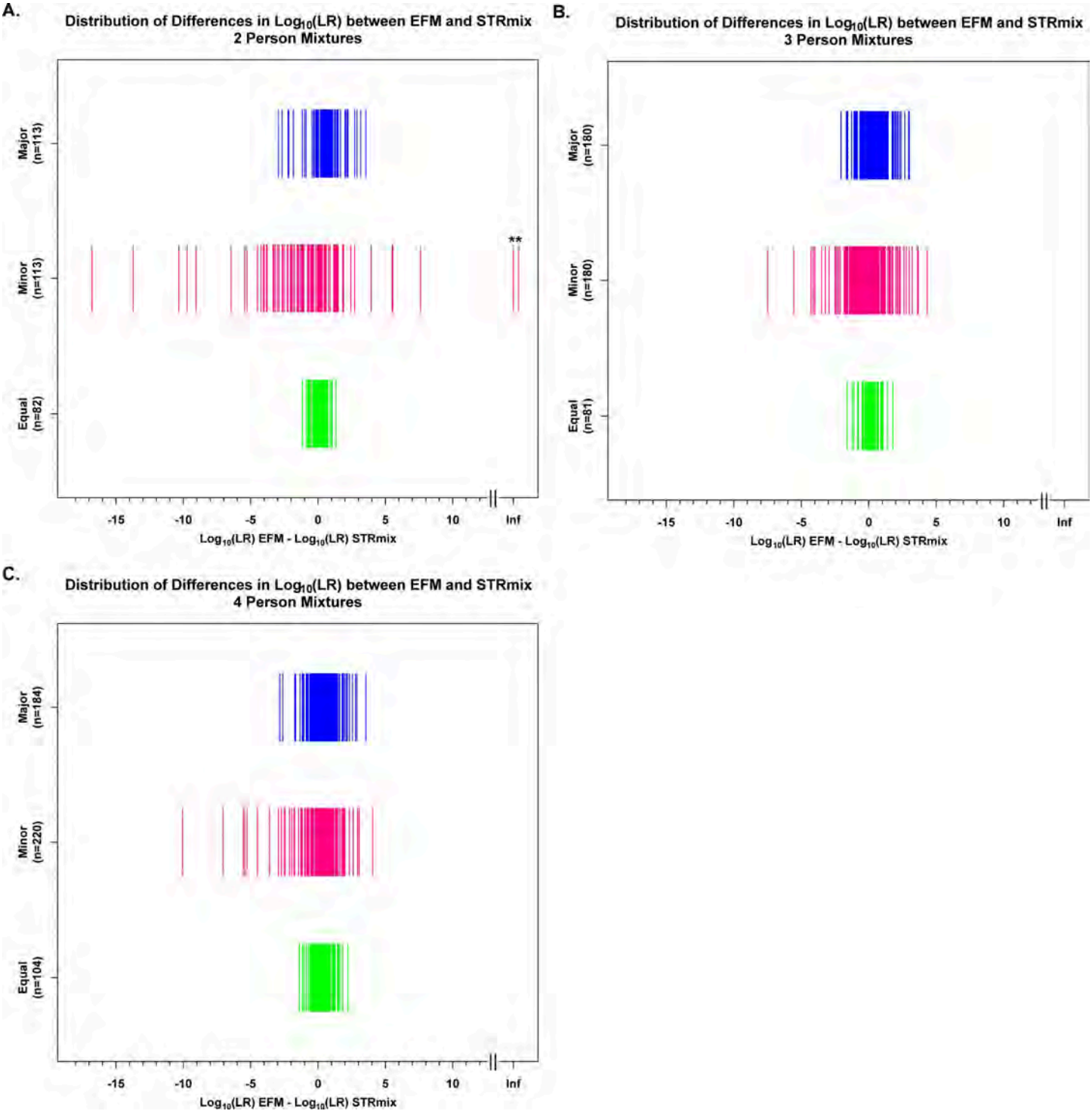


Fig 7. Distribution of differences in $\log_{10}$(LRs) across major, minor, and equal contributors. The differences in $\log_{10}$(LRs) here shown between EFM and STRmix ($\log_{10}(LR)_{EFM} - \log_{10}(LR)_{STRmix}$) are plotted on the x-axis in $\log_{10}$ scale. The y-axis shows the labels of the types of POI with their corresponding number of observations. ** are the two 2P H1-true test interpretations for which STRmix assigned profile LRs of 0 (plotted in magenta at Infinity (Inf) on the $\log_{10}$ scale and discussed in detail in Section 3.6).

https://doi.org/10.1371/journal.pone.0256714.g007

discussed in detail in Russell et al. [59]. In actual casework, every analysis should be subjected to diagnostic checks. But in this study and for practical reasons only cases where STRmix and EFM differed by a factor of $\geq 10^3$ were inspected for genotypic weights, mixture proportions, per-locus LRs, log(likelihood), peak height variance parameters, and Gelman-Rubin (GR) statistics.

Two extreme differences observed between STRmix and EFM were with the 2P mixture profiles, C02_RD14-0003-40_41–1;4-M2U15-0.315GF (herein referred to as "C02") and H06_RD14-0003-48_49–1;4-M2e-0.315GF (referred to as "H06") (S8 Table). C02 and H06 generated profile LR of 0 in STRmix when compared to true known minor contributors, 40 and 48, respectively. A locus LR value of 0 will lead to a profile LR of 0. The $\log_{10}$(LR) assessments for these profiles in EFM were 27.6 for C02 and 19.0 for H06. Unlike STRmix, EFM displays low to very low LRs for exclusionary loci but does not provide a zero locus LR. A review of the per locus LRs (S8 Table) assigned to the evaluation of the POIs in STRmix indicated that almost all loci favor inclusion (LR > 1) except for a single locus displaying an LR of 0 in each interpretation, D1S1656 in C02 and D3S1358 in H06. Instances of single locus LR = 0 have been observed using different data from different studies [2, 36, 71]. In such cases and if samples are sufficient, either replicate analysis or sample reamplification is used. Otherwise, options are to either ignore that locus during deconvolution, or repeat the deconvolution in STRmix with: a random starting seed for the MCMC different than the one that gave LR = 0, or an increase in number of MCMC accepts, or a larger Random Walk Standard Deviation (RWSD) [2, 7, 36, 59, 71]. Here, we repeated the runs in STRmix with more MCMC accepts (as discussed in Section 2.7) and the repeated interpretations generated non-zero LRs for the affected loci, and profile $\log_{10}$(LRs) of 24.8 and 19.6 (S8 Table). It is to note that these two discussed 2P H1-true test interpretations with profile LRs of 0 assigned by STRmix were plotted: (i) at −125 on the $\log_{10}$ scale in Fig 2 and S1 and S2 Files; (ii) at–Infinity (-Inf) in Figs 4A and 6A; (iii) at Infinity (Inf) in Figs 7A and 8A; and were binned into the exclusionary verbal category (Table A in Fig 8).

Another extreme difference observed between STRmix and EFM was with E04_RD14-0003-42_43–1;9-M2U105-0.15GF, a 2P mixture profile of which comparison to the minor contributor in STRmix and EFM, yielded $\log_{10}$(LRs) of -1.3 and 4.1, respectively (EPG and data shown in S4 Table). A review of the STRmix output indicated negative log(likelihood), which might be due to several reasons including "flawed input data" [2, 59]. Inspection of the DNA typing results, ground truth genotypes of the POIs, and deconvolution results indicated retained artifact peaks binned into alleles at two loci, D19S433 "18.2" and D5S818 "14" (EPG shown in S9 Table). The artifacts were each modelled in STRmix as being allelic in origin and were included in the genotypic combinations thus leading to exclusion after comparing the resolved profile to the true contributors [7, 71]. In such cases, mixture samples can be re-injected or reamplified [2]. However, since only the electronic data was accessible for this study, the artifacts were removed and the input file were re-interpreted in both STRmix and EFM, generating profile $\log_{10}$(LRs) of 0.8 and 4.6, respectively (S9 Table).

A GR > 1.2 might be an indication that more MCMC runs may be needed for convergence [7, 59, 72]. Profiles indicating discrepancies of $\geq 10^3$ and with a GR > 1.2 (a total of 6 out of 53) were reinterpreted in STRmix using higher number of burn-in and post burn-in accepts [7]. The repeated LR computations resulted in lower GR. Although the GR decreased, there was either no effect or a slight increase by a factor of 10 in the profile overall LRs (S10 Table) and did not substantially alter the observed LR differences in these cases.

EFM provides an option for selection of one of four models (turning on either or both of the degradation and stutter models) under H1 and H2 hypothesis and generates a Probability-Probability (PP) plot to examine if the model selected explain the observed data adequately

[56, 68, 73]. A linear trend of PP plots within 99% Bonferroni band indicates that the assumed continuous models may be adequate for the data of the observed peak heights above the detection threshold [56, 73]. Herein, we selected the model with both degradation and back-stutter options turned on and cross checked a total of four mixture profiles out of 53 interpretations showing discrepancies (S3 File). The PP plots showed that models selected (i.e., degradation ON and stutter ON) appear to adequately explain the data.

These observations indicate that non-convergence of MCMC and the inability of the software to describe the observed profile given the provided information are one of the reasons behind the observed differences.

II. Decision to provide identical EPGs for both LR systems

Instances of the underestimation of LR values observed in EFM as compared to STRmix was primarily due to the unmodelled stutter type peaks not filtered from the input files (S11 Table). Stutter models for B2, B1, and F1 were applied to the mixture deconvolutions performed in STRmix. EFM v2.1.0 used in this work only models stutter peak heights in the -1 repeat unit position [54, 56, 73]. The B2 and F1 peaks retained after applying the analytical thresholds and not pre-filtered from the DNA profiles before analysis in EFMv2.1.0 led to instances of smaller LRs than those assigned by STRmix (S11 Table). As reflected in (S11 Table), differences were highest with minor contributors in major/minor mixture profiles where allele peak heights from a minor contributor can have the same size and height as stutter peaks of major contributors [54, 59].

To further examine this hypothesis, we removed the retained (i.e., above AT) unmodelled F1 and B2 stutter peaks from the profiles that showed a difference of factor of $\geq 10^3$ and reinterpreted the analysis in EFM v2.1.0 for the 2P, 3P, and 4P mixtures. The LRs of the minor contributors in the repeated profiles increased substantially, thus decreasing the differences in $\log_{10}$(LR) values observed between STRmix and original EFM runs (S11 Table).

Our intentions of leaving in the unmodelled stutters (F1 and B2) were to have identical EPGs as input files for both software especially since according to certain publications any unmodelled stutter could be explained as drop-in allelic events [39, 54]. For example, according to You and Balding [39], "All the alleles explained by the over-stutter (OS) or double-stutter (DS) models could also be explained by the drop-in model, and so it is unclear whether or not there is a material benefit from modelling DS and OS in addition to drop-in, an option that is available in likeLTD". According to Bleka et al. on the effect of applying the drop-in model to accommodate an extra allele in [54]: "Hence we observe that the implemented drop-in model in EuroForMix accommodates spurious alleles very efficiently—there is a small decrease in the LR. As expected, the larger the peak height, the greater the reduction in LR, because it impacts on heterozygote balance with other alleles." These unmodelled stutter peaks were considered in certain cases less likely to be drop-ins than alleles and therefore were considered alleles instead as observed from the profile LRs, per locus LRs, and deconvolution. A new EFM version 3.2.0 [17, 74] is now available and accounts for forward stutter in LR calculations. This new version was not available during the time of the analysis.

Unmodelled stutter peaks (F1 and B2) can be removed before interpretation to improve the fit of the model to the observed data by using stutter-type specific thresholds [68]. However, there is no guarantee that the stutter thresholds will work all the time across all the cases due to false positives (stutter peaks are left in as alleles) and/or false negatives (removing low-level alleles of the minor contributors).

We discuss an illustration in S4 File on one of the profiles shown in S11 Table. D05_RD14-0003-48_49–1;4-M3a-0.315GF is a two-person mixed GlobalFiler (GF) DNA profile with major and minor contributors from the PROVEDIt dataset with pristine DNA (a) of total DNA amount of 315pg and mixture ratio of 1:4. When the POI corresponded to the major

contributor, STRmix and EFM gave near identical profile $\log_{10}$(LR) values of 27.6 and 27.9, respectively. However, for the minor contributor position, STRmix and EFM gave profile $\log_{10}$(LR) values of 27.4 and 21.9, respectively, leading to a 5.4 difference in $\log_{10}$ scale (S4 File). A further review of the per-Locus LR tables obtained from STRmix and EFM for the minor contributor indicated that all loci had LR values favoring inclusion (i.e., LR $>$ 1), except for the D22S1045 in EFM that has been assigned a locus LR $<$ 1 (i.e., 0.001139) (S4 File). A review of the mixture profile (S4 File) indicated that the exclusionary LR at D22S1045 generated from EFM is likely due to a peak at "16" at D22S1045 which is likely an F1 of allele "15". EFMv2.1.0 did not model F1 and had accounted for "16" as being allelic in origin instead of being modeled as "drop-in" (S4 File). We removed the "16" from the input file and reinterpreted in EFMv2.1.0. The rerun gave a D22S1045 locus LR of 16.2 (S4 File) and a profile $\log_{10}$(LR) of 26.1 (S4 File), thus decreasing the discrepancy between EFM and STRmix to a factor of approximately 10.

There were cases (e.g. A03-40_41–1;4-M2U105-0.315GF; H03-48_49_50_29–1;4;4;4-M3I22-0.75GF; E03-48_49_50_29–1;4;4;4-M2I15-0.75GF; D01-50_29_30_31–1;1;2;1-M2a-0.155GF), that did not contain any instances of F1 or B2 and differed by a factor of $\geq 10^3$ when compared to the profile LR generated in EFM (highlighted in red in S11 Table). A plausible explanation for these differences will be discussed below.

III. Different modeling assumptions and parameters settings between the two software

There were instances in which EFM assigned larger LR values than STRmix (S12 Table) and cases of which STRmix profile LRs were greater than EFM LRs (highlighted in red in S11 Table and as mentioned above not due to F1 or B2). Some of these profiles in which EFM assigned larger LR values than STRmix contained instances of F1 and B2. Reinterpreting those profiles in EFMv2.1.0 with F1 and B2 removed resulted in a slight increase or had no effect on the profile LRs (S12 Table). Larger differences between the two LR systems were observed when comparing minor contributors (in most cases) with mixture profiles composed of low total template amount, low minor template amount, and/or degraded DNA (as reflected in S12 Table). In these cases, there is increase in stochastic effects, variation in peak heights, and drop-out events.

As an illustration we discuss one of the profiles shown in S12 Table. B07_RD14-0003-48_49–1;4-M3e-0.075GF is a two-person mixed GlobalFiler (GF) DNA profile with major and minor contributors from the PROVEDIt dataset with degraded DNA (DNA treated with DNase I) of total DNA amount of 75 pg, minor template amount of 15 pg, and mixture ratio of 1:4. For the minor contributor, EFM and STRmix gave profile $\log_{10}$(LR) values of 11.3 and 3.6, respectively, leading to a 7.6 difference in $\log_{10}$ scale (S5 File). A further review of the per-Locus LR tables obtained from EFM and STRmix for the minor contributor indicated that the LR of D1S1656 had the largest difference (S5 File). The known genotypes at this locus for major and minor contributors were (12,15) and (13,14), respectively, showing that allele "13" dropped-out. A review of the STRmix deconvolution indicated that the genotype at that locus (Q,14) is accepted with a low assigned weight (S5 File). The weights in STRmix are used for LR assignments [59], hence the low D1S1656 LR value.

Therefore, differences observed in profile LRs between the STRmix and EFM maybe partly influenced by the analyst's review of data and analyst's decisions when interpreting DNA typing results, different modeling assumptions and statistical models between the two software (e.g. degradation's effect on peak height, peak height variability, heterozygote balance, drop-in/drop-out, and different stutter types), parameter values settings, and how each software is implementing deconvolution and LR calculations [67]. Different analysts may make different decisions when interpreting the same EPG, thus leading to different LRs even if using the same software [37]. Upon changing models (e.g. modeling double-back and forward stutter) and/or

changing parameter values (e.g. adding a per-dye detection thresholds in EFM v3.2.0, parameters from model maker and profiling kit in STRmix generated from internal validation studies) the resulting LRs will vary to some degree. Different algorithms will also lead to different deconvolution and LR values for the same DNA profile; EFM uses maximum likelihood approaches and STRmix uses Bayesian or MCMC approaches [56].

3.7. The verbal equivalents resulting from the numeric LR values from STRmix and EFM

The numeric LR values can be accompanied by a verbal expression, a qualitative statement used in court to describe the degree of support of the findings for one of the propositions relative to the alternative proposition [75–77]. As an exercise for this study we assessed if differences in the quantitative LRs assigned by the two different LR systems resulted in the same or different verbal expressions for both the H1-true tests and H2-true tests. The LR values assigned by STRmix and EFM were binned into their corresponding verbal categories based on the verbal convention recommendations set by the Scientific Working Group on DNA Analysis Methods (SWGDAM) [78] (shown in S13 Table). The SWGDAM verbal scale is composed of 5 verbal categories: 'uninformative', 'limited', 'moderate', 'strong', and 'very strong' for both H1 and H2 support. Each category is associated with a bracket of numerical range of LR values as shown in S8 Table.

For the H1-true tests (Tables A, B, and C in Fig 8 and S14 Table), the changes in the verbal statements increased with an increase in the number of contributors. The following analysis were binned into identical verbal categories: 96.42% (297 out of 308) of the LRs from 2P mixtures, 89.11% (393 out of 441) of the LR values of 3P mixtures, and 86.61% (440 out of 508) of the LRs of the 4P mixtures. Hence, (11 out of 308) of the LRs of 2P samples, (48 out of 441) of the LRs of 3P samples, and (68 out of 508) of the LRs of 4P samples were classified into different categories (Tables A, B, and C in Fig 8). For the 11 2P cases that were different verbally, 6 were placed in the neighboring categories (for example, for the same 2P profile, an LR from one software was binned into 'moderate support' and the LR from the other software was placed in the 'strong support' category). The other 5 cases were located in non-adjacent categories and differed by two or more than two verbal categories (e.g. 'moderate support' and 'very strong support' or 'exclusionary' and 'limited' or 'exclusionary' and 'strong support' or 'exclusionary' and 'very strong support') (Table A in Fig 8). With 3P analysis, (6 cases out of 48) were classified into non-adjacent categories: 4 cases were two categories away ('exclusion' and 'limited support' or 'limited support' and 'strong support') and 2 cases were different by three categories ('exclusion' and 'moderate support') (Table B in Fig 8). For the LRs of the 4P (Table C in Fig 8) analysis that fell in different categories, only 7 out of 68 cases were different by more than one verbal category: 5 cases were different by two categories ("Exclusion" and "Limited Support" or "Limited Support" and "Strong Support" or "Very Strong Support" and "Moderate Support"), and 2 cases were three categories away: ("Exclusion" and "Moderate Support" or "Very Strong Support" and "Limited Support"). Cases of LRs with more than one category difference corresponded to H1-true tests in which POI was a minor contributor and/or had low template amount (S14 Table).

The categories used for the binning the LRs of the H2-true tests are in favor of H2 over H1 (i.e., mirror image of the verbal scale of the H1-true tests). For the H2-true tests, similarly as for the H1-true tests, as the number of contributors increased the differences in the verbal statements increased as well (Tables D, E, and F in Fig 8 and S15 Table). The following analysis were binned into the same verbal category: 80.51% (248 out of 308) of the LRs from 2P mixtures, 27.21% (120 out of 441) of the LR values of 3P mixtures, and 8.07% (41 out of 508) of the

1 bin apart | 2 bins apart | 3 bins apart | 4 bins apart | 5 bins apart

A. *2 person mixtures (H1-True Tests) N=308*

STRmix \ EFM	Exclusionary	Uninformative	Limited	Moderate	Strong	Very Strong
Exclusionary	0	0	1	0	1	[illegible]
Uninformative	1	0	0	0	0	0
Limited	0	0	2	0	0	0
Moderate	0	0	0	7	1	1
Strong	0	0	0	0	9	1
Very Strong	[illegible]	0	0	0	3	279

B. *3 person mixtures (H1-True Tests) N=441*

STRmix \ EFM	Exclusionary	Uninformative	Limited	Moderate	Strong	Very Strong
Exclusionary	3	5	2	2	0	[illegible]
Uninformative	0	0	0	0	0	0
Limited	0	1	15	8	2	0
Moderate	0	0	2	18	8	0
Strong	0	0	0	6	23	7
Very Strong	[illegible]	0	0	0	5	334

C. *4 person mixtures (H1-True Tests) N=508*

STRmix \ EFM	Exclusionary	Uninformative	Limited	Moderate	Strong	Very Strong
Exclusionary	2	4	3	1	0	[illegible]
Uninformative	0	0	7	0	0	0
Limited	0	1	27	9	1	0
Moderate	0	0	1	64	10	0
Strong	0	0	0	4	59	13
Very Strong	[illegible]	0	1	1	12	288

D. *2 person mixtures (H2-True Tests) N=308*

STRmix \ EFM	Very Strong	Strong	Moderate	Limited	Uninformative	Inclusionary
Very Strong	239	21	17	8	1	[illegible]
Strong	0	1	2	3	0	3
Moderate	0	0	2	1	1	0
Limited	0	0	0	5	1	1
Uninformative	0	0	0	0	0	0
Inclusionary	[illegible]	0	0	0	0	1

E. *3 person mixtures (H2-True Tests) N=441*

STRmix \ EFM	Very Strong	Strong	Moderate	Limited	Uninformative	Inclusionary
Very Strong	107	53	77	74	9	[illegible]
Strong	0	0	7	21	8	2
Moderate	0	0	1	12	14	14
Limited	0	0	0	10	7	13
Uninformative	0	0	0	0	0	0
Inclusionary	[illegible]	0	0	0	0	2

F. *4 person mixtures (H2-True Tests) N=508*

STRmix \ EFM	Very Strong	Strong	Moderate	Limited	Uninformative	Inclusionary
Very Strong	11	15	33	71	32	[illegible]
Strong	2	0	2	34	15	14
Moderate	0	0	2	41	37	27
Limited	1	0	0	16	32	50
Uninformative	0	0	0	0	0	6
Inclusionary	[illegible]	0	0	0	2	12

Fig 8. Concordance/discordance tables of the binned LR values assigned by STRmix and EFM into their verbal equivalents. The tables display the results of the categorization of the LRs for both the H1-true tests of (A) 2P where ** are the two 2P H1-true test interpretations for which STRmix assigned profile LRs of 0 (binned into the Exclusionary category and discussed in detail in Section 3.6), (B) 3P, and (C) 4P and H2-true tests of (D) 2P, (E) 3P, and (F) 4P generated in STRmix and EFM into their corresponding verbal expression. Also, the tables demonstrate the observed differences in the verbal expressions between the two LR systems. The number of cases that resulted in same verbal expression between STRmix and EFM fell inside the diagonal (white cells). All the numbers outside the diagonal (shaded cells) are indication of cases where LRs from both software were classified into different categories and resulted in shifting by one or more than one verbal category (indicated by different shades as shown by the legend). Values in and above the diagonal are the results of the verbal expression of LRs produced in EFM while values in and below the diagonals are the results of the verbal expression of LR values assigned by STRmix. The verbal expressions are shown at the top and left edges of the tables.

https://doi.org/10.1371/journal.pone.0256714.g008

LRs of the 4P mixtures (Tables D, E, and F in Fig 8). For the 60 2P cases that were different verbally, 35 were placed in non-neighboring categories (Table D in Fig 8). With 3P and 4P analysis, (242 cases out of 321) and (367 cases out of 467), respectively, were classified into non-adjacent categories (Tables E and F in Fig 8).

4. Conclusion

In this independent study, we examined the discrimination performance as well as LR values assigned by two LR systems using two continuous PGS built on different modelling assumptions, STRmix (proprietary) and EFM (open-source) [7, 56]. We use the term LR system deliberately to emphasize that the assigned LR values are a product of the decisions that went into the interpretation process of the LR system and not solely the PGS. For example, our specific choice of the PROVEDIt filtered files, protocols used for the data analysis in both STRmix and EFM, decision to use the known NOC, and to provide similar data (EPGs) into both software are specific to "our" LR system used in this study. We recognize that alternative decisions could have been made, and thus different LR values could have been assigned. We described the degree of differences in the LR values, where the differences occur, and the potential explanations for the observed differences. We analyzed 154 2P, 147 3P, and 127 4P mixture profiles from PROVEDIt database [43, 44] of varying DNA quality, DNA quantity, and mixture ratios (shown in S4 and S5 Tables). Both H1-true tests (S4 Table) and H2-true tests (S5 Table) for the 2P, 3P, and 4P were analyzed in both STRmix and EFM yielding a total of 1,257 of known-contributor LRs and 1,257 of known non-contributor LRs from each software.

The discrimination performance was evaluated qualitatively (Fig 2) and quantitatively (Fig 4) by checking the ability of each LR system in discriminating between H1-true and H2-true scenarios. The overall distribution plots (Fig 2) and ROC plots (Fig 4) suggest that the ability of the two LR systems to discriminate between known contributors and known non-contributors in aggregate are statistically indistinguishable for the data we considered.

Although both LR systems had similar discrimination performance, that did not imply that STRmix and EFM assigned equal LR values on a case-by-case basis even though LR computations were based on same/fixed EPG features, same pair of propositions, NOC, theta, and population allele frequency (Fig 5). The magnitude of differences was broken down into factor of 10 bins (Fig 6) and stratified by the type of POI (Fig 7). Differences in LR values greater than or equal to 3 on the $\log_{10}$ scale (as discussed in Section 3.6) were investigated and could occur due to one or more of the following reasons:

1. decisions made during parameters settings (e.g. choice of profiles for Model Maker interpretation and choice of settings for analysis such as analytical thresholds and drop-in parameters)
2. decision to analyze the same input files in both STRmix and EFM of which some of these profiles contained stutter peaks (F1 and B2) that were not modelled by EFM v2.1.0
3. non-convergence of the MCMC algorithms
4. differences in modelling assumptions of peak height information and variability, degradation, heterozygote balance, and allelic drop-outs/drop-ins

It is important to note that the apparent differences observed due to mentioned factors (2) and (3) were reduced upon re-interpretation of data both manually and in the software (e.g. re-interpreting profiles in EFM after removing the unmodelled F1 and B2 (S11 Table) or repeating analysis in STRmix with higher number of accepts (S8 Table).

Irrespective of the quantitative differences observed in certain cases between the LR systems (Fig 5), there seems to be a pattern observed in this study. Differences in LR values were observed in both directions (e.g., when LR STRmix $\geq$ 1000*LR EFM or when LR EFM $\geq$ 1000*LR STRmix). The magnitude of the differences was greater with minor donors than with equal or major contributors (Fig 7 and S11 and S12 Tables). Similar observations were documented in [34, 42, 62] when comparing LRs from various models.

Both LR systems showed adventitious exclusionary LR values (LR < 1) for H1-true tests (mainly with minor contributors) (Fig 3 and S6 Table) and adventitious inclusionary LR values (LR > 1) for H2-true tests (Fig 3 and S7 Table). The largest LR assigned using our LR systems and dataset was 587 from STRmix and 167 from EFM for a known unrelated non-contributor in the 1,257 H2-true tests (Fig 3C and S7 Table).

We observed that in certain cases differences in numerical LR values from both software resulted in differences in one or more than one verbal categories (Fig 8). These differences were substantially more with low template minor contributors and higher NOC (Fig 8 and S14 and S15 Tables); observations that have as well been examined in Swaminathan et al. [34]. Also, the cases of differences in the numerical LR values and verbal classification of the H2-true tests between the two models were higher than the ones observed with H1-true tests (Figs 6–8), thus showing the differences in the ability of both models to evaluate/measure the strength of evidence. The comparison of the assigned LR values in the verbal scale framework was included to provide some context to the observed differences. Although interesting, observed differences greater than 10^3 may have less practical impact for large LRs (e.g. 10^{15} versus 10^{18}) as compared to smaller LRs (e.g. 10^1 versus 10^4).

The findings of this study are specific to the LR systems (Fig 1) used in our study: (i) data chosen to generate parameter values and settings for analysis (e.g. Model Maker, analytical thresholds, drop-in, stutter settings), (ii) decisions made prior to the analysis of the mixture profiles in both software, and (iii) mixture profiles used for LR assessments. The profiles used for generating parameter values are shown in S1, S2, and S3 Tables. We also share with the forensic community the mixture profiles used for H1-true and H2-true tests with their corresponding LR values from both LR systems (S4 and S5 Tables). The comparisons performed in this study are more extensive than any software comparisons previously reported [34, 39–42, 54, 61, 62, 64, 68, 79]. The included supplementary tables and figures are intended to provide an example of the level of information and transparency we desire to see in similar DNA mixture publications. This provides the opportunity to review a specific mixture profile and further examine the assigned LR value(s). We believe that sharing the assigned LR values correlated with each mixture vs POI comparison complements the global aggregate level ROC and scatter plots used to assess the LR systems. This was further enabled by using the publicly available and consented PROVEDIt mixture profiles (i.e., the sharing of DNA profiles was not an issue). We encourage other investigators to assess the PROVEDIt profiles with *their* LR systems, compare their assigned LR values to those obtained in this study, and/or develop further visualization tools.

To sum up, "there are no true likelihood ratios, just like there are no true models" [80] and "no model perfectly incorporates all sources of uncertainty" [67]. The focus of this study is not to suggest that any one of the software is based on a true or best model. Our intent is to (i) understand the variability in LR values across different PG models, (ii) demonstrate the value of using a publicly available ground truth known mixture data [44] to assess performance of any LR system, (iii) describe how examining more than one PGS with similar discrimination power can be beneficial and an additional empirical diagnostic check even if software in use does contain certain diagnostic statistics as part of the output, (iv) share our observations with the forensic community that can lead to improving one or both models, and (v) address "Under what circumstances—and why—does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?", as recommended by the President's Council of Advisors on Science and Technology (PCAST) report [81].

Supporting information

S1 File. Distribution of $\log_{10}$(LR) values for H1-true and H2-true tests of two, three, and four person mixtures by software and mixture ratios.
(PPTX)

S2 File. Distribution of $\log_{10}$(LR) values for H1-true and H2-true tests of two, three, and four person mixtures by software and varying DNA treatments.
(PPTX)

S3 File. Model examination.
(PPTX)

S4 File. An illustration of an example of a 2-person mixture profile of which LR (STRmix) > 1000*LR (EFM).
(PPTX)

S5 File. An illustration of an example of a 2-person mixture profile of which LR (EFM) > 1000*LR (STRmix).
(PPTX)

S1 Table. Single source sample profiles used in determining AT values.
(XLSX)

S2 Table. Negative control profiles used in determining drop-in parameters.
(XLSX)

S3 Table. Single source profiles included in the Model Maker analysis with varying range of DNA quality and quantity.
(XLSX)

S4 Table. Total H1-true calculations using mixtures with different ground truth number of contributors (NOC), total template amounts, type of POI (major, minor, and equal), and mixture ratios analyzed in both STRmix and EFM.
(XLSX)

S5 Table. Total H2-true calculations using mixtures with different ground truth number of contributors (NOC), total template amounts, and mixture ratios analyzed in both STRmix and EFM.
(XLSX)

S6 Table. Cases of 2P, 3P, and 4P mixture profiles with adventitious exclusionary LRs that resulted from H1-true tests in STRmix and EFM.
(XLSX)

S7 Table. Cases of 2P, 3P, and 4P mixture profiles with adventitious inclusionary LRs that resulted from H2-true tests in STRmix and EFM.
(XLSX)

S8 Table. Profile Log10(LRs) and per locus LRs of the H1-true tests that generated adventitious exclusionary LR values (LR = 0 or Log10(LR) = undefined) in STRmix when compared to known minor contributors due to a zero LR at a single locus.
(XLSX)

S9 Table. Profile Log10(LRs) of a 2P mixture profile that generated adventitious exclusionary LR value (LR < 1) in STRmix when compared to known minor contributor "42" due to

artifact peaks retained in the input file.
(XLSX)

S10 Table. Diagnostics of Gelman-Rubin (GR) statistics.
(XLSX)

S11 Table. Overview of the H1-true calculations where LR (STRmix) $\geq$ 1000*LR (EFM).
(XLSX)

S12 Table. Overview of the H1-true calculations where LR (EFM) $\geq$ 1000*LR (STRmix).
(XLSX)

S13 Table. The SWGDAM verbal scale for the expression of the likelihood ratios.
(XLSX)

S14 Table. Verbal equivalents of the numeric LR values assigned by STRmix and EFM for the 2P, 3P, and 4P true contributor analysis (H1-true tests) based on the verbal convention recommendations set by the SWGDAM.
(XLSX)

S15 Table. Verbal equivalents of the numeric LR values assigned by STRmix and EFM for the 2P, 3P, and 4P true non-contributor analysis (H2-true tests) based on the verbal convention recommendations set by the SWGDAM.
(XLSX)

Acknowledgments

The authors would like to thank John Butler and Arun Moorthy at NIST for critically reading the manuscript. The authors would also like to thank Øyvind Bleka (Oslo University Hospital), Zane Kerr (Environmental Science and Research), and Steven Myers (CAL DOJ) for their input on using the software and meaningful discussions on mixture data analysis.

Author Contributions

Conceptualization: Sarah Riman, Hari Iyer, Peter M. Vallone.

Data curation: Hari Iyer.

Formal analysis: Sarah Riman, Hari Iyer.

Funding acquisition: Peter M. Vallone.

Investigation: Sarah Riman.

Methodology: Sarah Riman, Hari Iyer, Peter M. Vallone.

Resources: Peter M. Vallone.

Software: Hari Iyer.

Supervision: Peter M. Vallone.

Visualization: Sarah Riman, Hari Iyer, Peter M. Vallone.

Writing – original draft: Sarah Riman.

Writing – review & editing: Sarah Riman, Hari Iyer, Peter M. Vallone.

References

1. SWGDAM. Guidelines for the validation of probabilistic genotyping systems. 2015.
2. Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright JA, et al. Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. Forensic science international Genetics. 2017; 29:126–44. https://doi.org/10.1016/j.fsigen.2017.04.004 PMID: 28504203
3. Bright JA, Taylor D, Gittelson S, Buckleton J. The paradigm shift in DNA profile interpretation. Forensic science international Genetics. 2017; 31:e24–e32. https://doi.org/10.1016/j.fsigen.2017.08.005 PMID: 28838643
4. Buckleton JS, Bright JA, Gittelson S, Moretti TR, Onorato AJ, Bieber FR, et al. The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity. Journal of forensic sciences. 2019; 64 (2):393–405. https://doi.org/10.1111/1556-4029.13898 PMID: 30132900
5. Kelly H, Bright JA, Buckleton JS, Curran JM. A comparison of statistical models for the analysis of complex forensic DNA profiles. Science & justice: journal of the Forensic Science Society. 2014; 54(1):66–70. https://doi.org/10.1016/j.scijus.2013.07.003 PMID: 24438780
6. Coble MD, Bright JA. Probabilistic genotyping software: An overview. Forensic science international Genetics. 2019; 38:219–24. https://doi.org/10.1016/j.fsigen.2018.11.009 PMID: 30458407
7. Bright JA, Taylor D, McGovern C, Cooper S, Russell L, Abarno D, et al. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic science international Genetics. 2016; 23:226–39. https://doi.org/10.1016/j.fsigen.2016.05.007 PMID: 27235797
8. Taylor D, Bright JA, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic science international Genetics. 2013; 7(5):516–28. https://doi.org/10.1016/j.fsigen.2013.05.011 PMID: 23948322
9. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele® DNA mixture interpretation. Journal of forensic sciences. 2011; 56(6):1430–47. https://doi.org/10.1111/j.1556-4029.2011.01859.x PMID: 21827458
10. Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic science international Genetics. 2009; 4(1):1–10. https://doi.org/10.1016/j.fsigen.2009.03.003 PMID: 19948328
11. STRmix™ forensic software Available from: https://www.strmix.com/.
12. TrueAllele ® DNA Interpretation. Available from: https://www.cybgen.com/.
13. MaSTR™ Software. Available from: https://softgenetics.com/MaSTR.php.
14. GenoProof Mixture 3. Available from: https://www.qualitype.de/en/solutions/products/evaluation-software/genoproof-mixture/.
15. The DNA·VIEW® Mixture Solution. Available from: http://dna-view.com/.
16. LiRa. Available from: https://cdnmedia.eurofins.com/european-west/media/1418957/lgc_lira_fact_sheet_en_0815_90.pdf.
17. EuroForMix. Available from: http://www.euroformix.com/.
18. Swaminathan H, Garg A, Grgicak CM, Medard M, Lun DS. CEESIt: A computational tool for the interpretation of STR mixtures. Forensic science international Genetics. 2016; 22:149–60. https://doi.org/10.1016/j.fsigen.2016.02.005 PMID: 26946255
19. likeLTD (likelihoods for low-template DNA profiles). Available from: https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.
20. DNAmixtures. Available from: http://dnamixtures.r-forge.r-project.org/.
21. Kongoh. Available from: https://github.com/manabe0322/Kongoh.
22. BulletProof probabilistic genotyping software. Available from: http://ednalims.com/probabilistic-genotyping/.
23. DNAxs/DNAStatistX. Available from: https://www.forensicinstitute.nl/research-and-innovation/international-projects/dnaxs.
24. Bright JA, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic science international Genetics. 2013; 7(2):296–304. https://doi.org/10.1016/j.fsigen.2012.11.013 PMID: 23317914
25. Puch-Solis R, Rodgers L, Mazumder A, Pope S, Evett I, Curran J, et al. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic science international Genetics. 2013; 7(5):555–63. https://doi.org/10.1016/j.fsigen.2013.05.009 PMID: 23948327
26. Cowell RG, Lauritzen SL, Mortera J. A gamma model for DNA mixture analyses. Bayesian Anal. 2007; 2(2):333–48.
27. Cowell R, Graversen T, Lauritzen S, Mortera J. Analysis of forensic DNA mixtures with artefacts. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2015; 64(1):1–48.

28. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic science international. 1994; 64(2–3):125–40. https://doi.org/10.1016/0379-0738(94)90222-4 PMID: 8175083

29. Lindley DV. A problem in forensic science. Biometrika. 1977; 64(2):207–13.

30. Evett IW, Buffery C, Willott G, Stoney D. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. Journal—Forensic Science Society. 1991; 31(1):41–7. https://doi.org/10.1016/s0015-7368(91)73116-2 PMID: 1856673

31. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr W, et al. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. Forensic science international. 2006; 160(2–3):90–101. https://doi.org/10.1016/j.forsciint.2006.04.009 PMID: 16750605

32. Jackson G, Jones S, Booth G, Champod C, Evett IW. The nature of forensic science opinion—a possible framework to guide thinking and practicce in investigation and in court proceedings. Science & Justice. 2006; 46(1):33–44.

33. Cooper S, McGovern C, Bright JA, Taylor D, Buckleton J. Investigating a common approach to DNA profile interpretation using probabilistic software. Forensic science international Genetics. 2015; 16:121–31. https://doi.org/10.1016/j.fsigen.2014.12.009 PMID: 25596557

34. Swaminathan H, Qureshi MO, Grgicak CM, Duffy K, Lun DS. Four model variants within a continuous forensic DNA mixture interpretation framework: Effects on evidential inference and reporting. 2018; 13 (11):e0207599.

35. Hannig J, Riman S, Iyer H, Vallone PM. Are reported likelihood ratios well calibrated? Forensic Science International: Genetics Supplement Series. 2019; 7(1):572–4.

36. Kelly H, Bright JA, Kruijver M, Cooper S, Taylor D, Duke K, et al. A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration. Forensic science international Genetics. 2018; 35:113–22. https://doi.org/10.1016/j.fsigen.2018.04.009 PMID: 29727813

37. Barrio PA, Crespillo M, Luque JA, Aler M, Baeza-Richer C, Baldassarri L, et al. GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: Results and evaluation. Forensic science international Genetics. 2018; 35:156–63. https://doi.org/10.1016/j.fsigen.2018.05.005 PMID: 29783171

38. Bright JA, Cheng K, Kerr Z, McGovern C, Kelly H, Moretti TR, et al. STRmix™ collaborative exercise on DNA mixture interpretation. Forensic science international Genetics. 2019; 40:1–8. https://doi.org/10.1016/j.fsigen.2019.01.006 PMID: 30665115

39. You Y, Balding D. A comparison of software for the evaluation of complex DNA profiles. Forensic science international Genetics. 2019; 40:114–9. https://doi.org/10.1016/j.fsigen.2019.02.014 PMID: 30798114

40. Alladio E, Omedei M, Cisana S, D'Amico G, Caneparo D, Vincenti M, et al. DNA mixtures interpretation—A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples. Forensic science international Genetics. 2018; 37:143–50. https://doi.org/10.1016/j.fsigen.2018.08.002 PMID: 30173123

41. Buckleton JS, Bright JA, Cheng K, Budowle B, Coble MD. NIST interlaboratory studies involving DNA mixtures (MIX13): A modern analysis. Forensic science international Genetics. 2018; 37:172–9. https://doi.org/10.1016/j.fsigen.2018.08.014 PMID: 30176439

42. Manabe S, Morimoto C, Hamano Y, Fujimoto S, Tamaki K. Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model. 2017; 12(11): e0188183.

43. Alfonse LE, Garrett AD, Lun DS, Duffy KR, Grgicak CM. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. Forensic science international Genetics. 2018; 32:62–70. https://doi.org/10.1016/j.fsigen.2017.10.006 PMID: 29091906

44. PROVEDIt Database. Available from: https://lftdi.camden.rutgers.edu/provedit/files/.

45. Manual for EuroForMix v2.1 (2019). Available from: http://www.euroformix.com/sites/default/files/euroformixManual_v2_1.pdf.

46. STRmix v2.6.0 Operation Manual (2018). Available from: https://support.strmix.com/.

47. Butler JM. Advanced Topics in Forensic DNA Typing: Interpretation: Elsevier, Amsterdam; 2015.

48. Butler JM, Iyer HK. Validation, Principles, Practices, Parameters, Performance, Evaluations, and Protocols ISHI 2020 Validation Workshop. Available from: https://strbase.nist.gov/pub_pres/ISHI2020-ValidationWorkshop-Butler_Iyer-Slides.pdf.

49. Gilder JR, Doom TE, Inman K, Krane DE. Run-specific limits of detection and quantitation for STR-based DNA testing. Journal of forensic sciences. 2007; 52(1):97–101. https://doi.org/10.1111/j.1556-4029.2006.00318.x PMID: 17209918

50. Bregu J, Conklin D, Coronado E, Terrill M, Cotton RW, Grgicak CM. Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis. Journal of forensic sciences. 2013; 58 (1):120–9. https://doi.org/10.1111/1556-4029.12008 PMID: 23130820

51. Taylor D, Bright JA, McGoven C, Hefford C, Kalafut T, Buckleton J. Validating multiplexes for use in conjunction with modern interpretation strategies. Forensic science international Genetics. 2016; 20:6–19. https://doi.org/10.1016/j.fsigen.2015.09.011 PMID: 26433484

52. Mönich UJ, Duffy K, Médard M, Cadambe V, Alfonse LE, Grgicak C. Probabilistic characterisation of baseline noise in STR profiles. Forensic science international Genetics. 2015; 19:107–22. https://doi.org/10.1016/j.fsigen.2015.07.001 PMID: 26218981

53. STRmix v2.6.0 User Manual and Implementation/Validation Guide (2018). Available from: https://support.strmix.com/.

54. Bleka Ø, Benschop CCG, Storvik G, Gill P. A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles. Forensic science international Genetics. 2016; 25:85–96. https://doi.org/10.1016/j.fsigen.2016.07.016 PMID: 27529774

55. Bright JA, Curran JM. Investigation into stutter ratio variability between different laboratories. Forensic science international Genetics. 2014; 13:79–81. https://doi.org/10.1016/j.fsigen.2014.07.003 PMID: 25082139

56. Bleka Ø, Storvik G, Gill P. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. Forensic science international Genetics. 2016; 21:35–44. https://doi.org/10.1016/j.fsigen.2015.11.008 PMID: 26720812

57. Steffen CR, Coble MD, Gettings KB, Vallone PM. Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci' [Forensic Sci. Int. Genet. 7 (2013) e82-e83]. Forensic science international Genetics. 2017; 31:e36–e40. https://doi.org/10.1016/j.fsigen.2017.08.011 PMID: 28867528

58. Buckleton J, Curran J, Goudet J, Taylor D, Thiery A, Weir BS. Population-specific FST values for forensic STR markers: A worldwide survey. Forensic science international Genetics. 2016; 23:91–100. https://doi.org/10.1016/j.fsigen.2016.03.004 PMID: 27082756

59. Russell L, Cooper S, Wivell R, Kerr Z, Taylor D, Buckleton J, et al. A guide to results and diagnostics within a STRmix™ report. WIREs Forensic Science. 2019; 1(6):e1354.

60. R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2020). Available from: https://www.r-project.org/.

61. Steele CD, Greenhalgh M, Balding DJ. Evaluation of low-template DNA profiles using peak heights. Statistical applications in genetics and molecular biology. 2016; 15(5):431–45. https://doi.org/10.1515/sagmb-2016-0038 PMID: 27416618

62. Slooten K. The information gain from peak height data in DNA mixtures. Forensic science international Genetics. 2018; 36:119–23. https://doi.org/10.1016/j.fsigen.2018.06.009 PMID: 29990823

63. Taylor D, Buckleton J. Do low template DNA profiles have useful quantitative data? Forensic science international Genetics. 2015; 16:13–6. https://doi.org/10.1016/j.fsigen.2014.11.001 PMID: 25474687

64. Bille TW, Weitz SM, Coble MD, Buckleton J, Bright JA. Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. Electrophoresis. 2014; 35(21–22):3125–33. https://doi.org/10.1002/elps.201400110 PMID: 25168355

65. Taylor D. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic science international Genetics. 2014; 11:144–53. https://doi.org/10.1016/j.fsigen.2014.03.008 PMID: 24727432

66. Noël S, Noël J, Granger D, Lefebvre JF, Séguin D. STRmix(™) put to the test: 300 000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates. Forensic science international Genetics. 2019; 41:24–31. https://doi.org/10.1016/j.fsigen.2019.03.017 PMID: 30947115

67. Taylor D, Balding D. How can courts take into account the uncertainty in a likelihood ratio? Forensic science international Genetics. 2020; 48:102361. https://doi.org/10.1016/j.fsigen.2020.102361 PMID: 32769057

68. Benschop CCG, Nijveld A, Duijs FE, Sijen T. An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors. Forensic science international Genetics. 2019; 42:31–8. https://doi.org/10.1016/j.fsigen.2019.06.005 PMID: 31212207

69. Green DM, Swets JA. Signal detection theory and psychophysics. Oxford, England: John Wiley; 1966. xi, 455–xi, p.

70. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44(3):837–45. PMID: 3203132

71. Bright JA, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™ - A multi laboratory response to PCAST. Forensic science international Genetics. 2018; 34:11–24. https://doi.org/10.1016/j.fsigen.2018.01.003 PMID: 29367014

72. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statist Sci. 1992; 7 (4):457–72.

73. Bleka Ø. An introduction to EuroForMix 2017. Available from: http://euroformix.com/sites/default/files/EuroForMixTheory_ISFG17.pdf.

74. Bleka Ø. New update of EuroForMix Version 3.0.0.

75. Gill P, Hicks T, Butler JM, Connolly E, Gusmão L, Kokshoorn B, et al. DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence—Guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositions. Forensic science international Genetics. 2020; 44:102186. https://doi.org/10.1016/j.fsigen.2019.102186 PMID: 31677444

76. Aitken C, Taroni F. A verbal scale for the interpretation of evidence. Science & Justice. 1998; 8:279–81.

77. Arscott E, Morgan R, Meakin G, French J. Understanding forensic expert evaluative evidence: A study of the perception of verbal expressions of the strength of evidence. Science & justice: journal of the Forensic Science Society. 2017; 57(3):221–7. https://doi.org/10.1016/j.scijus.2017.02.002 PMID: 28454631

78. SWGDAM. Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios (2018). Available from: https://docs.wixstatic.com/ugd/4344b0_dd5221694d1448588dcd0937738c9e46.pdf.

79. Bright JA, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic science international Genetics. 2015; 14:125–31. https://doi.org/10.1016/j.fsigen.2014.09.019 PMID: 25450783

80. Gill P, Hicks T, Butler JM, Connolly E, Gusmão L, Kokshoorn B, et al. DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence—Guidelines highlighting the importance of propositions: Part I: evaluation of DNA profiling comparisons given (sub-) source propositions. Forensic science international Genetics. 2018; 36:189–202. https://doi.org/10.1016/j.fsigen.2018.07.003 PMID: 30041098

81. President's Council of Advisors on Science and Technology. Report to the President—Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016). Available from: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Forensic Science International: Genetics 22 (2016) 54–63

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[f], Katherine B. Gettings[f], Peter Gill[g,h], Leonor Gusmão[i,j,k], Douglas R. Hares[l], Jodi A. Irwin[l], Jonathan L. King[d], Peter de Knijff[m], Niels Morling[n], Mechthild Prinz[o], Peter M. Schneider[p], Christophe Van Neste[q], Sascha Willuweit[r], Christopher Phillips[s]

[a] Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria
[b] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA
[c] Faculty of Life Sciences, King's College, London, UK
[d] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia
[f] National Institute of Standards and Technology, Gaithersburg, MD, USA
[g] Norwegian Institute of Public Health, Department of Forensic Biology, Oslo, Norway
[h] Department of Forensic Medicine, University of Oslo, Oslo, Norway
[i] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil
[j] IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal
[k] Instituto de Investigação e Inovação em Saúde, University of Porto, Portugal
[l] FBI Laboratory, Quantico, VA, USA
[m] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[n] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[o] Department of Sciences, John Jay College for Criminal Justice, New York, NY, USA
[p] Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne, Germany
[q] Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
[r] Institute of Legal Medicine, Humboldt University, Berlin, Germany
[s] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 14 January 2016
Accepted 16 January 2016
Available online 21 January 2016

Keywords:
Massively parallel sequencing
MPS
Next generation sequencing
NGS
Short tandem repeats
STRs
Nomenclature

ABSTRACT

The DNA Commission of the International Society for Forensic Genetics (ISFG) is reviewing factors that need to be considered ahead of the adoption by the forensic community of short tandem repeat (STR) genotyping by massively parallel sequencing (MPS) technologies. MPS produces sequence data that provide a precise description of the repeat allele structure of a STR marker and variants that may reside in the flanking areas of the repeat region. When a STR contains a complex arrangement of repeat motifs, the level of genetic polymorphism revealed by the sequence data can increase substantially. As repeat structures can be complex and include substitutions, insertions, deletions, variable tandem repeat arrangements of multiple nucleotide motifs, and flanking region SNPs, established capillary electrophoresis (CE) allele descriptions must be supplemented by a new system of STR allele nomenclature, which retains backward compatibility with the CE data that currently populate national DNA databases and that will continue to be produced for the coming years. Thus, there is a pressing need to produce a standardized framework for describing complex sequences that enable comparison with currently used repeat allele nomenclature derived from conventional CE systems. It is important to discern three levels of information in hierarchical order (i) the sequence, (ii) the alignment, and (iii) the nomenclature of STR sequence data. We propose a sequence (text) string format the minimal requirement of data storage that laboratories should follow when adopting MPS of STRs. We further discuss the variant annotation and sequence comparison framework necessary to maintain compatibility among established and future data. This system must be easy to use and interpret by the DNA specialist,

* Corresponding author at: Medical University of Innsbruck, Muellerstr. 44, Innsbruck 6020, Austria.
E-mail address: walther.parson@i-med.ac.at (W. Parson).

http://dx.doi.org/10.1016/j.fsigen.2016.01.009

based on a universally accessible genome assembly, and in place before the uptake of MPS by the general forensic community starts to generate sequence data on a large scale. While the established nomenclature for CE-based STR analysis will remain unchanged in the future, the nomenclature of sequence-based STR genotypes will need to follow updated rules and be generated by expert systems that translate MPS sequences to match CE conventions in order to guarantee compatibility between the different generations of STR data.

1. Introduction

Short tandem repeats (STRs) were introduced as polymorphic DNA loci in the forensic field in the early 1990s [1,2] and have become the primary workhorse for individual identification in criminal casework, paternity analyses, and identification of missing persons [3,4]. The STR loci used in forensic DNA analysis were selected using stringent criteria (e.g. [5]). Later, core loci were defined with broad overlap among international legislations [6]. Allele categories have been identified by PCR-based amplicon sizing methods and gel or capillary electrophoretic (CE) systems [3] following simple nomenclature convention [7–9]. Size categories were operationally called relative to sequenced alleles that made up the allelic ladders, with integer values indicating the number of complete repeat motifs and additional nucleotides (i.e. incomplete repeats) separated by a decimal point (e.g. TH01 9.3 [7]). This convention was based on the observed variation generated by CE systems; however, it does not account for sequence differences between alleles that may be caused by transversions, transitions, insertions, deletions, and inversions of one or more nucleotides, including repetitive motifs. Nevertheless, this nomenclature is quite robust, having been adopted universally. In addition, the discrimination power of size-based alleles has proved to be sufficiently high to give useful information for forensic genetic purposes, and even more so with the introduction of large multiplexes [10,11].

Massively parallel sequencing (MPS) is adding a new dimension to the field of forensic genetics, providing distinct advantages over CE systems in terms of captured information, multiplex sizes, and analyzing highly degraded samples [12–14]. In recent years, MPS has been applied to the generation of STR sequence data [15–19] with the general outcome that STRs can be successfully typed producing genotypes compatible with those of CE analyses, even from compromised forensic samples [20]. Furthermore, MPS derived STR genotypes provide additional information to that generated by CE separation by capturing the full nucleotide sequence underlying the repeat units and nearby flanking regions. It was demonstrated by earlier studies using mass spectrometric (MS) systems that the discrimination power of STR typing could be increased by differentiating the nucleotide sequences of alleles with identical size [21–23]. With MPS, forensic tests will further discern STR variants that cannot be distinguished by MS, e.g. repeat motifs that are shifted relative to each other in the repeat region [22]. Early assessments of MPS STR typing show it will be highly beneficial to routine casework by increasing the discrimination power, improving resolution of mixtures, and enhancing the identification of stutter peaks and artifacts [12,18].

However, MPS STR analysis poses challenges to the forensic practitioner. The new technology will affect how the data are analyzed and reported, as well as how they should be stored and searched in databases. This is on top of the necessity to store raw MPS data at the laboratory level. Sequence-based STR variants are more complex and the previously defined nomenclature guidelines do not accommodate the additional variation. While the field is still learning about the sequence variation observed to date and has begun to develop strategies to harmonize nomenclature [24] some laboratories are starting to develop their own large-scale population studies to provide a basis for the introduction of MPS into forensic practice.

For the above reasons, the executive board of the ISFG decided to introduce a DNA commission to evaluate initial considerations regarding STR nomenclature. The primary goal is to define minimum criteria for data analyses and database storage. Ultimately, this should facilitate compatibility between MPS STR data generated currently and the data that will inevitably follow with wider adoption, while ensuring backward and parallel compatibility to the millions of profiles derived from CE-based STR typing in national DNA databases as well as published population data. At present, it can be expected that both CE- and MPS-based STR typing methods will continue to coexist. Their application to casework will depend on laboratory-specific considerations, such as resources, ease of use, speed of analysis, the value of the increased resolution power, and each technique's relevance to complex and challenging cases.

This paper discusses the scientific issues concerning the use of MPS technology for STR typing in forensics and highlights relevant points that should be considered to maintain compatibility of data between technological generations and within and among countries. The adoption of sequenced STR alleles in practical forensic work requires considerations at three hierarchical levels: the full sequence, i.e. the sequence string (Section 2), alignment of sequences relative to a reference sequence (Section 3), and annotation of alleles (Section 4).

2. MPS STR typing and sequence strings

With the application of MPS, the molecular genetic analysis of forensically relevant STR loci results in full nucleotide sequences that harbor the maximum discrimination power possible with DNA-based analyses. The most comprehensive representation of such data is the entire text string of sequenced nucleotides capturing all the information—the sequence string. This string is often referred to as the 'FASTA format', which derives from a more comprehensive and complex 'FASTQ format' that is produced from the raw data of MPS analysis software. It has already been demonstrated that the sequence string is the most convenient and reliable system for storing mitochondrial DNA sequences in database format, as both storage and search tasks become disentangled from alignment and notation (see [25] for mitochondrial DNA sequence strings held in EMPOP [26]). The established analysis regimes for mitochondrial DNA data demonstrate that sequences are not missed in searches performed with an alignment-free format [25], a feature that is particularly desirable and relevant in the forensic field. However, the format of sequence strings is unwieldy when reporting mitochondrial or STR variation in expert reports and cannot be communicated and compared easily without dedicated software.

Consideration 1. MPS analysis should be performed with software that allows STR sequences to be exported and stored in databases as sequence (text) strings to capture the maximum consensus sequence information.

3. Alignment of STR sequences

The forensic community is currently discussing diverse approaches to designate new MPS-based STR data in a suitably compact format. The proposed systems for defining STR sequence variation vary with respect to their complexity and information content. They share the common requirement that they must all be compatible with the existing CE-based STR data (backward compatibility) that populate current forensic databases worldwide. These approaches involve comparison to a reference sequence, a feature that is common practice in the field of mitochondrial DNA sequencing.

3.1. Reference sequences

3.1.1. Lessons learned from mitochondrial DNA

In a discussion about the use of reference sequences to report STR variability, the experience gained with other markers historically reported with respect to a reference sequence is worth revisiting. In the 1990s, the forensic community successfully adopted the concept of using a reference sequence to communicate and report mitochondrial DNA haplotypes [27,28]. The decision to use the first human mitochondrial sequence produced in 1981 [29] as the reference was practically based and was compatible with other fields of research. Every newly generated (partial) mitochondrial DNA sequence was reported relative to this first mitochondrial sequence, known as the Cambridge Reference Sequence (CRS). Eighteen years later, the same source DNA was re-sequenced with improved sequencing technology and alignment software, which resulted in the publication of the revised Cambridge Reference Sequence (rCRS, [30]). The rCRS contains corrections at eleven positions, ten of which were base substitutions at positions 3423T, 4985A, 9559C, 11335C, 13702C, 14199T, 14272C, 14365C, 14368C, and 14766C relative to the CRS. One additional difference was observed at positions 3106 and 3107, where two Cs were recorded in the CRS but only one C was determined in the rCRS. Practically, this means that the rCRS is shorter than the CRS by one nucleotide (16,568 vs. 16,569 total nucleotides). Instead of adjusting all positions downstream of 3107 (or 3106) in their numbering, this position is indicated in the rCRS as a gap [30]. This pragmatic decision allows the numbering system employed for the CRS and by the body of earlier established data to continue to be used unadjusted with the rCRS and subsequent studies.

More recently, the switch to a new mitochondrial DNA reference sequence was proposed. In contrast to the phylogenetically modern rCRS, the proposed sequence represents the deepest root in the known human mtDNA phylogeny (Reconstructed Sapiens Reference Sequence; RSRS [31]). Despite some appealing features of the RSRS, especially with respect to the interpretation of ancient and derived mutations, the forensic community has not adopted it for a number of reasons [32]. Most importantly, lack of adoption eliminates the risk of introducing error as a consequence of the translation between different versions of the mitochondrial reference sequence, especially when comparisons are performed manually. However, the decision was also based on the potential lack of stability of the RSRS that could produce unforeseen consequences for the forensic field [33].

The lessons learned in the field of mitochondrial DNA demonstrate that an established nomenclature system can remain stable and be employed by the forensic community even though (length) changes in the reference sequence were detected (in the shift from CRS to rCRS). As more laboratories begin to use MPS, numerous new STR variants will be discovered. Therefore, it is important to stress that an adapted STR allele nomenclature framework needs to be both flexible and stable in the forensic field. This functionality is easiest to achieve if the nomenclature is 'natural', i.e. is derived from the sequence of the allele.

3.1.2. Choice of a reference framework to define STR sequence variation

For any future STR nomenclature scheme, it is necessary to define which of the two DNA strands is reported and to harmonize this criterion so that a universal approach is applied to sequence alignment and comparisons. In contrast to earlier STR nomenclature guidelines that gave general preference to reporting of the coding region strand [7], we propose standardized use of one strand direction. This approach can be framed in a straightforward way by reference to the current standardized genome assembly (the term 'build' also is used for a full genome sequence construction, but builds can be short-lived and create multiple numbers within one assembly). A genome assembly assigns each nucleotide a unique chromosome coordinate that positions it precisely in the sequence and follows the system universally applied to locating genomic features such as Single Nucleotide Polymorphisms (SNPs) and Insertions/Deletions (InDels). Genomic coordinates are coded by integers denoting chromosome:position and in the human genome run from the start of the chromosome 1 p-arm to the end of the chromosome 22 q-arm (i.e. 1:1 to 1:248956422 through to 22:1 to 22:50818468 in the autosomal sequences of the most recent genome assembly GRCh38) with equivalent values for the X and Y chromosomes. These genomic coordinates dictate that the strand direction be reported for the human genome as 5′ to 3′—often referred to as "forward" or "positive". Although strand selection is sometimes arbitrary for other species (i.e. the coordinates can start at the q-arm and go towards the p-arm), in human genome mapping there is a single universal sequence direction dictated by chromosome arm length.

Use of an agreed standard human reference sequence (the reference assembly) for the nuclear portion of the genome provides the key framework from which to generate nucleotide difference-coded genotypes and to designate variants in the sequence string. At the time of writing, the current published genome assembly will be the best framework, as it represents the most accurate sequence curation, i.e. taking into account the precise mapping of complex sequence segments such as duplications and inversions. During the last three to four years, the human genetics community has worked with two human genome assemblies termed GRCh37 and GRCh38. Both GRCh37 and GRCh38 are referenced in the three main human genome databases (NCBI Genome Browser: http://www.ncbi.nlm.nih.gov; UCSC Genome Browser: http://genome.ucsc.edu; and 1000 Genomes Browser: http://browser.1000genomes.org/Homo_sapiens/Info/Index) with data consisting of both sets of coordinates. Although the 1000 Genomes data are still aligned to the GRCh37 assembly [34], at the time of writing, all sequence data from this project are undergoing the transition to map the full human sequence and its variant positions onto the GRCh38 assembly. Therefore, the GRCh38 genome assembly currently is recommended to be the reference sequence adopted by the forensic community and the nucleotide coordinates of this assembly used to map each sequence feature when describing STR variants, whether they are differences in sequence motif, SNPs, or InDels.

Of relevance here is the fact that each MPS platform has analysis software that generates sequence alignments of forensic loci from a standardized assembly. Therefore, agreement between the forensic community and MPS system suppliers about the appropriate assembly used for sequence alignments and annotation becomes a key objective for the DNA Commission on forensic STR sequence nomenclature.

Since the translation of one set of integer values to another is relatively straightforward, it is feasible to have in place an agreed

genome assembly for all forensic markers, and retain references to the coordinates of previous assemblies. This compatibility need is important as the entire catalog of SNPs, InDels and microsatellite variants currently accessible from the 1000 Genomes variant database is positioned according to GRCh37 genomic coordinates. When the current GRCh38 assembly is eventually replaced with a new one, the (potentially) necessary transition in coordinate data can be organized within the forensic community while retaining the previous GRCh37 and GRCh38 nucleotide position data. Although genotypes based on previous assemblies could, in principle, be re-coded, the reference assembly difference between any two genotypes could instead be handled bioinformatically when necessary—e.g. at the time of a comparison between two samples. Human genome assembly changes became less frequent in recent years: GRCh38 (hg38) was introduced in December 2013; GRCh37 (hg19) February 2009; NCBI36 (hg18) March 2006; NCBI35 (hg17) May 2004; NCBI34 (hg16) February 2003. Nevertheless, the data processing infrastructure organized for forensic analysis should be prepared to accommodate inevitable changes. Future developments in genome assemblies will be monitored by the Commission and the decision whether or not to adapt the reference sequence to a new assembly will be subject to later discussion.

Consideration 2. The forward strand direction assigned in the human genome has been constant for all assemblies published since the first draft in 2001 and can be used to align STR sequences.

Consideration 3. The choice of reference sequence is crucial for standardizing STR nomenclature systems. At the time of writing, GRCh38 is the most up-to-date sequence assembly and is recommended as the framework with which to define repeat region structure for sequence alignment and for the mapping of sequence features such as SNPs. Software will be required to handle comparisons between multiple reference sequences, particularly in the short term, where sequence variants listed by 1000 Genomes currently retain GRCh37 coordinates. Continued discussions are necessary to decide whether or not to adapt to novel genome assemblies

3.2. Findings from early research on alignment

Having one agreed-upon and up-to-date genome assembly with a unified strand direction presents a logical format as the coordinate integers are ascending values that can be tracked by all forensic scientists using online access to public domain genomic databases. However, this approach is not without complications, as demonstrated by the following examples indicating that more research is required.

Out of 58 STR loci for which MPS designs have become available at the time of this writing (listed in Tables 2–4 of [35]), 23 have been designated historically on the reverse strand. In 17 of these loci, the change to the forward strand for repeat region designation results in a potential shift of the reading frame (Table 1). This shift of reading frame would be consistent with the earlier ISFG

Table 1
Twenty-three STR loci previously aligned relative to the reverse strand (past repeat region sequence) with coordinates and sequences from the current human genome reference GRCh38 [34]. Bolded nucleotides are not counted for the repeat number designation. Seventeen loci for which a potential frameshift exists when converting to forward strand are denoted with "*". The repeat region sequence based on the reference sequence direction (future repeat region sequence) maintains the same location on the reference assembly and is recommended to facilitate comparison to existing sequence data and to length-based STR types. DYS385a/b and DYF387S1a/b: when reporting the forward strand, one allele will contain the reverse complement motif of the other allele, reflecting the occurrence of inversions in each STR.

STR	Chr.	Human reference genome assembly GRCh38					Potential frameshift exists
		Location of repeat region start	Location of repeat region stop	Repeat no.	Past repeat region sequence summary	Future repeat region sequence summary	
D1S1656	1	230769616	230769683	17	[TAGA]16 [TAGG] **[TG]5**	**[CA]5** [CCTA] [TCTA]16	*
D2S1338	2	218014859	218014950	23	[TGCC]7 [TTCC]13 [GTCC] [TTCC]2	[GGAA]2 [GGAC] [GGAA]13 [GGCA]7	
FGA	4	154587736	154587823	22	[TTTC]3 [TTTT] [TTCT] [CTTT]14 [CTCC] [TTCC]2	[GGAA]2 [GGAG] [AAAG]14[AGAA] [AAAA] [GAAA]3	*
D5S818	5	123775556	123775599	11	[AGAT]11	[ATCT]11	*
CSF1PO	5	150076324	150076375	13	[AGAT]13	[ATCT]13	*
D6S1043	6	91740225	91740272	12	[AGAT]12	[ATCT]12	*
D7S820	7	84160226	84160277	13	[GATA]13	[TATC]13	
VWA	12	5983977	5984044	17	[TCTA] [TCTG]5 [TCTA]11 **TCCA TCTA**	**TAGA TGGA** [TAGA]11 [CAGA]5 [TAGA]	*
Penta E	15	96831015	96831039	5	[AAAGA]5	[TCTTT]5	*
D19S433	19	29926235	29926298	16	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]12	[CCTT]12 **CCTA** [CCTT] **CTTT** [CCTT]	*
DYS19	Y	9684380	9684443	15	[TAGA]3 **TAGG** [TAGA]12	[TCTA]12 **CCTA** [TCTA]3	*
DYS635	Y	12258860	12258951	23	[TCTA]4 [TGTA]2 [TCTA]2 [TGTA]2 [TCTA]2 [TGTA]2 [TCTA]9	[TAGA]9 [TACA]2 [TAGA]2 [TACA]2 [TAGA]2 [TACA]2 [TAGA]4	*
DYS389I	Y	12500448	12500495	12	[TCTG]3 [TCTA]9	[TAGA]9 [CAGA]3	*
DYS389II	Y	12500448	12500611	29	[TCTG]5 [TCTA]12 **48 nt.** [TCTG]3 [TCTA] 9	[TAGA]9 [CAGA]3 **48 nt.** [TAGA]12 [CAGA]5	*
DYS390	Y	15163067	15163162	24	**[TCTA]2** [TCTG]8 [TCTA]11 TCTG [TCTA] 4	[TAGA]4CAGA [TAGA]11 [CAGA]8 **[TAGA]2**	*
Y-GATA-H4	Y	16631673	16631720	12	[TAGA]12	[TCTA]12	
DYS385ab	Y	18639713	18639756	11	[GAAA]11	[TTTC]11	*
		18680632	18680687	14	[GAAA]14	[GAAA]14	
DYS460	Y	18888810	18888849	10	[GATA]10	[TATC]10	*
DYS392	Y	20471987	20472025	13	[TAT]13	[ATA]13	*
DYF387S1ab	Y	23785361	23785500	35	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13	
		25884581	25884724	36	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]13	[CTTT]13 [CTTC]10 [CTTT]2CTTC [CTTT] 2 [CTTC]4CTAC [CTTT]3	*
DXS8378	X	9402262	9402301	10	[CTAT]10	[ATAG]10	
HPRTB	X	134481506	134481561	13	[TAGA]14	[TCTA]14	
DXS7423	X	150542522	150542589	15	[TCCA]3 **TCTGTCCT** [TCCA]12	[TGGA]12 **AGGACAGA** [TGGA]3	

recommendations [7] that the repeat region begins with the first possible repeat motif. This change can cause a shift in the position of features within the motif and/or an increase in the number of apparent repeats. For example, the D19S433 locus historically has been reported on the reverse strand as an AAGG repeat interspersed with one AAAG and one TAGG that are uncounted (see first example sequence below, underlined bases are counted while bolded bases are not counted). The reverse complement consists of a CCTT repeat interspersed with one CCTA and one CTTT that are uncounted (second example sequence below). However, under earlier recommendations, the first possible repeat motif of TCCT would be reported (one nucleotide shift to the left, third example sequence below), and the interspersed feature becomes ACCT TCTT. This change could complicate comparisons to existing sequence data.

1. TGTTG AAGG **AAAG** AAGG **TAGG** AAGG AAGG AAGG AAGG AAGG AAGG AGAGA

2. TCTCT CCTT CCTT CCTT CCTT CCTT CCTT **CCTA** CCTT **CTTT** CCTT CAACA

3. TCTC TCCT TCCT TCCT TCCT TCCT TCCT TCCT **ACCT TCTT** TCCT TCAACA

At the DYS389I/II loci, the potential exists for a two nucleotide shift, which would result in the appearance of one extra repeat in the larger allele. The first two bracketed sequences below show the change from reverse to forward strand maintaining identical repeat region positions on GRCh38, while the third bracketed sequence shows the change of strand with a shifted motif, yielding an extra repeat at the 3′ end. If sequence based analysis counted this repeat while traditional CE assays did not, the results would appear discordant by one repeat unit.

Previously reported reverse strand:	$[TCTG]_5$ $[TCTA]_{12}$ 48 nt. $[TCTG]_3$ $[TCTA]_9$
Forward strand, no frame shift:	$[TAGA]_9$ $[CAGA]_3$ 48 nt. $[TAGA]_{12}$ $[CAGA]_5$
Forward strand, frame shift:	$[GATA]_9$ $[GACA]_3$ 48 nt. $[GATA]_{12}$ $[GACA]_6$

Lastly, the DYS385 a/b marker has two repeat regions located in the most recent human reference sequence at Y:18639713-18639756 and Y:18680632-18680687 (Table 1). On the forward strand the first fragment has TTTC motifs while the second one comprises an inversion of the same sequence presenting GAAA motifs. In this case, using the forward strand, it is not possible to summarize DYS385 a/b repeats by a uniform motif description as was reported in the past. In addition, it is expected that some individuals will exhibit a larger first fragment and a smaller second fragment, resulting in a genotype of, e.g. 14, 11.

These examples aptly demonstrate potential complications arising from conversion of STR loci to the forward strand. It is clearly indicated that this conversion needs to be performed by designed software once MPS has reached routine application, and not manually, as the risk of introducing error would be too high. Also, it is imperative that repeat region start and end locations be strictly defined for all STR loci employed in MPS. This work is underway in various laboratories and updates will be made available to the forensic community.

As a simple guide to the human genome reference sequence, Supplementary file S1 outlines the reference strings of the repeat regions plus 50 nucleotides of each flanking sequence of STRs that will form the next generation of MPS multiplexes or have already become established for this type of forensic DNA analysis. Supplementary file S1A details 35 autosomal STRs (12 ESS, 20 CODIS markers) in common use, and Supplementary file S1B details 29 Y-STRs plus 7 X-STRs. The SNPs and InDels currently recorded by 1000 Genomes are identified in the flanking sequences, and the most polymorphic of these flanking region variants (>10% minor allele frequencies) are summarized with pie charts.

Although the human genome assembly coordinates of GRCh37 and GRCh38 can be translated in a straightforward way, three common STRs have nucleotide differences in the repeat region sequences reported by each assembly. These are for the loci DYS437 (GRCh38 one less repeat), DYS438 (two more repeats), and DYS439 (one less repeat), each reference sequence is summarized in Supplementary file S2. These nucleotide differences illustrate the challenges that must be addressed when future human genome assemblies are published and used for STR sequence alignments of MPS data.

Lastly, during detailed examination of the human genome assembly sequences at each STR, it emerged that the forensic marker named D5S2500 is represented by two different microsatellites that each form separate components in commercial CE multiplexes (e.g. Qiagen's HD-plex (Hilden, Germany) and AGCU ScienTech's 21-plex (Wuxi, China)). Investigations of both sites reveal that D5S2500 in Qiagen's HD-plex is the correctly assigned STR name. The microsatellite targeted in AGCU ScienTech's 21-plex is not a named microsatellite at the time of writing, being positioned 1688 nucleotides further upstream. The microsatellite in the AGCU kit was originally developed as a miniSTR, incorrectly named D5S2500 and reported by Hill et al. [36]. To avoid confusion while including sequence details of each of these important forensic STRs, the locus used in Qiagen's HD-plex is labeled with its NCBI accession number D5S2500.G08468, while the locus used in AGCU ScienTech's 21-plex is coded as D5S2500.AC008791(Supplementary file S1C). Details of both D5S2500 markers are summarized in the same way as the other STRs but placed in a separate Supplementary File S1C. More thorough characterization of these two microsatellites is the subject of a separate paper in preparation.

Consideration 4. Further work is needed to translate the nomenclature of STR loci thus far coded relative to the reverse strand and repeat region start and end points. There is a need to strictly define these and other anchor points to specify the repeat regions.

4. Annotation of STR alleles—nomenclature systems

Established conventions for the nomenclature of forensic CE-based STR genotypes will remain unchanged. Updated and extended nomenclature systems that can be performed by expert systems will be required for STR sequences that can be performed by specifically designed software. It is crucial that this software allow for translation of MPS-derived genotypes to the CE-based nomenclature convention to stay compatible with established STR databases and future CE-based STR results. We note that it is too early to set strict guidelines for new nomenclature formats for MPS. The following exemplar systems are presented here to explore different ways to call MPS-based STR results and can serve as the basis for further discussion and development.

4.1. Comprehensive (high level) STR nomenclature systems

Comprehensive STR nomenclature systems capture the majority, preferably all, of the information present in the STR sequence string and can be delineated from the recommendations of the human genome variation society (http://www.hgvs.org). A comprehensive format includes the STR locus information, the size-based allele category, which provides backward compatibility to existing STR databases, and an unambiguous description of the

sequence variation of each allele. An example of a minimum nomenclature format that could be used in the case of the D13S317 locus is shown in Textbox 1. When a particular genome assembly is used as the reference for the sequence alignment, the assembly version should be stated. Information must be also compiled on the chromosome number and coordinates relating to the whole STR amplicon to compare alleles generated with different primer pairs and the repeat region to differentiate identical repeat and flanking sequence motifs, from which the allele designation was made. Finally, the repeat motif should be fully described with the relevant nucleotide 'blocks' and repeat numbers in brackets as well as SNPs and/or InDels described by genome coordinates or rs-numbers. Common SNP and InDel variants, including those in repeat regions, typically have been identified already and have rs-numbers. Novel variants not yet catalogued tend to keep their chromosome coordinates as identifiers until an rs-number is assigned. This process of rs-number assignment is becoming an increasingly difficult process to complete as a large proportion of SNP variation is unique to an individual [34].

Comprehensive STR nomenclature systems are informative and can be translated to lower level nomenclature systems at any time to maintain backward compatibility with existing databases. However, they cannot easily be applied for communication among forensic analysts and stakeholders as is currently practiced with simple repeat number notation. To facilitate communication and maintain backwards compatibility, any nomenclature system will need to take into account the number of repeats presented in the human reference sequence.

4.2. Simple (low level) STR nomenclature systems

Low-level STR nomenclature systems are based on the translation of sequence strings or comprehensive STR nomenclature systems and typically represent easy-to-read unique identifiers. They typically consist of the STR locus name and the operationally-defined repeat-based allele designation derived from CE. This approach makes the data directly compatible with those of existing STR databases. In order to capture the additional sequence information, accompanying letters have been proposed or numbers and letters in alternating order could be applied, a system that is currently used to display the phylogenetic relationship between linearly inherited markers [37,38]. Simple STR nomenclature systems are easy to communicate and therefore

Textbox 1. An example of a possible sequence nomenclature regime using the example STR D13S317 allele 12 ([CE12]) compared to the reference allele 11 (Ref [11]). Sequence descriptions include the following bolded components: (1) the reference genome assembly sequence (includes allele 11); (2) locus name and CE allele number; (3) chromosome number and reference genome assembly used; (4) repeat region coordinates of the reference allele (start-end nucleotide positions, but eventually to also include the reported region start-end coordinates); (5) description of the repeat motifs; and (6) location of flanking region variants. See D13S317 in Supplementary file S1A for more details of the reference sequence.

```
D13S317 Ref (11)  TCTAACGCCT ATCTGTATTT ACAAATACAT TATC TATC TATC TATC
D13S317 [CE12]    ......A... .......... .......... .... .... .... ....

D13S317 Ref (11)  TATC TATC TATC TATC TATC TATC TATC ++++ AATCAATCAT
D13S317 [CE12]    .... .... .... .... .... .... .... TATC T.........

D13S317 Ref (11)  CTATCTATCT TTCTGTCTGT
D13S317 [CE12]    .......... ..........
```

G Known polymorphic sites

++++ Additional nucleotides compared to reference sequence

1. Bold segment = the reference genome assembly sequence description
D13S317 Ref (11) -Chr13-GRCh38 82148025-82148068 [TATC]$_{11}$
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 [TATC]$_{12}$ 82148001-A; 82148069-T

2. Locus name and capillary electrophoresis allele name
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 [TATC]$_{12}$ 82148001-A; 82148069-T

3. Chromosome and human genome assembly version
D13S317[CE12]-**Chr13-GRCh38** 82148025-82148068 [TATC]$_{12}$ 82148001-A; 82148069-T

4. STR repeat region co-ordinates (start-end) for reference allele
D13S317[CE12]-Chr13-GRCh38 **82148025-82148068** [TATC]$_{12}$ 82148001-A; 82148069-T

5. Description of STR motifs
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 **[TATC]$_{12}$** 82148001-A; 82148069-T

6. Location of flanking region variants
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 [TATC]$_{12}$ **82148001-A; 82148069-T**

preferred for routine exchange of STR data between analysts and stakeholders and may be easier to apply to existing software packages that perform various population genetic and statistical analyses. However, the translation process will have to be managed by a centralized nomenclature commission to avoid ambiguous or imprecise allele names being adopted, or assigning different names to identical alleles. It has been suggested that an online system could be used that is curated by a nomenclature commission, which would be responsible for new allele designations upon validation of the observed sequence variation. Criteria for the validation of sequence variation and its comparison with existing variants need to be defined in more detail. Numerous new variants will be discovered; hence, it is necessary to automate the process as much as possible. If a 'natural' nomenclature is adopted, then cataloguing of variants can be accommodated by an open source algorithm, which should be a key aim of the community.

Fig. 1 illustrates examples of potential difficulties that can arise from the more detailed characterization of STR sequences that MPS provides. There can be unforeseen challenges when aligning the sequence generated by MPS to the established repeat motif description of any STR. Each of the three STRs is described by its respective human reference sequences, which include the repeat regions plus the short segments of the flanking regions.

The D18S51 reference sequence comprises 18 AGAA repeat motifs (ten nucleotides of flanking region also displayed). Two repeat region InDels create intermediate repeats: x.3 (rs572637907); x.2 (rs575219471); or x.1 (presence of both deletions or another unmapped deletion). Furthermore, the flanking A/G SNP rs535823682 potentially complicates the alignment of the repeat sequence.

The D13S317 reference sequence comprises 11 TATC repeat motifs (extended flanking regions displayed). The two 3′ flanking region A/T SNPs, rs9546005 and rs202043589, create TATC tetra-nucleotides matching the repeat motifs, but these are not counted when deriving the total repeat number. The rs561167308 TCTG deletion potentially creates a four-nucleotide fragment size disparity with CE-based allele descriptions depending on the position of the 3′ primer-binding site. The 5′ SNP rs146621667 is the site of the '82148001-A' variant described in Textbox 1.

The D19S433 reference sequence comprises 14CCTT repeat motifs, which contain two 'punctuated' stable repeat motifs, CCTA and CTTT, that should be counted, but in the initial development of forensic CE kits for D19S433 were not. The D19S433 STRbase (http://www.cstl.nist.gov/strbase/) fact sheet therefore provides a cautionary note to highlight that current allelic ladders retain the numbering system first used that did not count the above two non-standard motifs in combination with the CCTT motifs. The 16 nucleotide 5′ flanking sequence also shows permutations on the CCTT motif that have no sequence variants but can present alignment challenges for analysis of MPS sequence data.

The above examples illustrate that when characterization of repeat regions does not follow previously agreed nomenclature rules [7] it potentially creates discrepancies between CE-based repeat counts and MPS sequence analyses made from the same amplified fragments. In this case, a nomenclature commission can preempt potential issues by harmonizing CE numbering systems and repeat region sequence descriptions. However, since STR types based on CE already populate national DNA databases, the existing nomenclature rules must be applied to MPS sequence data to prevent data mismatches, even though they may not follow common logic.

Consideration 5. Although simple STR nomenclature systems may be required at some point in the future to facilitate communication and data exchange, comprehensive STR nomenclature systems are preferred for early adopters of STR MPS analysis in order to ensure compatibility with MPS data generated in the future. Backward compatibility to the

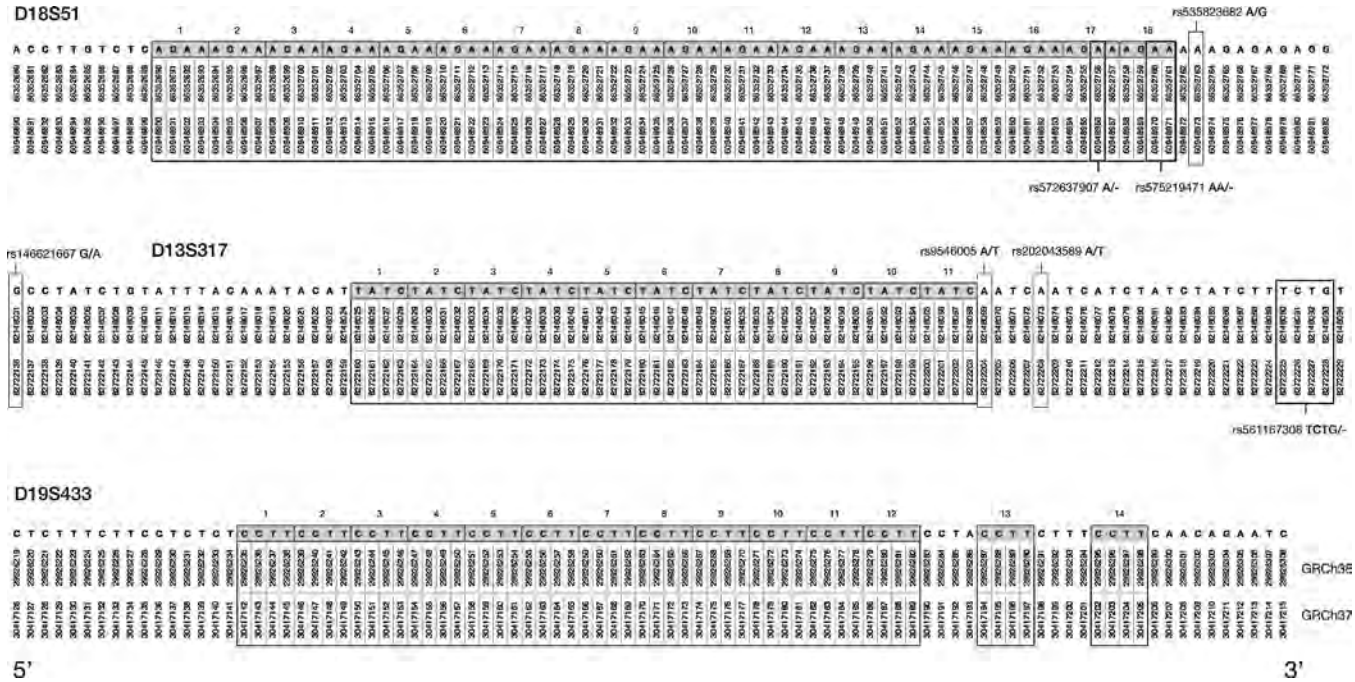


Fig. 1. Three examples of STR repeat regions plus the short segments of their 5′ and 3′ flanking sequences that illustrate potential difficulties with repeat motif description. All sequences are taken from the current human reference genome assembly and coordinates are given for both GRCh37 and GRCh38. Repeat regions are denoted by thin black boxes, InDels by thick black boxes, and SNPs by grey boxes. For a more detailed description of each STR sequence see [17]. D18S51 reference sequence of 18 AGAA repeat motifs and ten nucleotides of flanking region. D13S317 reference sequence of 11 TATC repeat motifs with extended flanking regions. In both STRs InDel polymorphisms and/or SNPs in the 3′ flanking region create intermediate alleles but these sequence changes can mimic repeat motifs not included in the CE-based nomenclature. D19S433 reference sequence of 14CCTT repeat motifs and flanking regions. In this STR not all tandemly-arranged tetra-nucleotide motifs are counted in the description of the repeat region.

repeat-based nomenclature derived from CE needs to be maintained to preserve the universal applicability of established national STR databases

4.3. Flanking regions

The inclusion of flanking region sequence variants (between primer binding sites and the repeat region) in compiled MPS data is important for several reasons. First, it provides additional informative polymorphisms with which to differentiate alleles that have identical repeat region sequences. Second, the mapping of InDel variants informs the assignment of size-based allele designations from CE analyses, where the total fragment size is altered by the presence of the variant. One example is the occurrence of a four-nucleotide deletion (rs561167308) close to the repeat region of the D18S51 locus that changes the repeat length but is not a detected repeat itself [18]. This is also the case with the DXS10148 locus, which has a variable motif of eight bases adjacent to the core tetra-nucleotide repeat region [39]. Third, it is likely that a small but regular proportion of novel rare variants will be discovered in full STR sequence segments that potentially provide additional ways to differentiate STR alleles amongst related individuals, but which have no previously defined frequency data. In these instances, it is important to compare the novel variants with a database of established flanking region variants including sample population sizes to provide allele frequencies. As flanking region variants and repeat region sequence variants are present on one DNA fragment, the database must compile all variation in the sequence string from any one sample. Novel variants can be described by their genome coordinates, while recognized variants that already are catalogued will have rs-numbers. To ensure compatibility between/among different primer sets used for library preparation and sequencing, it is mandatory to provide genome coordinates of the sequence read start and end points similar to current practices with difference-coded variants describing mtDNA haplotypes [28]. This procedure should cover annotation of InDels, as it is possible that some MPS primer sets will be positioned inside those used for CE analysis such that InDel sites may escape detection by sequencing and create discordant fragment sizes. Such checks have been made successfully, e.g. the concordance studies of MiniFiler systems, where modified primer positions did influence the observed repeat numbers [40].

Supplementary file S1 illustrates seven common flanking region SNPs within 50 nucleotides flanking region of the listed autosomal STRs. The SNPs are shown with population frequency data from 1000 Genomes samples and represent the most informative levels of flanking region variation, defined here as having minor allele frequencies of 10% or more in most populations (average heterozygosities of 18% or higher). These SNPs are: rs4847015 in the D1S1656 locus; rs6736691 in the D2S1338 locus; rs25768 in the D5S818 locus; rs16887642 in the D7S820 locus; rs75219269 in the VWA locus; rs9546005 in the D13S317 locus, and rs11642858 in the D16539 locus. However, their detection is dependent on the amplified fragment sizes of each locus (i.e. the position of the primers). For example, certain SNPs within 50 nucleotides of the repeat region will not be genotyped when much shorter STR fragment lengths are generated by MPS primer sets.

Consideration 6. To account for relevant genetic variation outside common repeat regions, STR sequences stored as sequence strings should include flanking sequences as well as the genome coordinates of the sequence read start and end points.

5. Updated allele frequencies

Current allele frequency tables are not sufficient to quantify any new variation gained by sequencing of STRs. Preliminary studies indicate that the number of rare STR alleles will increase substantially with MPS [18,41,42]. Thus, comprehensive MPS databasing will be required to characterize the extent of STR sequence variation for use in STR frequency estimates. Therefore, there is a particular need to promptly harmonize nomenclature frameworks, since a coordinated effort is required to collate the sequence variation found by early adopters, before this process reaches the wider community of forensic laboratories.

From data published so far [18,41,42] and from previous assessments of sequence variation with ICEMS technology [22,23,43] it is certain that many common STRs (e.g. D12S391, D21S11) will require large-scale efforts to compile representative samples of their variation, while other STRs such as FGA appear to have largely unchanged levels of polymorphism. In addition, flanking sequence variation will show a proportion of 'private' variants at $<1\%$ frequencies that have not been previously described [34]. Thus, the community must adopt a nomenclature framework that captures variation within the repeats and a framework for flanking SNPs lacking rs-numbers. Prompt standardization of nomenclature will facilitate the development of large-scale sequence databases and expedite the collection of rare variant allele frequencies, much of which may be population-specific.

Consideration 7. Updated allele frequency databases will be necessary to take full advantage of the increased power of discrimination offered by MPS generated STR data. A unified nomenclature system is needed to ensure compatibility of worldwide population databases.

6. Selection of STR loci

While the choice of the first forensic STR loci was previously driven by individual research groups (e.g. [44]) and later commercially produced (e.g. [45]), the addition of new forensically-relevant STR loci was led by world-wide forensic societies and working groups (e.g. [5,6,10]). This emphasis on localized needs was important for laboratories to meet legal requirements defined in their respective countries, with particular regard to database search strategies. It is desirable to continue dialogues between forensic groups and commercial suppliers to ensure provision of appropriate loci, chemistry, and software.

The variation of new STR loci should be tested with studies of populations from the main continental groups with particular emphasis on discrimination power, heterozygosity levels, sequence variation in the flanking regions, and inter- and intra-population variation. Given the complexities of STR sequence alignments and the current limitation of MPS read length, SE33 [46] is unlikely to be part of initial forensic MPS multiplexes. In its place many miniSTRs, newer to mainstream use, could be suitable alternatives and are certain to be incorporated into future MPS marker sets [36]. These STRs will require full characterization, including crucial information about possible linkage to the already well established STR markers [47]. so that frequency data and knowledge of sequence characteristics can be added to the extensive data in place for the commonly used loci.

At present, the key factors that must be considered in the application of sequencing technologies to STRs center on standardized representation of sequence variation. Until an appropriate, agreed upon framework for simplified STR nomenclature is established, STR sequence data should reflect the most detailed and inclusive level of information for any given allele, while still retaining compatibility with current CE-defined

variants. The likely near-term development of reference population data should serve to test the utility and robustness of the considerations presented here, and also provides the necessary data framework for refinement and establishment of a practical and durable simplified nomenclature scheme.

At a future point in time when MPS-based databases have grown in size, algorithms could be used to determine frequency databases without the need to annotate alleles. A strength-of-evidence calculation would follow without any reference to nomenclature. However, this approach would require a broad application of MPS-based STR typing by the forensic community.

Consideration 8. Future forensic MPS multiplexes would benefit from retention of past markers for backward compatibility and a marker selection process based on population data, molecular biology, sequencing chemistry, and a continued dialogue between the forensic community and commercial suppliers.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Concepts and ideas stated in this manuscript were stimulated by a panel discussion on STR nomenclature at the 26th Conference of the International Society for Forensic Genetics, 2 September 2015, Krakow, Poland. The authors are indebted to all those, who contributed to the discussions both during and after the panel's review of STR sequence nomenclature issues. The authors would like to thank Bettina Zimmermann (Innsbruck, Austria) for technical help and Chris Tyler Smith, Sanger Institute (Hinxton, UK) for valuable guidance on current 1000 Genomes data policies. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the National Institute of Standards and Technology (US). 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 endorsement by the National Institute of Standards and Technology, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. This is FBI Laboratory publication number 16-11.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.01.009.

References

[1] C. Puers, H.A. Hammond, L. Jin, C.T. Caskey, J.W. Schumm, Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder, Am. J. Hum. Genet. 53 (1993) 953–958.
[2] P. Gill, C. Kimpton, E. D'Aloja, J.F. Andersen, W. Bar, B. Brinkmann, et al., Report of the European DNA profiling group (EDNAP)—towards standardisation of short tandem repeat (STR) loci, Forensic Sci. Int. 65 (1994) 51–59.
[3] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci. Int. 89 (1997) 185–197.
[4] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S., Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians, J. Forensic Sci. 44 (1999) 1277–1286.
[5] P. Gill, E. d'Aloja, B. Dupuy, B. Eriksen, M. Jangblad, V. Johnsson, et al., Report of the European DNA profiling group (EDNAP)—an investigation of the hypervariable STR loci ACTBP2, APOAI1 and D11S554 and the compound loci D12S391 and D1S1656, Forensic Sci. Int. 98 (1998) 193–200.
[6] L.A. Welch, P. Gill, C. Phillips, R. Ansell, N. Morling, W. Parson, et al., European Network of Forensic Science Institutes (ENFSI): evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers, Forensic Sci. Int. Genet. 6 (2012) 819–826.
[7] W. Bär, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, et al., DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics, Int. J. Legal Med. 110 (1997) 175–176.
[8] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265.
[9] P.M. Schneider, Scientific standards for studies in forensic genetics, Forensic Sci. Int. 165 (2007) 238–243.
[10] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States, Forensic Sci. Int. Genet. 17 (2015) 33–34.
[11] P.M. Schneider, Expansion of the European Standard Set of DNA database loci—the current situation, Profiles DNA (2009) 6–7.
[12] C. Borsting, N. Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (2015) 78–89.
[13] W. Parson, G. Huber, L. Moreno, M.B. Madel, M.D. Brandhagen, S. Nagl, et al., Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples, Forensic Sci. Int. Genet. 15 (2015) 8–15.
[14] M. Eduardoff, C. Santos, M. de la Puente, T.E. Gross, M. Fondevila, C. Strobl, et al., Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM, Forensic Sci. Int. Genet. 17 (2015) 110–121.
[15] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, BioTechniques 51 (2011) 127–133.
[16] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417.
[17] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.
[18] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.
[19] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int. Genet. 6 (2012) 810–818.
[20] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers, Forensic Sci. Int. Genet. 12 (2014) 107–119.
[21] J.M. Butler, J. Li, T.A. Shaler, J.A. Monforte, C.H. Becker, Reliable genotyping of short tandem repeat loci without an allelic ladder using time-of-flight mass spectrometry, Int. J. Legal Med. 112 (1999) 45–49.
[22] F. Pitterl, H. Niederstatter, G. Huber, B. Zimmermann, H. Oberacher, W. Parson, The next generation of DNA profiling–STR typing by multiplexed PCR–ion-pair RP LC-ESI time-of-flight MS, Electrophoresis 29 (2008) 4739–4750.
[23] F. Pitterl, K. Schmidt, G. Huber, B. Zimmermann, R. Delport, S. Amory, et al., Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations, Int. J. Legal Med. 124 (2010) 551–558.
[24] K. van der Gaag, P. de Knijff, Forensic nomenclature for short tandem repeats updated for sequencing, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e542–e544.
[25] A. Röck, J. Irwin, A. Dur, T. Parsons, W. Parson, SAM: string-based sequence search algorithm for mitochondrial DNA database queries, Forensic Sci. Int. Genet. 5 (2011) 126–132.
[26] W. Parson, A. Dür, EMPOP–a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.
[27] A. Carracedo, W. Bär, P. Lincoln, W. Mayr, N. Morling, B. Olaisen, et al., DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing, Forensic Sci. Int. 110 (2000) 79–85.
[28] W. Parson, L. Gusmao, D.R. Hares, J.A. Irwin, W.R. Mayr, N. Morling, et al., DNA Commission of the International Society for forensic genetics: revised and extended guidelines for mitochondrial DNA typing, Forensic Sci. Int. Genet. 13 (2014) 134–142.
[29] S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, et al., Sequence and organization of the human mitochondrial genome, Nature 290 (1981) 457–465.
[30] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[31] D.M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.L. Loogvali, N.M. Silva, et al., A Copernican reassessment of the human mitochondrial DNA tree from its root, Am. J. Hum. Genet. 90 (2012) 675–684.
[32] A. Salas, M. Coble, S. Desmyter, T. Grzybowski, L. Gusmao, C. Hohoff, et al., A cautionary note on switching mitochondrial DNA reference sequences in forensic genetics, Forensic Sci. Int. Genet. 6 (2012) e182–e184.
[33] H.J. Bandelt, A. Kloss-Brandstätter, M.B. Richards, Y.G. Yao, I. Logan, The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and

the standardization of notation in human mitochondrial DNA studies, J. Hum. Genet. 59 (2014) 66–77.
[34] C. Genomes Project, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.
[35] ForenSeq DNA Signature Prep Reference Guide, Document No. 15049528 v.01, (2015), Illumina.
[36] C.R. Hill, J.M. Butler, P.M. Vallone, A 26plex autosomal STR assay to aid human identity testing, J. Forensic Sci. 54 (2009) 1008–1015.
[37] S.K. Lim, Y. Xue, E.J. Parkin, C. Tyler-Smith, Variation of 52 new Y-STR loci in the Y chromosome consortium worldwide panel of 76 diverse individuals, Int. J. Legal Med. 121 (2007) 124–127.
[38] A. Torroni, A. Achilli, V. Macaulay, M. Richards, H.J. Bandelt, Harvesting the fruit of the human mtDNA tree, Trends Genet. 22 (2006) 339–345.
[39] I. Gomes, A. Brehm, L. Gusmao, P.M. Schneider, New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci, Forensic Sci. Int. Genet. 20 (2016) 112–116.
[40] C.R. Hill, M.C. Kline, J.J. Mulero, R.E. Lagace, C.W. Chang, L.K. Hennessy, et al., Concordance study between the AmpFlSTR MiniFiler PCR amplification kit and conventional STR typing kits, J. Forensic Sci. 52 (2007) 870–873.
[41] E. Rockenbauer, S. Hansen, M. Mikkelsen, C. Borsting, N. Morling, Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing, Forensic Sci. Int. Genet. 8 (2014) 68–72.
[42] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.
[43] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, et al., Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, Forensic Sci. Int. Genet. 6 (2012) 594–606.
[44] J.E. Lygo, P.E. Johnson, D.J. Holdaway, S. Woodroffe, J.P. Whitaker, T.M. Clayton, et al., The validation of short tandem repeat (STR) loci for use in forensic casework, Int. J. Legal Med. 107 (1994) 77–89.
[45] B. Budowle, C.J. Sprecher, Concordance study on population database samples using the PowerPlex 16 kit and AmpFlSTR Profiler Plus kit and AmpFlSTR COfiler kit, J. Forensic Sci. 46 (2001) 637–641.
[46] D. Warne, C. Watkins, P. Bodfish, K. Nyberg, N.K. Spurr, Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene 2 (ACTBP2) detected using the polymerase chain reaction, Nucleic Acids Res. 19 (1991) 6980.
[47] C. Phillips, D. Ballard, P. Gill, D.S. Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.

Forensic Science International: Genetics 31 (2017) 111–117

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci

Katherine Butler Gettings[a,*], Lisa A. Borsuk[a], David Ballard[b], Martin Bodner[c], Bruce Budowle[d,e], Laurence Devesse[b], Jonathan King[d], Walther Parson[c,f], Christopher Phillips[g], Peter M. Vallone[a]

[a] U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[b] King's Forensics, King's College London, Franklin-Wilkins Building, 150 Stamford Street London, UK
[c] Institute of Legal Medicine, Medical University of Innsbruck, Austria
[d] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University Jeddah, Saudi Arabia
[f] Forensic Science Program, The Pennsylvania State University, USA
[g] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
Forensic STR
DNA sequencing
NGS
MPS
Nomenclature

ABSTRACT

The STR Sequencing Project (STRSeq) was initiated to facilitate the description of sequence-based alleles at the Short Tandem Repeat (STR) loci targeted in human identification assays. This international collaborative effort, which has been endorsed by the ISFG DNA Commission, provides a framework for communication among laboratories. The initial data used to populate the project are the aggregate alleles observed in targeted sequencing studies across four laboratories: National Institute of Standards and Technology (N = 1786), Kings College London (N = 1043), University of North Texas Health Sciences Center (N = 839), and University of Santiago de Compostela (N = 944), for a total of 4612 individuals. STRSeq data are maintained as GenBank records at the U.S. National Center for Biotechnology Information (NCBI), which participates in a daily data exchange with the DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA). Each GenBank record contains the observed sequence of a STR region, annotation ("bracketing") of the repeat region and flanking region polymorphisms, information regarding the sequencing assay and data quality, and backward compatible length-based allele designation. STRSeq GenBank records are organized within a BioProject at NCBI (https://www.ncbi.nlm.nih.gov/bioproject/380127), which is sub-divided into: commonly used autosomal STRs, alternate autosomal STRs, Y-chromosomal STRs, and X-chromosomal STRs. Each of these categories is further divided into locus-specific BioProjects. The BioProject hierarchy facilitates access to the GenBank records by browsing, BLAST searching, or ftp download. Future plans include user interface tools at strseq.nist.gov, a pathway for submission of additional allele records by laboratories performing population sample sequencing and interaction with the STRidER web portal for quality control (http://strider.online).

1. Introduction

As the forensic DNA community evaluates the potential of sequencing applications for Short Tandem Repeat (STR) loci, it is imperative to define the allelic diversity in these regions of the human genome. Large-scale sequencing projects within the broader genomics community may use shorter read chemistries (e.g. 100 bp) and may not describe repetitive regions due to their complexity and non-conformity to typical alignment parameters [1]. Additionally, knowledge of the forensic literature is needed to report STR sequences in the same manner established by the forensic community.

Even within forensic sequencing studies, there are differences in the reporting of sequence-based STR alleles. Names of convenience such as **20(a)** [2] or **FL1X20** [3] have not been standardized and may create confusion about the specific allele being reported. There may be differences in format for the compression or "bracketing" of STR sequences, such as **ATAG [9]** [4,5] or **[ATAG]$_9$** [6] or **[ATAG] 9** [7]. More importantly, there may be differences in strand reporting where choice of the forward strand will match the reference sequence direction, and choice of the reverse strand aligns the sequence in the opposite direction. The DNA Commission of the ISFG on minimal nomenclature requirements in 2016 recommended reporting all sequences

* Corresponding author at: National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA.
E-mail addresses: katherine.gettings@nist.gov (K.B. Gettings), lisa.borsuk@nist.gov (L.A. Borsuk), David.ballard@kcl.ac.uk (D. Ballard), peter.vallone@nist.gov (P.M. Vallone).

http://dx.doi.org/10.1016/j.fsigen.2017.08.017
Received 28 July 2017; Accepted 30 August 2017
Available online 01 September 2017

in the forward strand orientation [8]. However, some loci were historically reported on the reverse strand [9]. In particular, STRs for which the reported strand has changed over time may differ in reporting where the repeat region begins. This can result in shifted (different) allele number designations for the same sequence [8]. Lastly, the recovery and reporting of varying lengths of flanking regions (and hence flanking region variants) is inherent to differences in kit designs and bioinformatic pipelines.

The international forensic DNA community continues to develop guidance on STR sequence nomenclature, and additional resources for quality control of STR sequence data are being developed [10]. However, the need for standardization is immediate. A 2016 survey was recently published by the European Network of Forensic Science Institutes (ENFSI) DNA Working Group [11], in which over half of the 33 responding laboratories have already purchased at least one sequencing instrument. The respondents (primarily composed of government forensic laboratories across 25 countries) reported *lack of nomenclature and reporting standards* as the highest ranking scientific and legal challenge for the implementation of new sequencing technologies in forensic genetics. Also in 2016, the Applied Genetics Group of the U.S. National Institute of Standards and Technology (NIST) queried forensic laboratories to assess the utility of STR reference sequences for loci of forensic interest. The feedback received from 22 laboratories (representing 11 countries) mirrored the ENSFI survey with strong support for the development of STR sequence nomenclature resources.

In response to this need, NIST partnered with the U.S. National Center for Biotechnology Information (NCBI), leveraging NIST's over 20-year history supporting the forensic STR typing community [12] and NCBI's extensive infrastructure for accepting, maintaining and serving DNA sequence data. Through this partnership, the STR Sequencing Project (STRSeq) has been initiated to facilitate the description of sequence-based alleles at the STRs targeted in human identification assays. This resource consists of a curated catalog of sequence diversity at forensic STR loci, along with the key elements of nomenclature conforming to current guidelines [8], and will serve as the data backbone during this time of transition, as well as a stable resource for the future.

2. Samples and submission strategy

The initial data used to populate STRSeq are the aggregate alleles observed in targeted sequencing studies of single source samples across four laboratories: NIST, Kings College London (KCL), University of North Texas Health Science Center (UNT), and University of Santiago de Compostela (USC), for a total of 4612 individuals. The number of alleles aggregated differs by locus due to variable multiplex performance and quality requirements described in Section 3. As only aggregate alleles are displayed, the source of the alleles is anonymized. The targeted sequence data used in STRSeq either have been, or are expected to be published by the submitting laboratory ([6,13], additional manuscripts in preparation). Records will be added to the STRSeq BioProject in sets, largely coinciding with associated publications, as follows:

NIST: N = 1786 samples from multiple sources: 1) N = 665 liquid blood samples purchased from Interstate Blood Bank (Memphis, TN) and Millennium Biotech, Inc. (Ft. Lauderdale, FL) with self-declared ancestries from three U.S. population groups: Caucasian, African American, and Hispanic; 2) N = 781 buccal swabs provided by DNA Diagnostics Center (Fairfield, OH) from paternity testing samples with self-declared ancestries from four U.S. population groups: Caucasian, African American, Asian and Hispanic; 3) N = 297 buccal swabs collected from anonymous volunteers of self-reported, diverse ancestries, provided by the George Washington University; and 4) N = 43 control samples and reference materials. All samples have been sequenced with the ForenSeq system (Illumina) and a subset (> 600 samples) has overlapping sequence data from the PowerSeq Auto-Y assay (Promega). In addition, for the majority of these samples, capillary electrophoresis (CE) STR data is available at all ForenSeq and PowerSeq Auto-Y loci ([14,15] and unpublished data).

KCL: N = 1043 samples were obtained from consenting adult volunteers resident in the U.K. The samples relate to six U.K. population groups with self-declared ancestries of: White British, West African, North East African, South Asian, Chinese and Middle Eastern. All samples have been sequenced with the ForenSeq system and additionally genotyped with at least two commonly available CE kits.

UNT: N = 839 samples which have been described in associated sequence-based allele frequency publications and were sequenced with the ForenSeq system [6,13].

USC: N = 944 samples from the HGDP-CEPH diversity panel cell-line DNAs from 51 diverse populations were sequenced with the ForenSeq system.

Initially, STRSeq records will be created for the STR loci targeted in the aforementioned assays; additional records will be created as samples are sequenced with other available commercial assays, e.g. Precision ID GlobalFiler NGS STR Panel (Thermo Fisher Scientific). If new STR loci (see [16]) are targeted in commercially available assays launched in the future, additional records will be created.

A single laboratory will be indicated as having submitted each record. The association of a *submitting laboratory* with a record does not imply "discovery" of a sequence variant; rather the designation is simply the organization that initially provided the sequence and maintains the supporting data. For the initial data set, NIST will be the *submitting laboratory* of all sequences generated at NIST and the other laboratories will be the *submitting laboratory* of those sequences generated at that specific laboratory for which records do not already exist in the database. Duplicate records will not be created, which will generally result in a decreasing number of new sequence records as successive sample sets are added. Fig. 1 outlines an example submission strategy of non-duplicate allele records that might be expected from a typical highly polymorphic STR such as D12S391.

3. BioProject hierarchy and record format

The BioProject hierarchy serves to organize the GenBank records (Table 1). The highest-level STRSeq umbrella project contains four sub-umbrella projects: (a) **Commonly Used Autosomal STR Loci**, (b) **Alternate Autosomal STR Loci**, (c) **Y-Chromosomal STR Loci**, and (d) **X-Chromosomal STR Loci**. These sub-umbrella projects are divided further into locus-specific data-level projects which contain the GenBank sequence record data. Each umbrella and data-level project has a corresponding accession number, e.g. PRJNA380127 is the STRSeq umbrella project, PRJNA380345 is the **Commonly Used Autosomal STR Loci** sub-umbrella project, and PRJNA380554 is the **TPOX Sequence-Based Alleles** project (the common PRJNA prefix identifies the six-digit number as a BioProject). Entering one of these accession numbers at https://www.ncbi.nlm.nih.gov/bioproject allows direct access to the umbrella or data-level project of interest. Each BioProject page contains additional links for up, down, and cross navigation. Table 1 contains direct links to STRSeq umbrella and data-level projects.

The sequence records in GenBank are flat files of specified format that can be downloaded and parsed en masse (see Fig. 2 for an example record for the TPOX locus). Starting from the bottom of the record, in a section labeled **ORIGIN**, users will find the full sequence that was reported by the submitting laboratory. The length of reported sequence is dependent upon the assay and the quality of the flanking sequence data, but generally will be consistent with the assay-specific configuration files published in [17]. Above the sequence is the **FEATURES** table, which includes the position of the repeat region within the sequence, the position and dbSNP rs number of variations in the flanking regions (when applicable), and the subset of sequence that was observed with different commercial assays (when applicable). Each feature can be selected in order to highlight the appropriate region in the sequence

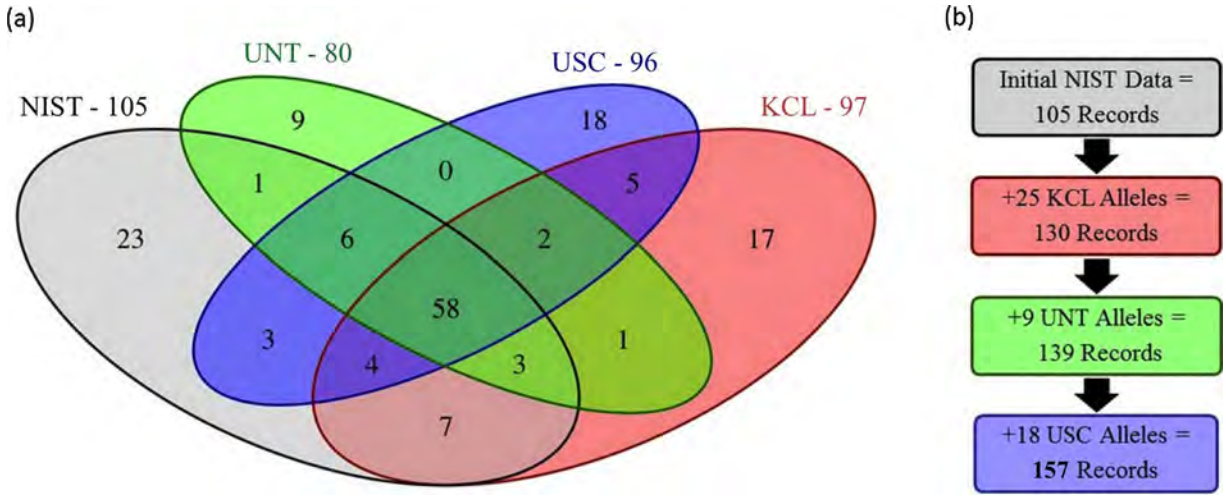


Fig. 1. (a) Venn diagram demonstrating the overlap of D12S391 sequence-based alleles observed among the four laboratories, and the total number of unique sequence-based alleles observed within each laboratory. (b) Submission strategy for 157 unique sequence-based alleles observed at the D12S391 locus. The 105 unique alleles generated at NIST form the basis of STRSeq records. Subsequent submissions from KCL, UNT, and USC will add records for sequences generated at each laboratory for which records do not already exist (25, 9, and 18 records, respectively).

string. SNP rs numbers are hyperlinked to dbSNP, allowing users to navigate and access frequency information quickly. If the polymorphism has not been assigned a dbSNP reference number, the GRCh38 coordinate is given, and the field will be updated if an rs number is assigned later or if the assembly is updated.

Above the **FEATURES** table is the *structured comments* section (offset with ##humanSTR-START## and ##humanSTR-END##), which contains field-based information relevant to STRSeq records. The given **Bracketed repeat** is intended to be consistent with the guidance of the ISFG nomenclature commission [8]. Specific to STRSeq records is the lower-case formatting of selected bases within the **Bracketed repeat**, which highlights sequence tracts that are not counted toward the length-based allele designation (when applicable, e.g. D19S433 14 allele will be presented as: [AAGG] aaag [AAGG] tagg [AAGG]12). The **Sequencing technology** field lists the commercial assay(s) and instrument(s) used to generate the sequence data. The **Coverage** field lists the minimum threshold of reads observed for the reported sequence. The current threshold for STRSeq record creation is > 30X. This is consistent with the default minimum "interpretation threshold" implemented in one commercial software, corresponding to the only relevant commercial assay with a published developmental validation [18] at the time of writing. This threshold will continue to be evaluated in the future as additional developmental validations are published. The **Length-based tech.** field lists the assay and instrument used to generate the **Length-based allele** given. Often a sequence will have been observed in multiple samples. The length-based information in each record indicates that, for at least one sample, the specified length-based allele was generated with the given length-based technology. This approach is not meant to be comprehensive; variation in the length-based allele among individuals or assays can result from indels in flanking regions. In some instances, length-based allele confirmation may not be possible, such as the lack of a CE assay for STRs targeted by commercial sequencing assays but not previously in common use. When a length-based allele confirmation has not been performed, the **Length-based allele** field will indicate e.g. "7 (Inferred from sequence)" and the **Length-based tech.** field will contain "Not reported". The remaining information in the *structured comments* section orients the sequence on the chromosome and will be updated along with the reference sequence assembly.

Above the *structured comments* section is the **COMMENT** block, which is identical across records and recapitulates this paper. Above the **COMMENT** block are references. **REFERENCE 1** will be this paper and **REFERENCE 2** identifies the submitting laboratory. The remaining topmost fields contain information for GenBank record organization. The **ACCESSION** and **VERSION** number is the GenBank sequence identifier (e.g. MF044256.1 in Fig. 2). If future commercial assay typing provides additional flanking sequence, the updated sequence will become e.g. MF044256.2 (coexisting with MF044256.1). If the additional flanking sequence reveals a polymorphism, the additional sequence consistent with the reference sequence becomes e.g. MF044256.2 and a new record is created for the additional sequence which differs from the reference sequence.

The **DEFINITION** line near the top of the record is the descriptor present in a list of sequences (see https://www.ncbi.nlm.nih.gov/nuccore/?term=strseq+tpox), and will uniquely identify each allele with components of the record itself. In addition, the top of each record contains hyperlinks to the **FASTA** sequence, which can be downloaded, and a **Graphics** view (Fig. 3). This graphical display presents an interactive version of the sequence (displaying forward and reverse strands) and the features identified in the GenBank record: the repeat region, the region(s) reported from each available sequencing technology, and any associated flanking region polymorphisms. The information shown in **Graphics** view is dependent on the **Tracks** selected in the viewer. All available information for the record is displayed simultaneously by selecting both the **Sequence** and **Aggregate features Track**. More information and tutorials on the NCBI Sequence Viewer can be found at https://www.ncbi.nlm.nih.gov/tools/sviewer.

4. Typical use cases

Several use cases for STRSeq have been identified based on feedback from the forensic community:

I. As a teaching tool to explore STR sequences. The STRSeq BioProject is expected to be useful to forensic operational, academic, and commercial laboratories interested in sequencing STRs as it allows the viewing and downloading of repeat region motifs, flanking region polymorphisms, and commercial assay overlap.
II. As the data backbone for software development. This catalog of sequences with associated forensic formatting and stable links to GenBank records facilitates development of STR sequencing methods and bioinformatic pipelines that conform to agreed variant data frameworks.
III. To provide a quality control function for the evaluation of rare sequences. When a sequence is observed in forensic casework that was not observed in initial validation studies or in the implemented

Table 1
STRSeq BioProject hierarchy, accession numbers, and direct links to all levels. The highest-level of organization is the STRSeq umbrella project (PRJNA380127, ncbi.nlm.nih.gov/bioproject/380127), containing four sub-umbrella projects: (a) Commonly Used Autosomal STR Loci, (b) Alternate Autosomal STR Loci, (c), Y-Chromosomal STR Loci and (d) X-Chromosomal STR Loci. Each of these contains locus-specific sub-projects, which are the data-level projects containing GenBank sequence records.

a

Commonly Used Autosomal STR Loci – PRJNA380345

ncbi.nlm.nih.gov/bioproject/380345

D1S1656	PRJNA380553	ncbi.nlm.nih.gov/bioproject/380553
TPOX	PRJNA380554	ncbi.nlm.nih.gov/bioproject/380554
D2S441	PRJNA380555	ncbi.nlm.nih.gov/bioproject/380555
D2S1338	PRJNA380556	ncbi.nlm.nih.gov/bioproject/380556
D3S1358	PRJNA380558	ncbi.nlm.nih.gov/bioproject/380558
FGA	PRJNA380559	ncbi.nlm.nih.gov/bioproject/380559
D5S818	PRJNA380560	ncbi.nlm.nih.gov/bioproject/380560
CSF1PO	PRJNA380561	ncbi.nlm.nih.gov/bioproject/380561
SE33	PRJNA380562	ncbi.nlm.nih.gov/bioproject/380562
D6S1043	PRJNA380563	ncbi.nlm.nih.gov/bioproject/380563
D7S820	PRJNA380564	ncbi.nlm.nih.gov/bioproject/380564
D8S1179	PRJNA380565	ncbi.nlm.nih.gov/bioproject/380565
D10S1248	PRJNA380566	ncbi.nlm.nih.gov/bioproject/380566
TH01	PRJNA380567	ncbi.nlm.nih.gov/bioproject/380567
vWA	PRJNA380568	ncbi.nlm.nih.gov/bioproject/380568
D12S391	PRJNA380569	ncbi.nlm.nih.gov/bioproject/380569
D13S317	PRJNA380570	ncbi.nlm.nih.gov/bioproject/380570
Penta E	PRJNA380571	ncbi.nlm.nih.gov/bioproject/380571
D16S539	PRJNA380572	ncbi.nlm.nih.gov/bioproject/380572
D18S51	PRJNA380573	ncbi.nlm.nih.gov/bioproject/380573
D19S433	PRJNA380574	ncbi.nlm.nih.gov/bioproject/380574
D21S11	PRJNA380575	ncbi.nlm.nih.gov/bioproject/380575
Penta D	PRJNA380576	ncbi.nlm.nih.gov/bioproject/380576
D22S1045	PRJNA380577	ncbi.nlm.nih.gov/bioproject/380577

b

Alternate Autosomal STR Loci – PRJNA380346

ncbi.nlm.nih.gov/bioproject/380346

D1S1677	PRJNA396107	ncbi.nlm.nih.gov/bioproject/396107
D2S1776	PRJNA396108	ncbi.nlm.nih.gov/bioproject/396108
D3S4529	PRJNA396109	ncbi.nlm.nih.gov/bioproject/396109
D4S2408	PRJNA396110	ncbi.nlm.nih.gov/bioproject/396110
D5S2800	PRJNA396111	ncbi.nlm.nih.gov/bioproject/396111
D6S474	PRJNA396112	ncbi.nlm.nih.gov/bioproject/396112
D9S1122	PRJNA396113	ncbi.nlm.nih.gov/bioproject/396113
D12ATA63	PRJNA396114	ncbi.nlm.nih.gov/bioproject/396114
D14S1434	PRJNA396115	ncbi.nlm.nih.gov/bioproject/396115
D17S1301	PRJNA396116	ncbi.nlm.nih.gov/bioproject/396116
D20S482	PRJNA396117	ncbi.nlm.nih.gov/bioproject/396117

c

Y-Chromosomal STR Loci – PRJNA380347

ncbi.nlm.nih.gov/bioproject/380347

DYF387S1	PRJNA396118	ncbi.nlm.nih.gov/bioproject/396118
DYS19	PRJNA396119	ncbi.nlm.nih.gov/bioproject/396119
DYS385 a/b	PRJNA396120	ncbi.nlm.nih.gov/bioproject/396120
DYS389 I/II	PRJNA396122	ncbi.nlm.nih.gov/bioproject/396122
DYS390	PRJNA396123	ncbi.nlm.nih.gov/bioproject/396123
DYS391	PRJNA396124	ncbi.nlm.nih.gov/bioproject/396124
DYS392	PRJNA396125	ncbi.nlm.nih.gov/bioproject/396125
DYS393	PRJNA396126	ncbi.nlm.nih.gov/bioproject/396126
DYS437	PRJNA396127	ncbi.nlm.nih.gov/bioproject/396127
DYS438	PRJNA396128	ncbi.nlm.nih.gov/bioproject/396128
DYS439	PRJNA396129	ncbi.nlm.nih.gov/bioproject/396129
DYS448	PRJNA396130	ncbi.nlm.nih.gov/bioproject/396130
DYS456	PRJNA396131	ncbi.nlm.nih.gov/bioproject/396131
DYS458	PRJNA396132	ncbi.nlm.nih.gov/bioproject/396132
DYS460	PRJNA396134	ncbi.nlm.nih.gov/bioproject/396134

Table 1 *(continued)*

c

Y-Chromosomal STR Loci – PRJNA380347

ncbi.nlm.nih.gov/bioproject/380347

DYS481	PRJNA396135	ncbi.nlm.nih.gov/bioproject/396135
DYS505	PRJNA396136	ncbi.nlm.nih.gov/bioproject/396136
DYS522	PRJNA396137	ncbi.nlm.nih.gov/bioproject/396137
DYS533	PRJNA396138	ncbi.nlm.nih.gov/bioproject/396138
DYS549	PRJNA396139	ncbi.nlm.nih.gov/bioproject/396139
DYS570	PRJNA396140	ncbi.nlm.nih.gov/bioproject/396140
DYS576	PRJNA396141	ncbi.nlm.nih.gov/bioproject/396141
DYS612	PRJNA396142	ncbi.nlm.nih.gov/bioproject/396142
DYS635	PRJNA396143	ncbi.nlm.nih.gov/bioproject/396143
DYS643	PRJNA396144	ncbi.nlm.nih.gov/bioproject/396144
Y-GATA-H4	PRJNA396145	ncbi.nlm.nih.gov/bioproject/396145

d

X-Chromosomal STR Loci – PRJNA380348

ncbi.nlm.nih.gov/bioproject/380348

DXS7132	PRJNA396146	ncbi.nlm.nih.gov/bioproject/396146
DXS7423	PRJNA396147	ncbi.nlm.nih.gov/bioproject/396147
DXS8378	PRJNA396148	ncbi.nlm.nih.gov/bioproject/396148
DXS10074	PRJNA396149	ncbi.nlm.nih.gov/bioproject/396149
DXS10103	PRJNA396150	ncbi.nlm.nih.gov/bioproject/396150
DXS10135	PRJNA396151	ncbi.nlm.nih.gov/bioproject/396151
HPRTB	PRJNA396152	ncbi.nlm.nih.gov/bioproject/396152

allele frequency database, a STRSeq BLAST search determines if a similar or identical sequence has been recorded. When a link to previous data is identified, STRSeq provides nomenclature information and leads the analyst to published allele frequency data (see Fig. 4).

5. Future directions for STRSeq

As previously described, sample sets and STRs will be added iteratively, allowing the BioProject to be built further and records to be released in phases. Once created, the GenBank records are expected to be stable but STRSeq should be viewed as a dynamic resource.

Some users will be familiar with NCBI interfaces and will quickly adapt their workflows to access, search, and download records contained in the STRSeq BioProject. While many tutorials exist to facilitate access to NCBI resources (see https://www.ncbi.nlm.nih.gov/guide/all/#howtos), it is likely that most users will prefer customized interface tools specific to this BioProject. Future plans include the development of such tools at strseq.nist.gov, in order to streamline BLAST searches and batch record downloads from the BioProject.

Additionally, we aim to provide a pathway for submission of new sequence records from laboratories performing population sample sequencing. We anticipate an integrated, seamless process whereby users upload population sample sequencing data to the STRidER web portal (http://strider.online) [10] for quality control, and STRidER queries STRSeq for a matching sequence accession number. In cases where the STRidER query finds no match in STRSeq, a process could be initiated to evaluate the sequence and then aim to create a new GenBank record. Such a process would strengthen the STRidER quality control function and expand STRSeq, while harmonizing nomenclature between both resources. This is particularly important for novel sequence variants likely to be encountered as population studies extend their geographic scope or sample numbers.

Fig. 2. Example STRSeq GenBank record, available online at https://www.ncbi.nlm.nih.gov/nuccore/1197990967.

Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence

GenBank: MF044247.1

FASTA Graphics

Go to:

```
LOCUS       MF044247                 163 bp    DNA     linear   PRI 30-MAY-2017
DEFINITION  Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence.
ACCESSION   MF044247
VERSION     MF044247.1
DBLINK      BioProject: PRJNA380554
KEYWORDS    STRSeq, STR, TPOX.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 163)
  AUTHORS   Gettings,K.B., Borsuk,L.A. and Vallone,P.M.
  TITLE     The STR Sequencing Project [manuscript in preparation]
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 163)
  AUTHORS   NIST,A.G.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAY-2017) Applied Genetics Group, National Institute
            of Standards and Technology, 100 Bureau Drive, MS-8314,
            Gaithersburg, MD 20899, USA
COMMENT     Annotation ('bracketing') of the repeat region is consistent with
            the guidance of the ISFG (International Society of Forensic
            Genetics), PMID: 26844919.  Lower case letters in the 'Bracketed
            repeat' region below denote uncounted bases.   The given
            length-based allele value was determined using the designated
            length-based technology.  Variation in the length-based allele
            between individuals or assays can result from indels in flanking
            regions.  The length of reported sequence is dependent on the assay
            (see 'Sequencing technology') and the quality of the flanking
            sequence. This information is provided as part of the STR
            Sequencing Project (STRseq), a collaborative effort of the
            international forensic DNA community.  The purpose of this project
            is to facilitate the description of sequence-based STR alleles.
            Additional resources can be found at strseq.nist.gov.  For
            questions or feedback, please contact strseq@nist.gov.  Allele
            frequency data can be accessed in the strider.online database.

            ##HumanSTR-START##
            STR locus name        :: TPOX
            Length-based allele   :: 7
            Bracketed repeat      :: [AATG]7
            Sequencing technology :: ForenSeq, MiSeq FGx; PowerSeq Auto, MiSeq
            Coverage              :: >30X
            Length-based tech.    :: PowerPlex Fusion, ABI3500xl
            Assembly              :: GRCh38 (GCF_000001405)
            Chromosome            :: 2
            RefSeq Accession      :: NC_000002.12
            Chrom. Location       :: 1489532..1489698
            Repeat Location       :: 1489653..1489684
            Cytogenetic Location  :: 2p25.3
            ##HumanSTR-END##
FEATURES             Location/Qualifiers
     source          1..163
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..163
                     /note="Promega PowerSeq Sequence"
     variation       25
                     /note="C/T SNP"
                     /db_xref="dbSNP:rs115644759"
     misc_feature    120..154
                     /note="Illumina ForenSeq Sequence"
     repeat_region   122..149
                     /rpt_type=tandem
                     /satellite="microsatellite:TPOX"
ORIGIN
        1 tggcctgtgg gtccccccat agattgtaag cccaggagga agggctgtgt ttcagggctg
       61 tgatcactag cacccagaac cgtcgactgg cacagaacag gcacttaggg aaccctcact
      121 gaatgaatga atgaatgaat gaatgaatgt ttgggcaaat aaa
//
```


Acknowledgements

The authors express gratitude to the NCBI staff who have facilitated development of the BioProject: Drs. Lori Black, Melissa Landrum, Ilene Mizrachi, Kim Pruitt, George Riley, and Steven Sherry. The authors also acknowledge the input of the European Commission project DNASEQEX (HOME/2014/ISFP/AG/LAWX/4000007135) and the support of the ENFSI DNA Working Group and thank the many practitioners and researchers who provided valuable feedback.

NIST funding sources and disclaimers: This work was funded in part

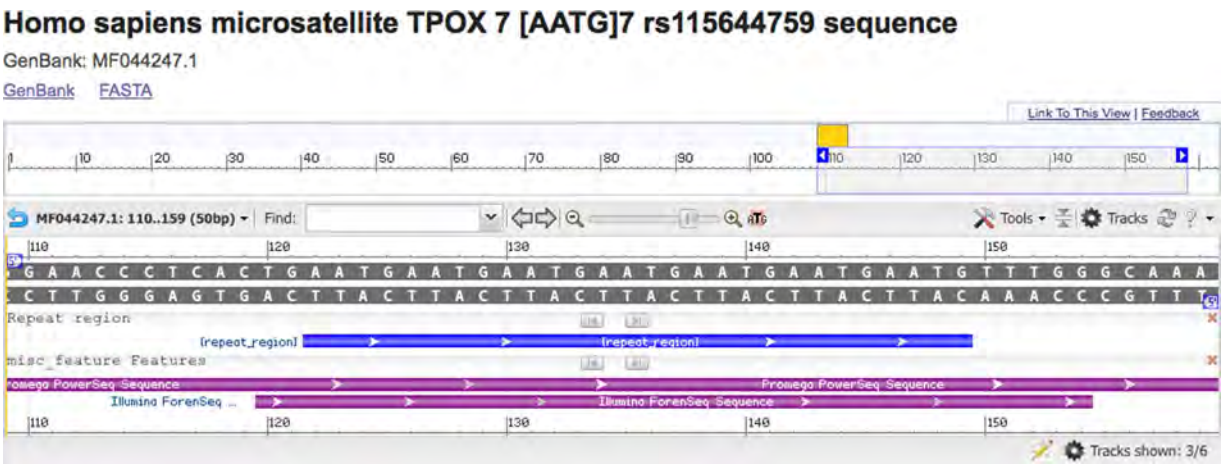


Fig. 3. Example Graphics view of STRSeq Genbank record, available and interactive online at https://www.ncbi.nlm.nih.gov/nuccore/1197990967?report=graph.

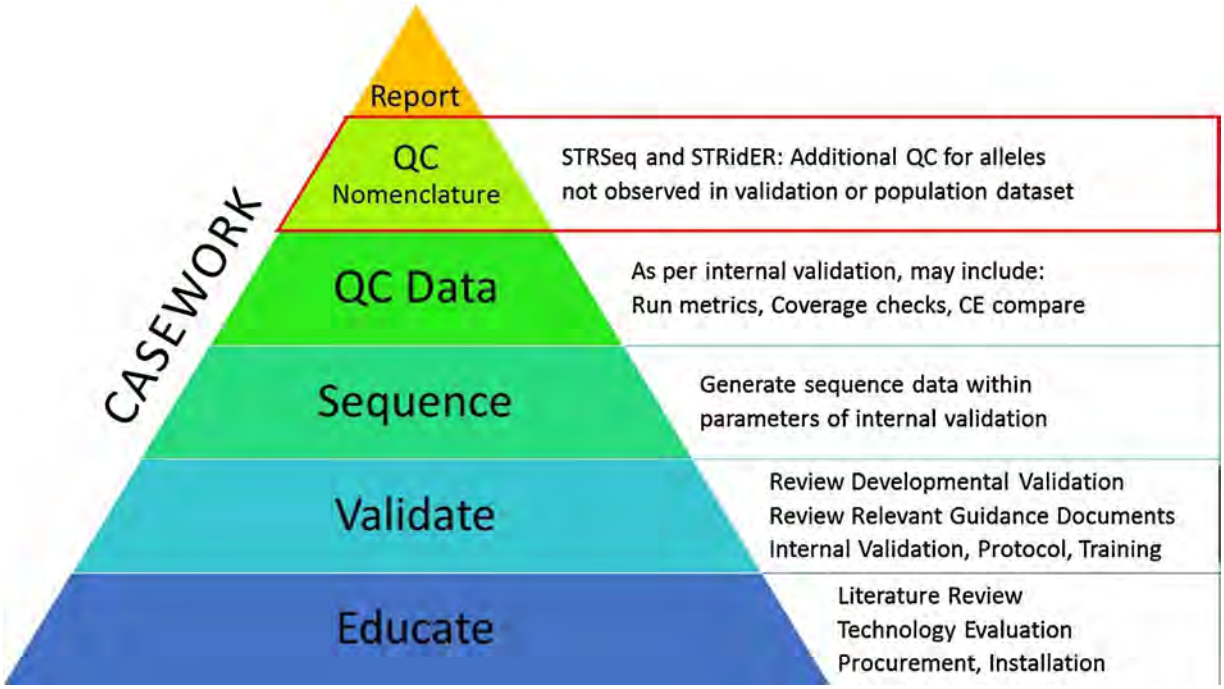


Fig. 4. Outline of the anticipated STRSeq use cases for evaluation of rare alleles in forensic casework, integrated into an overall quality assurance system.

by the National Institute of Justice (NIJ) interagency agreement 1609-602-18NIJ: "Forensic DNA Applications of Next Generation Sequencing". Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Departments of Commerce or Justice. 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 endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

UNT funding sources and disclaimers: This work was supported in part by award no. 2015-DN-BX- K067, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

References

[1] T.Z. Willems, D. Yuan, J. Gordon, A. Gymrek M, Y. Erlich, Genome-wide profiling of heritable and de novo STR variations, Nat. Methods 14 (6) (2017) 590–592.
[2] J.M. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier, USA, 2012.
[3] C. Van Neste, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, Forensic Loci Allele Database (FLAD): automatically generated permanent identifiers for sequenced forensic alleles, Forensic Sci. Int. Genet. 20 (2016) e1–3.
[4] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.
[5] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system, Forensic Sci. Int. Genet. 24 (2016) 86–96.
[6] N.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.
[7] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.
[8] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmao, D.R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs:

considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.
[9] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.
[10] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97–102.
[11] A. Alonso, P. Muller, L. Roewer, S. Willuweit, B. Budowle, W. Parson, European survey on forensic applications of massively parallel sequencing, Forensic Sci. Int. Genet. 29 (2017) e23–e25.
[12] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase A short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (1) (2017) 320–322.
[13] F.R. Wendt, J.L. King, N.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Flanking region variation of ForenSeq DNA signature prep kit STR and SNP loci in yavapai native americans, Forensic Sci. Int. Genet. 28 (2017) 146–154.
[14] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci. Int. Genet. 7 (3) (2013) e82–3.
[15] C.R. Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 MiniSTR loci for improved analysis of degraded DNA samples, J. Forensic Sci. 53 (1) (2008) 73–80.
[16] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, Forensic Sci. Int. Genet. 29 (2017) 193–204.
[17] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18–23.
[18] A.C. Jager, M.L. Alvarez, C.P. Davis, E. Guzman, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J. Pond, J. Varlaro, K.M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70.

Forensic Science International: Genetics 34 (2018) 162–169

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

"The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide

C. Phillips[a,*], K. Butler Gettings[b], J.L. King[c], D. Ballard[d], M. Bodner[e], L. Borsuk[b], W. Parson[e,f]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] National Institute of Standards and Technology, Biomolecular Measurement Division, Gaithersburg, MD, USA
[c] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA
[d] King's Forensics, King's College London, Franklin-Wilkins Building, London, UK
[e] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[f] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA, USA

ARTICLE INFO

Keywords:
Massively parallel sequencing MPS
Short tandem repeat STR
Indels
SNPs
dbSNP
Sequence alignment

ABSTRACT

The STR sequence template file published in 2016 as part of the considerations from the DNA Commission of the International Society for Forensic Genetics on minimal STR sequence nomenclature requirements, has been comprehensively revised and audited using the latest GRCh38 genome assembly. The list of forensic STRs characterized was expanded by including supplementary autosomal, X- and Y-chromosome microsatellites in less common use for routine DNA profiling, but some likely to be adopted in future massively parallel sequencing (MPS) STR panels. We outline several aspects of sequence alignment and annotation that required care and attention to detail when comparing sequences to GRCh37 and GRCh38 assemblies, as well as the necessary matching of MPS-based allele descriptions to previously established repeat region structures described in initial sequencing studies of the less well known forensic STRs. The revised sequence guide is now available in a dynamically updated FTP format from the STRidER website with a date-stamped change log to allow users to explore their own MPS data with the most up-to-date forensic STR sequence information compiled in a simple guide.

1. Introduction

In 2016, an Excel-based STR sequence template file accompanied the set of considerations published by the DNA Commission of the International Society for Forensic Genetics (ISFG) on minimal STR sequence nomenclature requirements [1]. The publication of these considerations was designed to foster consensus in the forensic community about the optimum way to arrange sequence alignments, variant annotation and an eventual allele nomenclature framework necessary for mainstream use of massively parallel sequencing (MPS) to genotype forensic STRs. The first principal guideline was a directive requiring STR sequences to conform to the standardized system, applied to all human microsatellites, of alignment to the genome reference sequence: a haploid, single-strand nucleotide string arranged in a unified p-arm to q-arm direction per chromosome. The second principal guideline recommended that variant annotation: the systematic description of genome sequence differences between individuals, should use the locus identifiers and novel variant reporting methods applied in the 1000 Genomes and NCBI dbSNP databases. It was recognized at the time of publication that sequence variation within the repeat regions of microsatellites presents particular challenges when tracking sequence changes relative to the human reference sequence, which would require care and a period of time to allow early adopters of forensic MPS systems to compile sufficient sequencing data. The STR sequence template file embodied these guidelines by summarizing each STR's sequence alignment and variant/repeat region annotations. As well as mapping the relevant segments of the human reference sequence for each STR, all recorded flanking region variants with more than 10% polymorphism (in one or more population groups) were placed in the context of the STR's repeat region. Annotation was extended to less frequent variants that become important when differentiating repeat region nucleotides from those in flanking regions (e.g. SNPs creating an uncounted repeat unit motif immediately next to the first or last true repeat – see Fig. 1 of [1]). Therefore, defining each STR's repeat region start and end points became the keystone for defining the allelic structure of the marker and protecting its backward compatibility to capillary electrophoresis (CE) genotypes populating all national DNA databases.

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2018.02.017
Received 27 November 2017; Received in revised form 11 February 2018; Accepted 18 February 2018
Available online 21 February 2018
1872-4973/

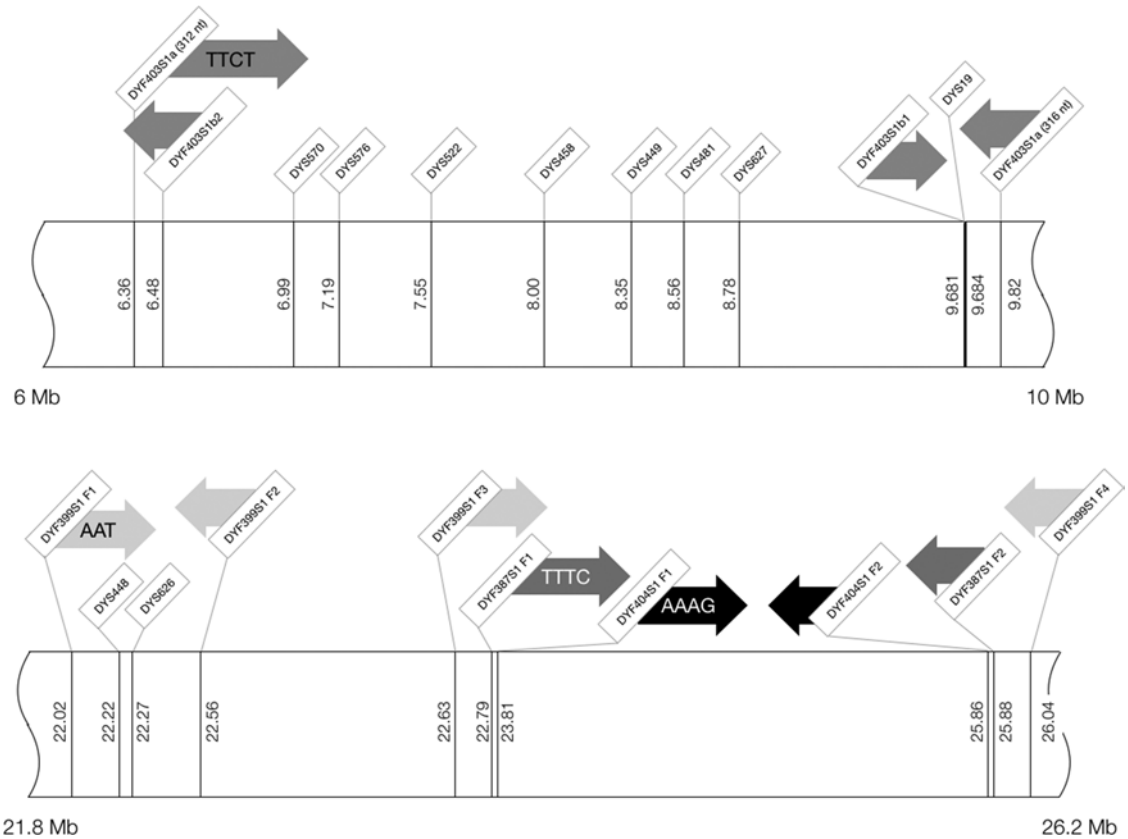


Fig. 1. Arrangement of inverted and replicated Y-STR sequences that are interspersed with single sequence loci in two 4-megabase (Mb) segments of the Y-chromosome. The repeat motif set by the reference sequence direction of the most 5′ fragment is indicated for each multiple allele Y-STR.

With the need for precision and detail in mind, the STR sequence working group (this authorship) have used the original template file as the main data-exchange facility in order to easily update or add variant annotations, as well as explore additional STR loci and re-align many of the originally published repeat region bounds. Periodic releases of a sequence template file continuously revised in this way would lead to confusion, multiple versions in common use and conflicting sequence descriptions. The template file also needs to be regularly calibrated to the most up-to-date human genome sequence build. Therefore, a clear need exists for a dynamic version of the sequence template file that can be placed in an open-access online file transfer scheme. Such a framework is best arranged in a dedicated FTP site with a date-stamped change-log updated at each file release. The STRidER database [2] has been set up to manage the compilation of forensic STR variation data in the MPS era – providing the obvious host site for a dynamic FTP version of the sequence template file.

This paper reports the release of a comprehensively revised sequence template file, herein the STR Sequence Guide, as an FTP file that can be downloaded periodically by forensic DNA practitioners to keep STR sequence information up-to-date. We briefly describe the surprisingly wide range of STR structure and sequence variation factors that the working group considered when revising the sequence data. Detailed descriptions of STR sequences are just as important for established CE genotyping regimes as MPS. The recent discovery of ambiguity in the genomic descriptions of forensic STRs when comparing CE and MPS data [3,4] highlights the need for greater care and accuracy when mapping microsatellites during their initial development for forensic adoption. For this reason, the listed STRs have been expanded to include an additional 45 autosomal, Y-chromosome and X-chromosome loci less commonly used in forensic DNA profiling.

2. Results and discussion

The revised STR Sequence Guide is available for download from the STRidER website at: https://strider.online

In addition to revisions of the STR sequence data and inclusion of less commonly used forensic STRs, a detailed change log has been added listing all changes made to the original sequence template file at the date the revision is checked, agreed (by the authorship) and compiled. An additional worksheet lists all STRs as simple FASTA sequence strings with their GRCh38 chromosome coordinates and individual GRCh38 nucleotides (i.e. one per Excel cell) within the stated 'bedfile' sequence segment bounds.

2.1. Extra forensic STRs added to the original sequence template file

Coverage of forensic STRs was expanded by adding: i. a significant number of autosomal STRs developed for supplemented forensic analyses (e.g. in complex kinship tests), comprising all the newly-adopted markers in supplementary CE kits that were compiled in the 2017 study of Phillips [5]; ii. a further five X-chromosome STRs of the panel of twelve analyzed by the Qiagen Argus X-12 CE kit; iii. additional rapidly-mutating Y-chromosome STRs (RM Y-STRs), yet to be adopted in MPS panels but of interest and forensic utility. All previously listed and additional forensic STRs now included in the revised STR Sequence Guide are detailed in Table 1.

Although MPS analysis of some of the above STRs may not be possible at present, the technology and sequence alignment algorithms to analyze sequences continue to improve to the point where the complex sequence structure of SE33 is likely to be amenable to genotyping by MPS in the near future. More importantly, many STRs are genotyped by CE in mainstream forensic analyses and interpretation of allelic patterns is improved when reference can be made to their repeat

Table 1
Forensic STRs compiled in the revised STR Sequence Guide. Black and gray text distinguishes previously and newly listed X-STRs/rapidly mutating (RM) Y-STRs respectively.

Previously listed common use autosomal STRs		Newly listed additional autosomal STRs	
D1S1677	D9S1122	D1GATA113	D10S2325
D1S1656	D9S2157	D1S1627	D11S2368
TPOX	D10S1248	D2S1360	D11S4463
D2S441	D12ATA63	D2S1772	D13S325
D2S1776	TH01	D3S3045	D14S608
D2S1338	VWA	D3S1744	D15S659
D3S1358	D12S391	D3S3053	D17S974
D3S4529	D13S317	D4S2366	D17S1290
D4S2408	D14S1434	D4S2364	D18S853
FGA	Penta E	D6S477	D18S535
D5S818	D16S539	D6S1017	D18S1364
D5S2500*	D17S1301	D7S3048	D19S253
D5S2800*	D18S51	D7S1517	D20S470
CSF1PO	D19S433	D8S1115	D20S1082
SE33	D20S482	D8S1132	D21S1270
D6S1043	D21S11	D9S925	D21S2055
D6S474	Penta D	D10S1435	D22GATA198
D7S820	D22S1045		
D8S1179			

* D5S2500 and D5S2800 now added to the common use A-STRs

X-STRs	RM Y-STRs	Previously listed Y-STRs (DYS458 newly listed)	
DXS10074	DYF387S1	DYS19	DYS456
DXS10103	DYS449	DYS385	DYS460 (DYS461)'
DXS10135	DYS518	DYS389-I /-II	DYS481
DXS7132	DYS570	DYS390	DYS505
DXS7423	DYS576	DYS391	DYS522
DXS8378	DYS612	DYS392	DYS533
HPRTB	DYS627	DYS393	DYS549
DXS10079	DYF399S1	DYS437	DYS635
DXS10101	DYF403S1	DYS438	DYS643
DXS10134	DYF404S1	DYS439	Y-GATA-H4
DXS10146	DYS526b#	DYS448	DYS458
DXS10148	DYS547		
	DYS626		

#DYS526a = DYS505.
'DYS461 = Incidental STR close to DYS460.

region sequence structures. Sequence data is now divided into four worksheets, adapting the original S1 designations: S1A, common use autosomal A-STRs (35 loci); S1B, common use XY-STRs (29 Y and 7 X loci); S1C, additional A-STRs (34 loci); and S1D, additional XY-STRs (6 RM Y-STRs and 5 X loci). The misidentified D5S2800 STR included in the Thermo Fisher Precision ID STR MPS panel has been placed in the common use A-STRs worksheet, alongside D5S2500 used in several CE kits [5].

2.2. Audit of GRCh38 reference genome builds released between 2013 and 2017

In the original template file, reference sequence was collected from the 1000 Genomes database and cross-checked against the chromosome coordinates of the two main genome assemblies of GRCh37 and GRCh38, using the *In Silico PCR* web-tool to map sequences in the first GRCh38 build (released 17-December-2013). The GRCh38 assembly has undergone periodic revisions that have identified a series of sequence inversions, segmental duplications and translocations with increasing precision. Therefore, a fresh review was made of the most recent stable GRCh38 genome build, GRCh38.p10 (released 1-June-2017), which is held in the Ensemble sequence repository [6]. A new build has since been published: GRCh38.p11, released 14-June-2017, but is not yet viewable in the Ensemble genome browser (each GRCh38 build is available to download at a dedicated NCBI site [7]).

The comparison of each GRCh38 genome build showed no nucleotide differences at any positions originally listed in the sequence template file.

Two additional XY-STRs showed differences in sequence arrangements between GRCh37 and GRCh38 assemblies, one similar in character and the other more complex, than the differences listed for DYS437, DYS438 and DYS439 in the original template file. Simple STR sequence differences were found in DXS10134, showing two more [GAAA] repeats in GRCh38. The DXS10146 sequence is more complex,

with GRCh37 an inversion of the GRCh38 sequence, and four sets of nucleotide differences, including an extra T nucleotide (nt) in the GRCh37 assembly at X:149584331; and an extra [AAAG] repeat unit in GRCh38 at X:150403993-150403996). As a result, 27 repeats for DXS10146 in GRCh38, differs from the Qiagen Argus X-12 CE kit guide listing of 26, based on the original X-STR sequence studies of Edelmann [8]. To match CE and MPS repeat region annotations, an uncounted sequence tract was extended to include the GRCh38-only AAAG nucleotides described above – losing one repeat (uncounted sequence revised to GRCh38 X:150403993-150404002). With these adjusted allele numbers in place for GRCh38, there are also contrasting repeat motifs between each assembly at repeats 13 and 19.

2.3. Revised STR repeat region sequence structure summaries

In the original sequence template file, the summary descriptions of each STR's repeat region structure (in column B) summarized the patterns of repeated motifs seen in the reference sequence. We took the opportunity to revise these summary descriptions to better reflect common variation in forensic STR repeat region structures seen in everyday practice, which may not always be represented by the repeat region in the reference sequence. In some loci there was also some inconsistency in the way complex repeat region structures were formatted to describe the repeat motifs and any uncounted nucleotide tracts between them. Simple format rules were refined for forensic STR repeat region descriptions and applied to their revised data, which now occupy column D, while columns B and C in the STR Sequence Guide now list the bounding GRCh38 coordinates for the whole sequence shown and the repeat region, plus the length of each sequence.

STR repeat region formatting rules apply to: 1) repeating elements (usually termed motifs) and; 2) uncounted nucleotide tracts imbedded within repeat regions. Table 2 outlines these formatting rules in detail with relevant examples. Rule 1.1 applies the established regime of square brackets around each distinct repetitive motif followed by the repeat number applicable. The Guide's repeat region structure summaries now have the suffix 'n' for each bracketed motif, signifying a repeat number would normally be relevant. Rule 1.2 applies a single space between multiple elements (separating each motif and uncounted tracts), if present in the repeat region. Rule 1.3 indicates non-repetitive motifs by an absence of brackets. This rule allows for flexibility in future STR reporting − should such motifs be observed in multiple copies; they can be bracketed. Two examples of adaptation of rule 1.3 in the STR Sequence Guide are the formatting of the first [ATGT]n motif in D6S1043 and the final [TCTA]n motif in D21S11. Multiple copies were observed for the first D6S1043 [ATGT] motif in a 20-repeat allele (KBG), and in the last D21S11 [TCTA] motif in multiple examples of 32.2 and 33.2 alleles (CP). Rule 1.4 describes non-standard motifs that create intermediate alleles, generally from deleted nucleotides, as separate elements. The widely observed TH01 9.3 allele is therefore properly described as: [AATG]6 ATG [AATG]3.

Rule 2 applies lower case formatting to uncounted nucleotide tracts within repeat regions. Uncounted nucleotides lack within-tract spaces and numerical descriptions. For brevity, long uncounted tracts of ten or more nucleotides are simply described with a number in the STR Sequence Guide, e.g. DYS448 [AGAGAT]n N42 [AGAGAT]n. However, proper reporting of such tracts in individual STR sequences should provide the full sequence string, as shown in the example of DYS449 in Table 2.

It is important to note that eleven STR repeat region descriptions now differ from those given in the original sequence template file, as it has been necessary to emphasize the description of common forensic STR variation in the repeat region structure summaries, rather than certain patterns seen in the reference sequence. The details of these STRs are shown in Table 3. The bulk of these revised repeat region descriptions indicate common intermediate alleles from shorter motifs, now included in the appropriate format, but D21S11 is a notable example of the repeat region variation at the 3′ endpoint of: AT [TCTA]n; commonly seen in longer D21S11 alleles, but absent from the reference sequence.

Table 2
Rules for the regularized description of repeat region sequence structures, summarized with examples of the formatting conventions recommended for reporting forensic loci listed in the STR Sequence Guide. Rule 1 items refer to the description of the *counted* elements or motifs in an STR's repeat region. Rule 2 items refer to *uncounted* nucleotides within a repeat region. It is important to note that motifs can have brackets in the Guide's STR repeat region structure summaries, but may not be present in any one individual sequence, or can be observed as a single copy of the motif, so are not bracketed in the STR description.

STR	Length designation	Examples of properly formatted repeat region sequence descriptions
1.1 Brackets and numerical designations are used to shorten the description of repetitive elements/motifs ('n' in the Guide)		
TPOX	6	[AATG]6
SE33	9	[AAAGA]9
1.2 A single space is used between multiple motifs		
D2S1338	18	[GGAA]11 [GGCA]7
D8S1179	15	[TCTA]2 [TCTG]2 [TCTA]11
1.3 Non-repetitive motifs are not bracketed or numbered (but may rarely be present in multiple copies)		
D3S1358	15	TCTA TCTG [TCTA]13
D3S1359	15	TCTA [TCTG]2 [TCTA]12
1.4 Nucleotide tracts creating intermediate length alleles (X.1, X.2, etc.) are formatted as separate elements		
TH01	9.3	[AATG]6 ATG [AATG]3
D1S1656	17.3	CCTA [TCTA]12 TCA TCTG [TCTA]3
FGA	17.2	[GGAA]2 GGAG [AAAG]10 AA AAAA [GAAA]3
D12S391	17.1	AGAT T [AGAT]9 [AGAC]6 AGAT
2. Nucleotide tracts within the repeat region which are not counted are given in lower case and lack spaces		
D19S433	14	[CCTT]12 ccta CCTT cttt CCTT
D19S434	14	[CCTT]9 TCTT [CCTT]2 ccta CCTT cttt CCTT
DXS101048	22	[GGAA]4 [AAGA]12 [AAAG]4 aaggaaag [AAGG]2
DYS449	33	[TTCT]15 ctctctcctcctctttctttcc [TTCT]3 tttcctctttcc [TTCT]15
Note 1: In the STR Sequence Guide, uncounted nucleotide tracts longer than 10 nt are given as numbers for brevity		
DYS449	33	[TTCT]15 N22 [TTCT]3 N12 [TTCT]15
Note 2: Motifs that have been observed to repeat are bracketed in the STR Sequence Guide, but individual sequences may not contain all the motifs of a repeat region, and/or these may not be repetitive (so should not be bracketed).		
D8S1179	Summary structure	[TCTA]n [TCTG]n [TCTA]n
D8S1179	11	TCTA TCTG [TCTA]9
Note 3: No font colours, bold type, italics, sub- or super-scripted characters are used in the repeat region descriptions		

2.4. Inverted multiple-allele Y-STRs

The most complex STR sequence patterns were seen in Y-STRs, in particular RM Y-STRs with multiple amplified sequences and therefore showing two, three or four alleles that may or may not have different sizes in CE. In all cases of multiple sequences, the most 'upstream' 5′ sequence position was used to set the direction for each read and therefore the repeat region structure (e.g. DYS385b is 5′ upstream of DYS385a and has [TTTC] repeats, so the DYS385a sequence is inverted to match). Additional XY-STRs with multiple inverted sequences were: DYF403S1 (4 fragments), DYF399S1 (4) and DYF404S1 (2). DYF399S1 and DYF404S1 were given provisional labels of "fragments" 1–4 and 1–2 respectively, based on their relative 5′ positions, as it is not viable to match the very similar repeat numbers in each sequence to patterns observed in CE genotyping of these loci.

The four sequences of DYF403S1 were labeled according to the

Table 3
Eleven STRs where the repeat region sequence structure summaries given in the STR Sequence Guide do not describe the human reference sequence patterns shown.

STR	STR Sequence Guide repeat region sequence structure summary	Reference Sequence repeat region sequence structure summary	Notes
D1S1656	CCTA [TCTA]n TCA [TCTA]n	CCTA [TCTA]n	TCA motif creates X.3 alleles
D2S441	[TCTA]n TCA [TCTA]n	[TCTA]n	TCA motif creates X.3 alleles
SE33	[CTTT]n TT CT [CTTT]n	[CTTT]n TT [CTTT]n	CT motif in several SE33 alleles
D6S1043	[ATCT]n [ATGT]n [ATCT]n ATGT [ATCT]n	[ATCT]n	[ATGT] motifs common in longer alleles
D6S474	[AGAT]n [GATA]n [GGTA]n [GACA]n	[AGAT]n [GATA]n	
D9S1122	TAGA [TCTG]n [TAGA]n	[TAGA]n	
TH01	[AATG]n ATG [AATG]n	[AATG]n	ATG motif creates X.3 alleles
D12S391	[AGAT]n GA T [AGAT]n [AGAC]n AGAT	[AGAT]n [AGAC]n AGAT	GAT/T motifs create X.3/X.1 alleles
D18S51	[AGAA]n AG	[AGAA]n	AG motif creates X.2 alleles
D21S11	[TCTA]n [TCTG]n [TCTA]n ta [TCTA]n tca [TCTA]n tccata [TCTA]n TA [TCTA]n	[TCTA]n [TCTG]n [TCTA]n ta [TCTA]n tca [TCTA]n tccata [TCTA]n	Final TA [TCTA]n motifs in X.2 alleles
DXS10074	[AAGA]n [AAGG]n [AAGA]n	[AAGA]n	

recommendations of Lee [9] as: DYF403S1a (312 nt in Lee's study); DYF403S1a (316 nt); DYF403S1b1 (341 nt); DYF403S1b2 (437 nt). Note that the two DYF403S1a fragments are identical in their flanking regions so are not be distinguishable as individual sequences. Although this is not a problem in routine genotyping, it shows that identical flanking sequences for two amplified fragments leads to an inability to identify them individually. During developmental studies of DYF403S1 in one contributing laboratory (DB), a fifth sequence fragment of 342 nt was detected using the primers of Lee [9]. Significantly, this fifth amplified sequence is only listed for GRCh38 with *In Silico PCR*, not for the GRCh37 genome assembly. Further investigation revealed the 342 nt fragment was actually the STR DYS627, where enough primer sequence homology has remained between each STR to allow some low level amplification (approximately 20%) of DYS627 from DYF403S1 primers (Supplementary file S1). Therefore, it can be concluded that a proportion of RM Y-STRs represent replicated sequences of previously established Y-chromosome microsatellites that have diverged sufficiently to become distinct STRs with locus-specific allelic variation and in certain cases differentiated repeat region structures.

Many multiple allele Y-STRs have relatively large distances between their sequence positions and are interspersed with single-sequence Y-STRs. Fig. 1 summarizes patterns of these Y-STR sequence positions in two 4-megabase sections of the Y-chromosome.

2.5. Incidental microsatellites

Two 'incidental' STRs were identified in the flanking region of the target STRs, making use of the comprehensive 2006 survey of forensic Y-chromosome STRs by Hansen and J Ballantyne [10]. First, STR DYS461 (GRCh38, Y:18888804-18888851) is separated by 104 nt on the 5′ side of DYS460 and has been annotated in the same way as the target STR. DYS460 primers used for CE analysis bind between each STR, so DYS461 does not influence DYS460 fragment length estimations. However for analysis of DYS460 with the Illumina Forenseq DNA Signature kit, both STRs are amplified together but DYS461 variation is not reported. Second, STR DYS467 is separated by 50 nt on the 3′ side of DYS389-II (GRCh38, Y:12500662-12500709). DYS467 is also circumvented by existing CE and MPS primers and may not be highly polymorphic, although it comprises 12/14 GATA repeats (two more repeats in GRCh37), suggesting a standard microsatellite locus. Although incidental STRs closely sited to the target STR may not always be amplified in forensic tests, we decided it is informative to track all polymorphisms found in flanking regions, not just SNPs and Indels.

The RM Y-STR DYS526 listed in additional XY-STRs, was reported in the 2010 study of K Ballantyne et al. [11] as two loci: DYS526a and DYS526b. However, Hansen and Ballantyne identified DYS526a as independent STR DYS505, separated by 93 nt on the 3′ side of the DYS526b locus [10]. Both STRs are included in the sequence details of DYS526.

2.6. Mobility-shift SNPs

Studies of SE33 and DYS481 sequence variation have identified a mobility shift effect from the presence of flanking region SNP variant alleles creating altered DNA folding patterns [12,13]. Although denaturing CE protocols should reduce formation of secondary structures, the effect appears to be consistent in certain kits and explainable from modeling the stem and loop structures formed by the sequence change (e.g. where a SNP variant allele forms a new C-G triple bond). The three SE33 SNP variants comprise: rs549958510-A; rs189881506-T; rs538644460-T; and the DYS481 SNP variant is rs368663163-A. As it is important to match sequence-based repeat number data with genotyping from CE fragment length estimations with the knowledge of potential discordant genotypes, we have highlighted the presence of the above SNPs with simple orange labels.

Two additional mobility shift SNPs have recently been identified in the Penta E and D2S441 flanking regions and were added to the Sequence Guide. The Penta E SNP variant is rs188309642-G, creating a −1 nt mobility shift confined to 11 repeat alleles in this STR [14]. The D2S441 flanking region variant is a G > C substitution creating a −2 nt mobility shift which does not have an rs-number, located at: GRCh38, 2:68011921 (personal communication, Rita Weispfenning, Promega).

2.7. Compiling insertion-deletion polymorphisms sited in repeat and flanking regions

It is important to track all insertion-deletion polymorphisms (Indels) within repeat regions as they can create a high level of variability in the sequence; are often population specific; can be ambiguously positioned; and when combined with Indels in flanking regions may create isometric fragments that go undetected by CE. We found it difficult to match the 1000 Genomes or dbSNP annotations of Indels sited in repeat regions with the accumulating knowledge of these variants from forensic MPS studies.

The most common forensic STR repeat region Indel that forms the TH01 9.3 allele, provides a good illustration of the difficulties of identifying such Indel positions. All TH01 9.3 MPS sequences collected to date show an [A/-] deletion in the seventh repeat which can be annotated as: [AATG]6 ATG [AATG]3. Previous sequence studies of rarer TH01 alleles indicated 6.3 = [AATG]3 ATG [AATG]3 [15]; 8.3 = [AATG]5 ATG [AATG]3; and 10.3 = [AATG]6 ATG [AATG]4 [16]. The reference sequence consists of 7 AATG repeats (GRCh38 11:2171088-2171115, placing the deleted A nucleotide in 9.3 alleles at 11:2171112). However, dbSNP reports the 9.3 sequence change as the [-/AATGAATGATG] 11 nt insertion rs763206927 (GRCh38

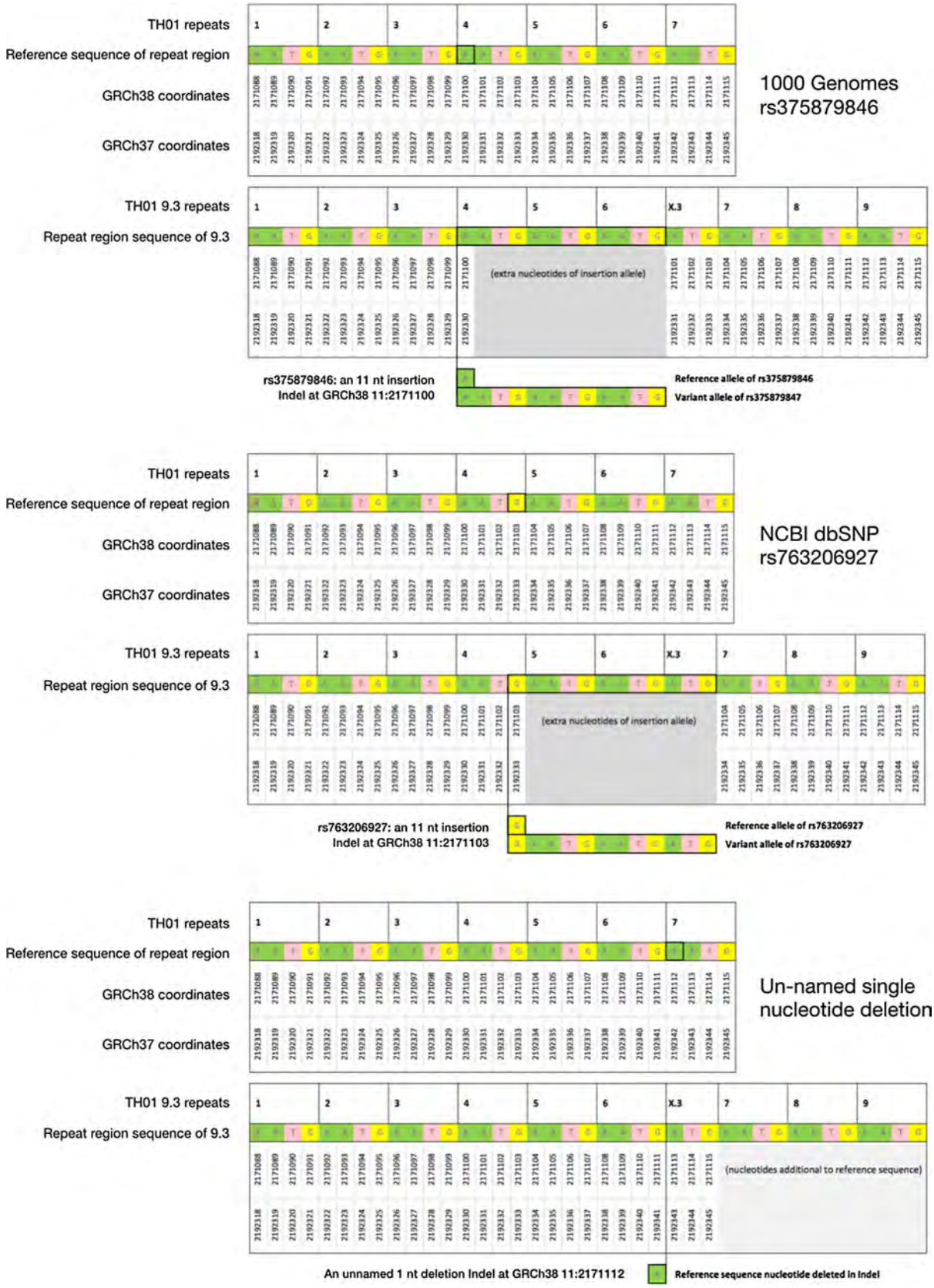


Fig. 2. Three alternative annotations of the TH01 9.3 repeat allele placed in relation to the reference sequence. Note that dbSNP describes the rs763206927 insertion Indel as [-/AATGAATGATG] without reference to the shared 5′ G nucleotide at 11:2171103, in contrast to the 1000 Genomes annotation system for Indel variants.

11:2171103). The 1000 Genomes annotations of TH01 consist of the [-/GTGAA/GTGAATGAA] 5 nt and 9 nt insertion alleles in rs554658416 (GRCh38 11:2171084, 5′ upstream of the repeat region), plus the [-/ATGAATGAATG] 11 nt insertion rs375879846 (GRCh38 11:2171100). The rs375879846 variant in 1000 Genomes matches the 9.3 allele frequency in Europeans (0.2783) and produces the correct sequence, but treating the 9.3 variant as an insertion in order to adjust the 6 repeats of the reference sequence seems counter intuitive when TH01 MPS sequences align in all positions apart from the deleted nucleotide at GRCh38 11:2171112.

Having three alternative ways to annotate the TH01 9.3 allele (Fig. 2) highlights the difficulty of describing copy number variation of one type located within copy number variation of another type, so it should be no surprise that current Indel annotation of repeat region reference sequence is not always consistent and remains incomplete. Therefore, it is a sensible policy to avoid overly precise descriptions of Indels in repeat regions, whether common or rare. Since forensic MPS STR data will be handled as full sequence strings, the annotation of Indels that are likely to occupy different positions in a range of alleles is unnecessary. The existing system of describing repeat regions with bracketed repeat motifs also captures any Indels that occur in these sequence tracts.

Despite the difficulty of annotating Indels, we retained details of two well-characterized Indel loci close to the 3′ repeat region endpoints of D18S51 (rs575219471) and D19S433 (rs147936416); because they are key to fixing the repeat region 3′ bounds in each STR. Indels positioned in a polymeric sequence tract were placed as insertions or deletions starting at the most 5′ nt, following 1000 Genomes and dbSNP conventions. A further two flanking region Indels have modified details from data given in the original template files: D19S433 has the 2 nt Indel rs745607776 moved to GRCh38 19:29926229-29926230; and Penta D now has an unnamed 13 nt [AAGAAAGAAAAAA/-] Indel deletion forming 2.2 and 3.2 alleles; changed from a 3 nt deletion placed in the first repeat of the reference sequence in the original template file. As an illustration of how knowledge of forensic STR sequence variation can contribute to a growing database of such variation in dbSNP, the 13 nt Penta D Indel has been given the provisional identifier ss2137535200 and this can provide a “place-holding” link to the variant until it is assigned an rs-number by dbSNP and this is added to the STR Sequence Guide.

A 4 nt deletion has been characterized in the D13S317 flanking region since the original template file was published. This unnamed [ATCT/-] Indel on the 3′ side of the D13S317 repeat region at GRCh38, 13:82148077-82148080, is an important factor influencing repeat allele size estimation and has been observed in multiple samples from a range of populations in two contributing laboratories (CP, KBG). Although the deleted nucleotide tract cannot be positioned exactly, we placed it at the start of the deletion at the most 5′ nucleotide coordinate. This 4 nt Indel has been given the provisional identifier ss2137543798 by dbSNP.

Two 3′ flanking region 4 nt deletions in SE33 that potentially influence repeat allele size estimation have been added at GRCh38 6:88277313-88277316 (provisional identifier ss2137535201) and GRCh38 6:88277355-88277358 (rs369314007). Two 5′ flanking region Indels creating intermediate alleles in D7S280 were also added, comprising the [T/-] deletion rs754976988 (GRCh38, 7:84160203; X.3 alleles) and the [T/TA] insertion, provisionally ss2137543824 (GRCh38, 7:84160204; X.1 alleles). Lastly, in D9S1122, the 5′ flanking region [TG/-] deletion rs754976988 was added (GRCh38, 9:77073816-77073817; creating X.2 alleles).

3. Concluding remarks: considerations for moving towards an agreed STR allele nomenclature system in the future

The phrase “the devil’s in the detail” describes how a seemingly simple task can turn out to be more complicated than supposed, as individual details produce unforeseen problems. This has often been the case during the compilation of sequence data, thoroughly revised here from the original sequence template file, in order to strengthen the foundations for a forensic STR allele nomenclature system. A persistent challenge has been the need to match repeat region structures found in the reference sequence and in MPS data, with the repeat allele numbers suggested by early Sanger sequencing analyses of STRs genotyped by CE. We have often used historical precedence, when the first published sequences of an STR allowed a repeat structure to be proposed. However, a period of comparative studies will be increasingly necessary for the less commonly used STRs compiled here. We place importance on the inclusion of as many forensic STRs as possible, since it is likely that MPS multiplexes will continue to expand and the compilation of often little used STRs provides a properly curated set of genomic details about their sequence characteristics alongside the core STRs. This is particularly important for RM Y-STRs that unsurprisingly, tend to be found in the more unstable regions of the Y-chromosome, which in turn may have influenced choice of these microsatellites in the first generation of forensic MPS STR panels. However, Y-STRs generally appear more prone to multiple sequences; Hansen and Ballantyne [10] observed ~13% of 417 forensic microsatellites on the Y-chromosome had two duplicated sequences (40 Y-STRs) or 3, 4, 5 and 9 duplications (11 Y-STRs). The close similarity in sequence between DYF403S1 and DYS627 we highlight in this report also suggests that replicated STR sequences eventually evolve into differentiated microsatellites. Such STRs can have distinct patterns of repeat variation, but may still retain enough sequence homology to cause problems in distinguishing their amplified fragments when they are combined in the same PCR.

The complexities revealed when GRCh37 and GRCh38 genome assemblies are compared, underlines the importance of a single stable reference sequence to act as the template on which all MPS sequence data can be reliably aligned. At all the forensic STR sequence tracts checked, GRCh38 has not changed in four years of re-assembly, but shows critical differences with GRCh37 at certain nucleotides. This issue is highlighted by the need to re-annotate the repeat region of DXS10146 because GRCh38 differs in sets of nucleotides in four separate positions, from the GRCh37 assembly originally used to map the repeats. We recommend exclusive use of the GRCh38 human genome sequence to align forensic MPS data, but retain the GRCh37 coordinates because publications still commonly map sequence variants to GRCh37 positions. In the case of DXS10146, the Argus X-12 kit’s ladder fragments, control genotypes and supporting literature made use of the GRCh37 assembly to name the repeat alleles in the component X-STRs. As the 1000 Genomes project has now officially completed its work, the transition to GRCh38 coordinates for all variants in this large-scale catalog of human variation is in process, as we expected to happen [1]. However, the 1000 Genomes Data Slicer tool that uses GRCh37 coordinates, combined with dbSNP that has both, currently provides the best way to check variation found in STR flanking regions.

Assessments of the range of sequence variation in forensic STRs from collective efforts such as STRSeq [17] will accelerate the progress towards an agreed sequence allele nomenclature framework, but these initiatives will be greatly helped by contributions from the whole community. A dynamically revised STR Sequence Guide makes the submission of new sequence discoveries from any forensic MPS practitioner wishing to compare their own data, much more straightforward. Recent discussions and exchange of details in the STR sequence working group have been prompted by revisiting CE information as much as new data generated from MPS. Therefore, anyone with an interest in understanding forensic STR sequences are free to access, and via STRidER contribute new variant annotations to, the revised STR Sequence Guide launched with this publication.

NIST funding sources and disclaimers

This work was funded in part by the National Institute of Justice

(NIJ) interagency agreement1609-602-18NIJ: "Forensic DNA Applications of Next Generation Sequencing". Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Departments of Commerce or Justice. 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 endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

UNT funding sources and disclaimers

This work was supported in part by award no. 2015-DN-BX- K067, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

Acknowledgements

The authors wish to thank Rita Weispfenning of Promega for sharing data on the D2S441 mobility shift variant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2018.02.017.

References

[1] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. de Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.

[2] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97–102.

[3] C. Phillips, W. Parson, J. Amigo, J.L. King, M.D. Coble, C.R. Steffen, P.M. Vallone, K.B. Gettings, J.M. Butler, B. Budowle, D5S2500 is an ambiguously characterized STR: identification and description of forensic microsatellites in the genomics age, Forensic Sci. Int. Genet. 23 (2016) 19–24.

[4] M. Whittle, More on the genomic identification of forensic STRs, Forensic Sci. Int. Genet. 25 (2016) e1–e2.

[5] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, Forensic Sci. Int. Genet. 29 (2017) 193–204.

[6] The Ensembl GRCh38 genome browser at: http://www.ensembl.org/index.html? 1000 Genomes GRCh37 variant catalogs, Data Slicer, etc. now hosted by Ensembl, at: http://grch37.ensembl.org/index.html Accessed November 2017.

[7] The NCBI GRCh38 genome assembly/build downloads and details page at: https://www.ncbi.nlm.nih.gov/assembly?term=GRCh38&cmd=DetailsSearch Accessed November 2017.

[8] J. Edelmann, S. Hering, C. Augustin, R. Szibor, Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28, Forensic Sci. Int. Genet. 2 (2008) 41–46.

[9] E.Y. Lee, H.Y. Lee, S.Y. Kwon, Y.N. Oha, W.I. Yang, K.J. Shin, A multiplex PCR system for 13 RM Y-STRs with separate amplification of two different repeat motif structures in DYF403S1a, Forensic Sci. Int. Genet. 26 (2017) 85–90.

[10] E.K. Hanson, J. Ballantyne, Comprehensive annotated STR physical map of the human Y chromosome: forensic implications, Leg. Med. 8 (2006) 110–120.

[11] K.N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, Y. Choi, K. van Duijn, M. Vermeulen, S. Brauer, R. Decorte, M. Poetsch, et al., Mutability of Y-chromosomal microsatellites: rates characteristics, molecular bases, and forensic implications, Am. J. Hum. Genet. 87 (2010) 341–353.

[12] D.Y. Wang, R.L. Green, R.E. Lagacé, N.J. Oldroyd, L.K. Hennessy, J.J. Mulero, Identification and secondary structure analysis of a region affecting electrophoretic mobility of the STR locus SE33, Forensic Sci. Int. Genet. 24 (2012) e7–e8.

[13] E.Y. Lee, H.Y. Lee, K.S. Shin, Off-ladder alleles due to a single nucleotide polymorphism in the flanking region at DYS481 detected by the PowerPlex® Y23 System, Forensic Sci. Int. Genet. 24 (2016) e7–e8.

[14] D. Ballard, A., Viggars, J., Stickley, L., Devesse, D. Syndercombe Court, C. Phillips, A SNP-based mobility shift in the Penta E short tandem repeat, submitted to Forensic Sci. Int. Genet.

[15] M. Klintschar, Z. Kozma, N. al Hammadi, F.A. Fatah, C. Nöhammer, A study on the short tandem repeat systems HumCD4, HumTH01 and HumFIBRA in population samples from Yemen and Egypt, Int. J. Legal Med. 111 (1998) 107–109.

[16] B. Brinkmann, A. Sajantila, H.W. Goedde, H. Matsumoto, K. Nishi, P. Wiegand, Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci, Eur. J. Hum. Genet. 4 (1996) 175–182.

[17] K.B. Gettings, D. Ballard, L. Devesse, J.L. King, W. Parson, C. Phillips, P. Vallone, STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.

Forensic Science International: Genetics 42 (2019) 14–20

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A nomenclature for sequence-based forensic DNA analysis

Brian Young*, Tom Faris, Luigi Armogida

NicheVision Forensics, LLC. 526 South Main St. Akron, OH, 44311, USA

ARTICLE INFO

Keywords:
Nomenclature
Forensic DNA analysis
Massively parallel sequencing
Short tandem repeat

ABSTRACT

Forensic DNA analysis of casework samples using massively parallel sequencing (MPS) technology requires a system of nomenclature for uniquely labeling sequence-based alleles and artifacts. The DNA Commission of the ISFG has published considerations concerning a nomenclature format that addresses the requirement for unique labeling of sequences. Nomenclatures based on this format can be used in databasing, or communicating sequence types, but the format is lengthy for software interfaces. The sequence identifier (SID) nomenclature addresses this gap by generating short labels able to uniquely identify all sequences (allelic and artifactual) in single-source or casework profiles. Sequences in casework profiles can be uniquely labeled with only two or three SID characters, making the format compact. SID labels can be used in algorithms for identifying and filtering artifacts, and for expressing associations between artifacts and their likely parent alleles. The nomenclature is suitable for use in downstream mixture analysis by any software able to accept character values rather than numeral values. The SID nomenclature is described, and its ability to discriminate sequence-based alleles and artifacts is demonstrated, and its applicability to forensic mixture analysis is demonstrated.

1. Introduction

Many of the functions in forensic DNA analysis of STR markers such as profiling, databasing and communication are critically dependent upon the availability of a suitable allele nomenclature system. Current PCR-CE profiling methods for STR markers measure length polymorphisms in DNA fragments [1–3]. Expressions of length are simple and compact, involving only a numerical description of the length feature of the fragments. New forensic methods are being introduced [4,5] which are based on PCR-MPS (massively parallel sequencing), and which measure the nucleotide sequence feature of DNA fragments. The sequence feature of DNA fragments is significantly more complex to describe than the length feature. One challenge is representing sequence based STR alleles[1] in a shorthand nomenclature that is i) simple enough for everyday communication in forensic laboratories; ii) compact enough for display in forensic software interfaces; and iii) informative enough to be usable in mixed casework samples.

1.1. The need for a practical sequence-based allele nomenclature

The lack of a universally accepted nomenclature system for sequence-based STR alleles has been cited as a barrier to implementing MPS technology in forensic genetics [4,6]. The nomenclature challenge can be divided into two different aspects: that of establishing a standard for databasing sequence-based alleles; and that of establishing a shorthand for practical representation of sequence-based alleles in everyday procedures performed in forensic DNA analysis.

Allele nomenclature can be arbitrarily complex in databasing applications because modern computers are able to handle enormous complexity. However, verbose nomenclatures can be impractical for routine operations in forensic DNA analysis such as comparing DNA profiles, interpreting mixed DNA samples, and describing profiles in court. A more compact nomenclature is needed for these activities.

1.2. Forensic marker nomenclature systems

Nomenclature for STR alleles has been a topic of considerable discussion for as long as STR markers have been used in forensic DNA analysis. Over this time the term nomenclature has been used with two related but slightly different emphases. One focuses on describing the repeat structure of the STR locus proper and defining what portions of that structure should be included or excluded when reporting the length feature of an STR allele (e.g. see [7]). The second emphasis focuses on uniquely discriminating each of the alleles in the set of possible alleles

* Corresponding author.
E-mail addresses: brian@nichevision.com (B. Young), tom@nichevision.com (T. Faris), luigi@nichevision.com (L. Armogida).

[1] DNA segments in forensic DNA analysis may include more than one STR, SNP or DIP marker. Multi-marker sequences are commonly termed haplotypes. However, the term allele will be used to refer to these sequence types in this paper.

https://doi.org/10.1016/j.fsigen.2019.06.001
Received 8 March 2019; Received in revised form 11 May 2019; Accepted 5 June 2019
Available online 05 June 2019
1872-4973/

at a locus (e.g. see [8,9]). Here we describe a novel nomenclature that focuses on the unique description of STR alleles. This nomenclature is intended for everyday operations in forensic DNA analysis. While the nomenclature can also be used in databasing, we do not emphasize that possible application. In order to clearly differentiate the proposed system from current practice, we briefly review selected relevant nomenclature systems.

1.2.1. Indexed bracket nomenclature

Indexed bracket nomenclature has achieved near-universal acceptance for communicating the repeat structure of STR alleles; although some variations exist such as bracketing the repeat numerals rather than the repeat motif [10–14]. Broad consensus has been achieved for other aspects of the indexed bracket shorthand including the strand to represent and the positions to begin and end bracketing within a DNA sequence [8,9]. Variations of this nomenclature have been developed to include indicators for sequence variants in flanking regions [14,15] or to improve stutter artifact labeling [16]. In its pure form, a weakness of the indexed bracket notation for sequence-based alleles is that it does not account for variation that may occur in PCR amplicons outside the STR variable region. Systems to account for this variation [8,9,17] can be complex to implement in software.

1.2.2. Allele number nomenclature

The allele number is a compact shorthand nomenclature used for routine description of alleles and for databasing and has gained universal acceptance in PCR-CE analysis. A weakness of the allele number nomenclature is that it cannot discriminate same-length but different-sequence alleles (aka isoalleles) except by resorting to additional indicators such as appended prime marks or letters [18]. A weakness of indicator systems is that they require interlaboratory coordinating mechanisms to avoid the use of the same indicators for isoalleles.

1.2.3. Database-managed nomenclature

Databases can be constructed where unique codes can serve as compact keys that point to distinct DNA sequence values (e.g. [19]). These systems require coordination and ongoing curation. A weakness of key-value databases is that artifactual sequences observed in casework samples may not be represented in the database, yet these artifacts require labeling and interpretation in mixture analysis.

1.2.4. ISFG DNA commission considerations for sequence-based nomenclature

The DNA Commission of the International Society for Forensic Genetics (ISFG) has published a nomenclature format [8,9] that incorporates both the indexed bracket and allele number nomenclatures while addressing some weaknesses of both for sequence-based alleles. Herein we refer to this format by its original authors (Parson et al., 2016). This nomenclature has been implemented in forensic DNA analysis software in parallel with allele number nomenclature where allele numbers are used in graphics requiring compact displays [20–22]. A weakness of the Parson et al. (2016) nomenclature is the relatively large number of characters needed to fully describe the sequence. A weakness of using allele numbers to label alleles in graphical displays is that isoalleles stack on top of one another. Stacking is manageable in single-source samples but can become complex in mixed samples where three or more isometric allelic or artifactual sequence types may stack at a given allele number position.

Here, we describe a sequence-based allele nomenclature for PCR-MPS data that has attractive features for implementation in software interfaces. The sequence identification (SID) nomenclature captures the sequence variation of entire PCR amplicon fragments, or substrings of them, yet is compact enough for use with complex forensic profile graphics exhibiting many alleles and artifacts. The method for generating SID labels is fully described for implementation in local bioinformatic pipelines, and an executable module for creating labels is available upon request. Application of SID labels to artifact management in mixed samples is described, and the use of SID labels in mixed DNA analysis software interfaces is demonstrated.

2. Materials and methods

2.1. Calculation of SID nomenclature labels

Nomenclature labels are calculated by the SID method in a series of steps as follows: 1) the SHA-256 hash function [23] is used to create a 256-bit digest of a DNA sequence of interest which is expressed in hexadecimal (base-16); 2) the hexadecimal output of the hash function is converted to hexavigesimal (base-26); 3) letters in the hexavigesimal number are capitalized, while all numerals are left unchanged; 4) the order of the characters is reversed so that the hexavigesimal digits appear left to right from least significant to most significant; 5) each digit is converted to its equivalent ASCII decimal number; 6) each decimal number is incremented using an offset of 10 (decimal) if the original hexavigesimal digit was a letter or an offset of 17 if the original hexavigesimal digit was a number; 7) each new decimal value is converted to the corresponding ASCII character. The method was implemented in a C# executable which is available upon request as an EXE or DLL file that can be incorporated into local pipelines. Optionally, the method can be implemented locally in a script because all the steps are outlined in a worked example provided in Supplementary Fig. 1. The SHA-256 hash algorithm is readily available as a module in many languages including Python, R and C#. An optional step is to dynamically allocate the minimum number of digits of the full SID label to distinctly identify all sequences within a scope of interest (aka a context). Allocation occurs left to right corresponding to least to most significant digit of the SID label. This (little-endian) order of characters (bytes) was chosen so that dynamic allocation of digits proceeds from in order of increasing significance from left to right. Character data are often left-aligned in tables. Thus, when left-aligned in genotype tables, SID labels will show equivalent significance in each character position even in cases where different SID labels have different numbers of characters due to dynamic allocation. Backward compatibility with the allele number nomenclature can be facilitated by prepending the SID labels with allele numbers.

2.2. DNA sequence discrimination testing

Ability to discriminate DNA sequences was demonstrated using sequences from NCBI BioProject PRJNA380127 [24] (STRSeq database, downloaded May 14, 2018). Each sequence was randomly mutated > 100X with point insertions, deletions and substitutions using a custom PowerShell script. Duplicate sequences were removed, and the total number of sequences was truncated to 100X the original number of distinct sequences at each locus. The resulting data set contained 114,500 distinct authentic and mutated sequences representing 28 STR loci. Sequences at six loci (D10S1248, D17S1301, D20S482, D4S2408, D9S1122, SE33) were too short to support 100X mutations per sequence and were therefore padded with 100 additional nucleotides from the GRCh38 reference sequence split between upstream and downstream. Entire human chromosome assemblies were used to demonstrate the ability of the method to handle arbitrarily large sequence strings (December 2013 assembly GRCh38 GCA_000001405.2 Downloaded December 30, 2018).

3. Results

3.1. Power of sequence discrimination

The power of the SID nomenclature system to discriminate distinct DNA sequences was demonstrated using three different test sets of DNA sequences referred to as A, B and C. Set A consisted of 114,500

authentic and mutated sequences derived from 28 STR markers in BioProject PRJNA380127. The BioProject sequences ranged in length from 50 to 309 nucleotides where the shortest sequence was a deletion mutant of TPOX GenBank Accession MG988076.1, and the longest sequence was an insertion mutant of PentaE GenBank Accession MH232669.1. The average length was 205 nucleotides. The SID method produced a distinct SID label for each of the 114,500 sequences in test set A (see Supplementary Table 1 for all sequences and SID labels). Each SID label contained either 54 or 55 digits. This variability in length is a consequence of converting hexadecimal digits to hexavigesimal. SID labels containing all 54 or 55 significant digits are capable of discriminating ~1.1×10^{77} different sequence strings. Far fewer than the full set of significant digits is necessary for discriminating all 1.2×10^5 distinct sequences in Set A. SID labels can be significantly compressed without loss of information by allocating the minimum number of digits required to discriminate sequences within a context. Within the context of Test Set A, 88% of the sequence types can be discriminated using just 4 significant digits. Distinct sequences that collide at an allocation of 4 digits, may be resolved by allocating more digits incrementally. Thus, within Test Set A, 88% of the sequences were uniquely labeled using only 4 SID digits, a further 11% were resolved using 5 digits and so on (Table 1). One sequence type required 8 SID digits. The allocation of more than 8 SID digits does not improve sequence discrimination because all sequences in this context are already uniquely identified with 8 digits. Dynamic allocation of digits results in SID labels of differing lengths but minimizes the total number of characters necessary to discriminate all sequences within the context.

Allelic profiles encountered in routine forensic DNA analysis scenarios typically contain far fewer distinct sequences than is present in Test Set A. Accordingly, fewer SID digits should be required to discriminate all sequences in a typical profile. Sequence discrimination within forensic profiles was modeled using 1000 profiles generated by randomly selecting sequences from Test Set A. Test set B consisted of 1000 random profiles in which each profile contained 10 sequences from each of 28 markers for a total of 280 sequences per profile. This level of sequence diversity is equivalent to that of a five-contributor mixture in which every contributor is heterozygous at each of 28 loci and every allelic sequence is distinct. An average of 82% of the sequences in the profiles were discriminated with allocations of just two SID digits, and 99% of the sequences were discriminated with three digits. No case was observed where greater than six SID digits was required to distinguish all 280 sequences across a model profile.

Test Set C consisted of all 1145 allelic sequences across all 28 loci downloaded from BioProject PRJNA380127. The context considered here is all BioProject PRJNA380127 sequences within a locus (at the time of download). Except for SE33, all sequences within a locus were discriminated using only 2 SID digits. Most allelic sequences at SE33 were also discriminated with just 2 SID digits, with three sequences requiring allocation of a third SID digit. This result has important implications for labeling of sequence-defined alleles in mixture analysis contexts. Allelic profiles of loci containing any number of contributors, across any number of casework samples cannot contain more allelic sequences than the total number of alleles in the human population. Hence, when the context is defined on a per-locus basis, virtually all sequence-defined alleles will be uniquely discriminated using just two SID digits with just a few requiring three digits. This affords a very desirable display in mixture analysis software. Slightly more digits will be required when the context is the entire profile (Test Set B). However, encoding alleles within loci has always been standard practice in forensics. For example, a length-10 allele at TH01 and a length-10 allele at CSF1PO are both encoded as "10" because the locus-association is usually provided separately.

3.2. Range of sequence input lengths

The SID method is accommodating of any length DNA sequence, ranging from a single nucleotide up to arbitrarily long sequences. This property of length-flexibility derives from the underlying SHA-256 hash function, for which the property has been well documented [25]. Inheritance of this property by the entire SID method was demonstrated using DNA sequences ranging from one nucleotide to entire chromosomes. At the single nucleotide level, the SID method generates 54- or 55-character SID codes for individual nucleotides A, C, G and T, and these four nucleotides can be discriminated by the SID method using two-digit SID codes of TZ, BO, XY and TW respectively. At the whole chromosome level, the SID method again generates 54- or 55-character SID codes for each human chromosome. The set of 25 human chromosome (22 autosomal, X, Y and M) sequences can be discriminated by allocating just three SID digits (Supplementary Table 2). While any length string can be accommodated by the method, the SID labels produced will depend upon the genomic extent of the substring selected. Therefore, it is important that the substring extent be communicated along with the SID labels. One approach is to communicate the laboratory protocol for string trimming. As an example, a laboratory may implement trimming consistent with the trim positions listed in the UAS software Flanking Regions Report (Verogen, Inc.).

3.3. Compression ratio

SID codes achieve extremely high compression ratios through a combination of the SHA-256 hash and the dynamic allocation of SID digits within analysis contexts. Within the context of the entire set of 114,500 test sequences, 99% of sequences were discriminated with just five digits. Within the context of individual allelic panels, all sequences were discriminated in 99% of the profiles with allocations of just 2 or 3 digits. Given an average length of 205 nucleotides, this represents a compression ratio of 98%.

3.4. Combining length and sequence labels

While sequence-based STR alleles have many advantages over length-based alleles, legacy databases are built on length-based alleles. Backward compatibility can be extended to the SID nomenclature by prepending allele number labels to SID labels (e.g. '9.3 YQ'). The allele number is not strictly necessary to discriminate alleles, or to know the length of an allele. The deterministic property of the SID nomenclature method means that the 9.3 YQ allele at the TH01 locus will always have the "YQ" SID code, whereas the 9 allele will always exhibit the CN SID

Table 1
Number of collisions observed in test sets of STR sequences. Set A consisted of 114,500 natural and randomly mutated sequences originating from NCBI BioProject PRJNA380127. Set B consisted of 1000 random profiles each containing ten distinct DNA sequences in each of 28 STR loci.

Context	SID Digits Allocated							
	1	2	3	4	5	6	7	8
Set A	26	676	17,555	101,424	113,932	114,473	114,499	114,500
SEQs Discriminated	(< 1%)	(< 1%)	(15%)	(88%)	(99%)	(> 99%)	(> 99%)	(100%)
Set B	26.00	229.10	277.84	279.91	279.99	280.00		
Avg. Number of Sequences Discriminated Per Profile	(9%)	(82%)	(99%)	(> 99%)	(> 99%)	(100%)		

code, assuming in both cases the segment analyzed corresponds to GRCh38 coordinates chr11:2,171,079..2,171,127 and the flanking regions correspond to the GenBank reference sequence. Thus, the allele length is always knowable from the SID label, however prepending allele numbers is a convenience to avoid the need for lookup tables. The combined allele number and SID label naturally disambiguates same-length but different-sequence alleles (i.e. isoalleles).

3.5. Artifact labeling

Sequence-based alleles present a more challenging artifact management scenario relative to length-based alleles. The principal artifact in fragment analysis by PCR-CE methods is stutter, which can be filtered using peak position and intensity.

By contrast, length measures cannot discriminate all stutter artifacts in PCR-MPS methods where distinct stutter artifacts of the same compound allele can exhibit identical lengths but distinct sequences. Moreover, counterbalancing stutter in separate motifs of compound alleles may yield artifacts with lengths identical to the parent allele.

In addition to stutter artifacts, PCR-MPS methods produce artifacts that arise from errors in sequencing or from base misincorporation during PCR. Here, we term artifactual sequences arising from non-stutter error "sequence artifacts". Sequence artifacts arising from nucleotide misincorporation or nucleotide substitution in sequencing are identical in length to the parent allele. Alternatively, sequence artifacts may arise from nucleotide insertion or deletion error result in artifactual sequences exhibiting lengths typically 1 or 2 nucleotides different from parent alleles.

The SID nomenclature method enables distinct labeling for all sequence-based artifacts appearing in a profile. For example, locus TH01 of sample 2800 M sequenced using the ForenSeq kit revealed a sequence-based genotype of 6 TK, 9.3 YQ. After applying an analytical threshold of 25 reads (0.2% of locus coverage) two stutter artifacts were clearly discernable as 5 LI and 8.3 WC as were three sequence artifacts 6 VS, 9.3 MS and 9.3 ZK (Table 2). Once sequence entities are classified as allele or artifact, the SID nomenclature can be used to uniquely and informatively label each sequence type. Type classifications can be performed manually or using software. Regardless of the classification method, combinations of SID labels can be used to depict the linkages between alleles and artifacts of those alleles. The depiction format used in the MixtureAce™ plugin to ArmedXpert™ (NicheVision, Akron, OH) links "parent" alleles to "child" stutter artifacts using a dot connector, and to non-stutter artifacts using a tick connector (Fig. 1a). The resulting label system effectively discriminates between all true alleles, and all possible artifacts using compact labels. A system of label connectors enables concise communication of allele-artifact associations.

3.6. Utility for mixture interpretation

Once stutter and non-stutter artifacts have been identified and labeled with SID labels, then these SID labels can be leveraged to aid mixture interpretation. SID labels can be used in the common operation of suppressing (aka clicking off) artifacts to reveal allelic profiles (Fig. 1b), and to filter artifacts from profiles prior to mixture interpretation operations such as profile matching (Fig. 1c) and mixture deconvolution.

4. Discussion

The SID method accepts arbitrary-length DNA sequence inputs and returns a fixed-length value (plus or minus one digit) that can be considered a digest or fingerprint of the original DNA sequence. The digest provides a practical handle or name for distinct sequences observed in sequence-based analysis of forensic DNA. This approach to producing distinct handles parallels the way that the numerical allele number nomenclature provides a practical handle for DNA fragment length-

Table 2
Alleles and artifacts observed at the TH01 locus for single-source sample 2800 M. The analyzed DNA segment corresponds to GRCh38 coordinates chr11:2,171,079..2,171,127 which includes the TH01 marker (underlined) and 9 and 13 nucleotides upstream and downstream respectively. Three different allele nomenclature systems are displayed side-by-side for two alleles and five other sequence types observed at the locus. Only two SID digits are required to discriminate the sequences. SID labels are prepended with allele numbers. Allele number nomenclature exhibits collisions between isometric sequences, and an arbitrary lettering system is used to discriminate alleles. The Parson et al. (2016) format descriptions are based on the ISFG DNA Commission considerations [8,9]. Single-nucleotide errors are shown in lower case.

Sequence Type	Read Count	Sequence	SID Code	Allele Number	Parson et al. (2016) Format Description
Allele	5,272	TCCATGGTGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG	6 TK	6a	TH01[CE6]-Chr11-GRCh38-2171088-2171115 [AATG]6
Sequence Artifact	35	TCCATGGTGAATGAAcGAATGAATGAATGAGGGAAATAAGG	6 VS	6b	TH01[CE6]-Chr11-GRCh38-2171088-2171115 [AATG]1[AACG]1[AATG]4
N-1 Stutter	368	TCCATGGTGAATGAATGAATGAATGAATGAGGGAAATAAGG	5 LI	5	TH01[CE5]-Chr11-GRCh38-2171088-2171115 [AATG]5
Allele	6,653	TCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG	9.3 YQ	9.3a	TH01[CE9.3]-Chr11-GRCh38-2171088-2171115 [AATG]6ATG[AATG]3
Sequence Artifact	11	TaCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG	9.3 MS	9.3b	TH01[CE9.3]-Chr11-GRCh38-2171088-2171115 [AATG]6ATG[AATG]3 2171080-A
Sequence Artifact	11	TCCATGGTGAATGAATGAATaAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG	9.3 ZK	9.3c	TH01[CE9.3]-Chr11-GRCh38-2171088-2171115 [AATG]2[AATA]1[AATG] 3ATG[AATG]3
N-1 Stutter	232	TCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG	8.3 WC	8.3	TH01[CE8.3]-Chr11-GRCh38-2171088-2171115 [AATG]5ATG[AATG]3

Data: Courtesy of Elisa Wurmbach, Office of Chief Medical Examiner, New York.

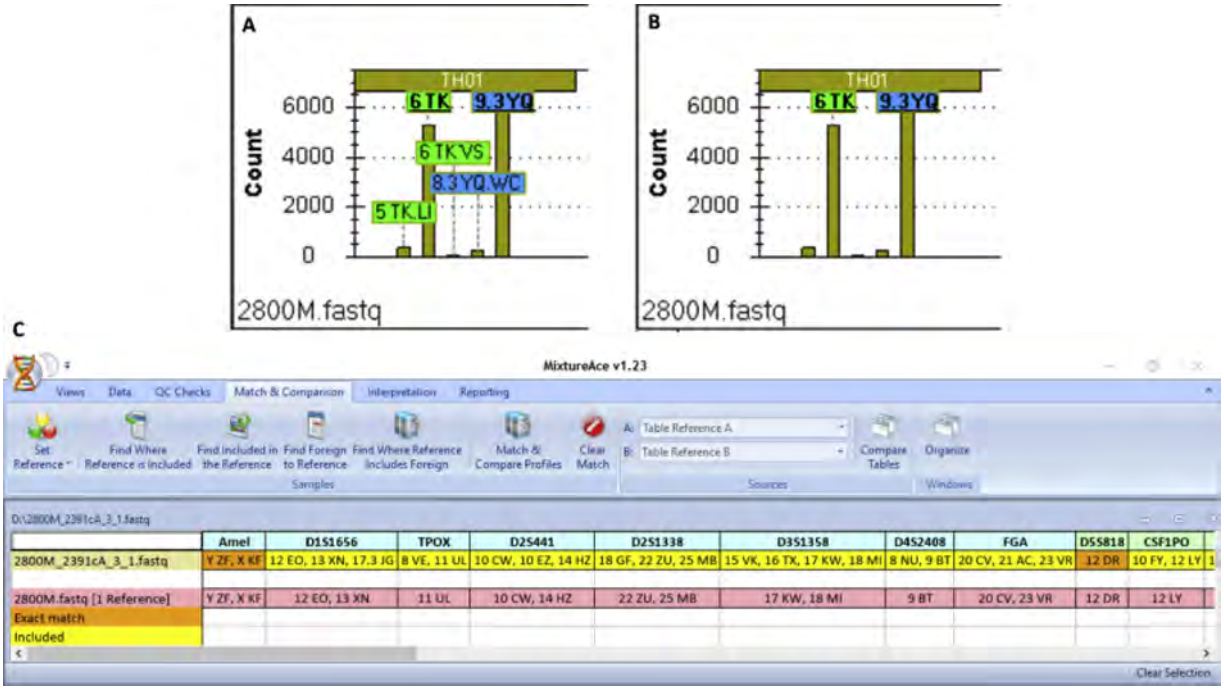


Fig. 1. Illustration of SID nomenclature in software interfaces using data generated with the ForenSeq kit and a MiSeq sequencer. A) Locus TH01 of Promega control DNA 2800 M exhibiting two alleles (6 TK, 9.3 YQ), two N-1 stutter artifacts (5 TK.LI, 8.3 YQ.WC) and one non-stutter artifact (6 TK`VS). SID nomenclature is used to distinctly label the observed sequence types, and dot and tick connectors are used to depict allele-artifact associations for stutter and non-stutter artifacts respectively. B) Artifacts once identified and labeled can be filtered revealing allelic profiles. C) The ArmedXpert match and compare tool is used to illustrate that mixture analysis of sequence-based alleles can proceed using conventional methods if software is able to utilize SID nomenclature. A 3:1 mixture of 2800 M and NIST SRM 2391c component A is compared to a 2800 M reference sample.

based analysis. In both cases, the handles are easy for humans to read and communicate and are short enough for practical use in software interfaces. In both cases, the handles themselves do not explicitly describe the underlying DNA sequence. However, knowledge of the sequence is not necessary during routine forensic analysis activities including comparison of profiles or analysis of mixtures, whether it be performed manually or using computer programs. The fundamental requirement for these operations is that a distinct feature of the DNA is available for analysis; and that a unique label is available for each distinct feature measurement. In PCR-CE analysis, the feature measured is fragment length, converted into units of full and partial STR repeat motifs. Every fragment in a forensic profile has a length feature and every distinct length feature can be labeled using the allele number nomenclature. In PCR-MPS methods, every read has a sequence feature and every distinct sequence can be labeled using SID nomenclature.

Moreover, the SID nomenclature system can be implemented in any laboratory without reference to external databases, genome assemblies or other resources. SID-labeled genotypes will be identical for identical samples in any laboratory anywhere. These features mirror the behavior of allele number nomenclature where any laboratory using any commercial forensic kit will obtain the same genotype for the same sample analyzed by PCR-CE. Just as PCR-CE analysis with commercial kits will always obtain a genotype of 6, 9.3 at the TH01 locus, PCR-MPS analysis with any commercial forensic kit and consistent read trim positions (in this case GRCh38 chr11:2,171,079..2,171,127) will always obtain a SID genotype of 6 TK, 9.3 YQ.

4.1. Discriminatory power

The discriminatory power of the SID nomenclature method is determined by the underlying SHA-256 hash function, which always produces a 256-bit hash value, usually expressed as a 64-digit hexadecimal number. Conversion from hexadecimal to hexavigesimal (base-26) results in a variable-length number of either 54 or 55 digits due to the higher capacity of the higher base. The conversion from a 256-bit binary hash to a base-26 number neither increases nor decreases discriminatory power. The range of SID method is 1.2×10^{77} labels (i.e. 2^{256}) representing a vast capacity many orders of magnitude beyond possible requirements with forensic DNA sequences. The theoretical number of possible sequences in a 200 nucleotide DNA segment is larger at 2.6×10^{120}. However, the maximum observable number of sequences is ultimately limited by the number of chromosomes in the human population (~1.5×10^{10}) [26]. This upper limit can never be reached due to evolutionary constraints. Population surveys of allele frequencies demonstrate that generally fewer than 100 alleles are observable at many forensic STR loci. An exception is the highly polymorphic SE33 locus for which 264 alleles are listed in the STRSeq database [24]. By these considerations, the SID nomenclature has enough safety margin for anticipated DNA variation within loci, across multi-locus profiles of single individuals, or even across multi-locus profiles of the entire human population.

4.2. Partitioning SID codes by locus

The limited sequence diversity across forensic STR loci can lead to SID collisions in the same profile in specific situations. This situation can arise when only the STR locus proper is the subject of analysis. For example, the genotype of Promega 2800 M control DNA is homozygous 12, 12 at both D5S818 and CSF1PO. All four chromosomes exhibit the sequence [ATCT]12, and the same SID code is generated for all four. This is the correct result, as the sequences are identical. When even a single dissimilar nucleotide from the flanking sequence is included in either locus, the SID codes for the D5S818 and CSF1PO alleles will diverge. In casework sample analysis, allele comparisons are made within loci and not across loci. Thus, identical SID codes for identical amplicon subsegments across loci is not an important constraint. Identical allele numbers at different loci within a forensic profile has always been a feature of fragment analysis by PCR-CE.

4.3. Separation of Databasing from routine forensic analysis

The SID nomenclature system is not intended for use in databasing sequence-based alleles. Rather, the SID system is intended to enable routine forensic DNA analysis of sequence-based alleles in computer interfaces including graphical displays of single-source and mixed profiles. The SID nomenclature system also enables artifact management in mixture interpretation scenarios (see Results § 3.5).

Using the SID nomenclature system for routine analysis effectively separates those activities from databasing activities, thereby allowing separate nomenclatures to be optimized for each. In routine analysis, the SID nomenclature permits unique labeling of allelic and artifactual sequences in profiles without complicating the analysis by maintaining sequence features that are not strictly necessary for analyzing profiles. For example, when performing profile comparisons, or mixture analysis it is unnecessary to show the indexed bracket notation at all steps. On the other hand, extracted profiles once ready for databasing can be annotated to any degree the databasing strategy requires.

4.4. Utility of deterministic algorithms in forensic typing

The SHA-256 hash function has been proven to be deterministic by theory and through extensive testing and validation [25]. Conversion of SHA256 digests to SID labels is a function in which elements of the hash function range are connected to elements of the SID label range in a one-to-one relationship. Therefore, SID labels are also deterministic. That is, a given sequence string will always produce the same SID label. This feature creates the opportunity to construct fast lookup tables of SID labels that correspond to specific sequences. For example, a TH01 allele with six repeats flanked by a given length of upstream and downstream nucleotides corresponding to the GRCh38 reference sequence will always yield the same SID code. In the case of the TH01 sequence discussed above (Table 1), the allelic sequence will always produce the SID label:

TKLWNTSSKKJKXAYYYKPTXHQDYPKCBTLUFYAZHCJRTJTYEHQPVBBZTWC. This means that observing the SID label is enough to know the sequence of the DNA fragment. When this label is observed, one knows the allele. The sets of alleles within a locus are relatively small. Observation of a SID label that is not in the lookup table can be an alert that an artifactual or novel allele sequence is present.

4.5. Dependency of SID nomenclature on DNA fragment extents

SID labels are dependent upon the extent of the underlying DNA fragment that is analyzed. This property derives from the discriminative nature of the SID label method wherein sequence fragments with even single nucleotide differences are accorded different SID labels. In PCR-MPS methods, read sequences of PCR amplicons may be bioinformatically trimmed as part of the analysis. When trim positions are changed, the resulting sequence changes through the addition or subtraction of nucleotide letters. This naturally leads to a different set of SID labels. Thus, it is critical that the locus-specific genomic extents used in forensic panels be decided prior to downstream analysis. This is the usual case in forensic analysis, where laboratory analysis conditions are described in protocols and configuration managed. The necessity of specifying extents in sequence-based allele comparisons has been emphasized previously [27].

4.6. Availability of the SID nomenclature method

The SID label generating method is intended for local implementation in bioinformatic pipelines. The steps of the method are fully described in the Materials and Methods section and in Supplementary Fig. 1, and the SHA-256 algorithm is readily available as modules in many major programming languages. Optionally, the method is available upon request as an EXE or DLL file that can be incorporated into local pipelines. For illustration purposes, and for parties not wanting to write a computer program implementing the method, an algorithm that executes the SID nomenclature method is available online at sid.nichevision.com

5. Conclusions

SID nomenclature system described here provides the features necessary to enable sequence-based forensic DNA analysis of mixed casework samples. Specifically, the SID system permits the identification of every distinct sequence in a profile including all alleles and artifacts. The SID nomenclature facilitates mixture interpretation by labeling artifacts distinctly from alleles. The nomenclature can be generated by any laboratory without need for external references or lookup tables. When sequence strings are consistently trimmed to the same genomic coordinates, then the same sample will yield the same SID nomenclature-based allelic profile.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors wish to acknowledge Dr. Elisa Wurmbach for access to data, and Nathanial Caldwell for his pivotal contributions to the SID nomenclature method.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.06.001.

References

[1] E. Buel, M.B. Schwartz, M.J. LaFountain, Capillary electrophoresis STR analysis: comparison to gel-based systems, J. Forensic Sci. 43 (1998) 164–170.
[2] C.J. Fregeau, R.M. Fourney, DNA typing with fluorescent tagged short tandem repeat : a sensitive and accurate approach to human identification, Biotechniques (1993).
[3] A. Jeffreys, V. Wilson, S. Thein, Individual-specific fingerprints of human DNA, Nature (1985).
[4] A. Alonso, P. Müller, L. Roewer, S. Willuweit, B. Budowle, W. Parson, European survey on forensic applications of massively parallel sequencing, Forensic Sci. Int. Genet. 29 (2017) e23–e25, https://doi.org/10.1016/j.fsigen.2017.04.017.
[5] A. Alonso, P.A. Barrio, P. Müller, S. Köcher, B. Berger, P. Martin, M. Bodner, S. Willuweit, W. Parson, L. Roewer, B. Budowle, Current state-of-art of STR sequencing in forensic genetics, Electrophoresis (2018) 1–14, https://doi.org/10.1002/elps.201800030.
[6] R.S. Just, J.A. Irwin, Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results, Forensic Sci. Int. Genet. 34 (2018) 197–205, https://doi.org/10.1016/j.fsigen.2018.02.016.
[7] W. Bär, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W.R. Mayr, B. Olaisen, DNA recommendations. Further report of the DNA Commission of the ISFG regarding the use of short tandem repeat systems, Forensic Sci. Int. 87 (1997) 179–184, https://doi.org/10.1007/s004140050061.
[8] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. De Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63, https://doi.org/10.1016/j.fsigen.2016.01.009.
[9] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The devil's in the detail": release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int. Genet. 34 (2018) 162–169, https://doi.org/10.1016/j.fsigen.2018.02.017.
[10] K.J. van der Gaag, P. de Knijff, Forensic nomenclature for short tandem repeats updated for sequencing, Forensic Sci, Int. Genet. Suppl. Ser. 5 (2015) e542–e544, https://doi.org/10.1016/j.fsigss.2015.09.214.
[11] S.Y. Anvar, K.J. Van Der Gaag, J.W.F. Van Der Heijden, M.H.A.M. Veltrop, R.H.A.M. Vossen, R.H. De Leeuw, C. Breukel, H.P.J. Buermans, J.S. Verbeek, P. De Knijff, J.T. Den Dunnen, J.F.J. Laros, TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes, Bioinformatics 30 (2014) 1651–1659, https://doi.org/10.1093/bioinformatics/btu068.

[12] T.I. Huszar, M.A. Jobling, J.I. Wetton, A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing, Forensic Sci. Int. Genet. 35 (2018) 97–106.
[13] S.L. Friis, A. Buchard, E. Rockenbauer, C. Børsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2016) 68–75, https://doi.org/10.1016/j.fsigen.2015.12.006.
[14] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Børsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41, https://doi.org/10.1016/j.fsigen.2014.04.016.
[15] L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Int. J. Legal Med. 120 (2006) 191–200, https://doi.org/10.1007/s00414-005-0026-1.
[16] S.B. Vilsen, T. Tvedebrink, P.S. Eriksen, C. Bøsting, C. Hussing, H.S. Mogensen, N. Morling, Stutter analysis of complex STR MPS data, Forensic Sci. Int. Genet. 35 (2018) 107–112, https://doi.org/10.1016/j.fsigen.2018.04.003.
[17] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Sci. Int. Genet. 5 (2011) 329–332, https://doi.org/10.1016/j.fsigen.2010.09.005.
[18] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier, 2011.
[19] C. Van Neste, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, Forensic Loci Allele Database (FLAD): automatically generated, permanent identifiers for sequenced forensic alleles, Forensic Sci. Int. Genet. 20 (2016) e1–e3, https://doi.org/10.1016/j.fsigen.2015.09.006.
[20] Anonymous, Converge NGS, (n.d.). https://www.thermofisher.com/us/en/home/technical-resources/software-downloads/converge-software.html.
[21] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F.J. Laros, FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40, https://doi.org/10.1016/j.fsigen.2016.11.007.
[22] Anonymous, ForenSeq Universal Analysis Software, (n.d.). https://www.illumina.com/systems/sequencing-platforms/miseq-fgx/products-services/forenseq-universal-analysis-software.html.
[23] G.M. Lilly, Device for and Method of One-way Cryptographic Hashing, US 2002/0122554 A1, (2002).
[24] NCBI, STRSeq, (n.d.). https://www.ncbi.nlm.nih.gov/bioproject/380127.
[25] Anonymous, Cryptographic Algorithm Validation Program, (n.d.). https://csrc.nist.gov/projects/cryptographic-algorithm-validation-program (accessed February 22, 2019).
[26] Anonymous, US Census Bureau, (n.d.). https://www.census.gov/popclock/.
[27] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130, https://doi.org/10.1016/j.fsigen.2015.06.005.

Forensic Science International: Genetics 43 (2019) 102165

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting

Katherine Butler Gettings[a,*], David Ballard[b], Martin Bodner[c], Lisa A. Borsuk[a], Jonathan L. King[d], Walther Parson[c,e], Christopher Phillips[f]

[a] U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD, 20899, USA
[b] King's Forensics, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London, UK
[c] Institute of Legal Medicine, Medical University of Innsbruck, Austria
[d] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX, 76107, USA
[e] Forensic Science Program, The Pennsylvania State University, USA
[f] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
STR
Sequence
Nomenclature
Bioinformatics

ABSTRACT

This report summarizes topics discussed at the STR sequence nomenclature meeting hosted by the STRAND Working Group in April 2019. Invited attendees for this meeting included researchers known-to-us to be developing STR sequence-based nomenclature schemata, scientific representatives from vendors developing STR sequence bioinformatic methods, DNA intelligence database curators, and academic experts in STR genomics. The goal of this meeting was to provide a forum for individuals developing nomenclature schemata to present and discuss their ideas, encouraging mutual awareness, identification of differences in approaches, opposing aspects, and opportunities for parallelization while some approaches are still under development.

1. Introduction

Since 2016, the *ad hoc* formed STR Sequence Working Group (the authorship of this publication) has been collaborating to harmonize related efforts across our respective laboratories, consisting of: STRidER STR sequence quality control [1], STRSeq catalog of sequences [2], STRait Razor bioinformatic freeware [3], the Forensic STR Sequence Structure Guide [4,5], and large-scale population sample sequencing efforts [6–9] (see [10] for a comprehensive review).

To address the more broadly reaching issue of STR sequence nomenclature, we formalized our group in 2018 as the STRAND Working Group (Short Tandem Repeat: Align, Name, Define). Subsequently, we received the endorsement of the ISFG Executive Board to organize an STR sequence nomenclature meeting, which was held in London on April 11th and 12th, 2019. Invited attendees for this meeting included researchers known-to-us to be developing STR sequence-based nomenclature schemata, scientific representatives from vendors developing STR sequence bioinformatic methods, DNA intelligence database curators, and academic experts in STR genomics. Attendees and affiliations were as follows:

Attendee Name	Affiliation
David Ballard	King's College London, UK
Pedro A. Barrio	National Institute of Toxicology and Forensic Science, Spain
Martin Bodner	Medical University of Innsbruck, Austria
Claus Børsting	University of Copenhagen, Denmark
Lisa Borsuk	National Institute of Standards and Technology, US
Laurence Devesse	King's College London, UK
Kristiaan van der Gaag	Netherlands Forensic Institute, Netherlands
Sebastian Ganschow	LABCON-OWL, Germany
Katherine Gettings	National Institute of Standards and Technology, US
Peter Gill	Norwegian Institute of Public Health, Norway
Theresa Gross	University of Cologne, Germany
Douglas Hares	Federal Bureau of Investigation, US
Cydne Holt	Verogen, US
Jerry Hoogenboom	Netherlands Forensic Institute, Netherlands
Tunde Huszar	University of Leicester, UK
Jodi Irwin	Federal Bureau of Investigation, US
Rebecca Just	Federal Bureau of Investigation, US
Jonathan King	University of North Texas Health Science Center, US
Peter de Knijff	Leiden University, Netherlands
Robert Lagacé	Thermo Fisher, US
Walther Parson	Medical University of Innsbruck, Austria
Christopher Phillips	University of Santiago de Compostela, Spain
Peter Schneider	University of Cologne, Germany
Christian Sell	BKA Wiesbaden, Germany
Sascha Willuweit	Charité University of Medicine Berlin, Germany
Brian Young	NicheVision, US

* Corresponding author at: National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD, 20899-8314, USA.
E-mail address: katherine.gettings@nist.gov (K.B. Gettings).

https://doi.org/10.1016/j.fsigen.2019.102165
Received 12 July 2019; Received in revised form 19 September 2019; Accepted 20 September 2019
Available online 21 September 2019

The goal of this meeting was to provide a forum for individuals developing nomenclature schemata to present and discuss their ideas. Thus, the first day of the meeting was dedicated to attendee presentations, and the second day consisted of group discussion (agenda and presentations permitted for distribution are included in Supplementary File 1). This forum encouraged mutual awareness, identification of differences in approaches, opposing aspects, and opportunities for parallelization while some approaches are still under development. The primary topics are outlined, and related discussions are summarized in this report, which we hope will advance this conversation toward the ultimate goal of an official (ISFG) recommendation on STR sequence nomenclature.

2. Formats for STR sequences

The first outcome of this meeting was consensus on the utility of three formats for STR sequences. The formats are described below, and the relevant presentations are summarized.

2.1. Short designator

For analyzing data within a case, databasing, and for common simple reference in discussion, a minimal code may be useful. Methods for generating such a code were presented and applications were discussed as follows:

1 Brian Young presented a process using the hash function SHA-256 that converts a DNA sequence into a 55 letter sequence identifier (SID) [11]. This SID can be truncated, depending on the application (e.g., identifying sequences within a sample/case may only require two letters). This method is available on GitHub (https://nichevision.github.io/sid.js/) and has been incorporated into ArmedXpert-MixtureAce software (NicheVision), where the SID is appended to the length-based allele and the locus name (e.g., TPOX 12 KG). Linking SIDs together with ticks or dots serves to identify artifacts and stutter, respectively, to primary allele **sequences**: The first outcome of this meeting was s in the software.

2 Sascha Willuweit presented NOMAUT, short for Nomenclature Authority, which is an online repository accessed at nomaut.org. The service allows users to upload a sequence, which is assigned a lower-case letter designator (e.g., TPOX 12 + b) when the submitted sequence is new to the database or is converted to upper-case if already submitted from another source (TPOX 12 + B). NOMAUT seeks to serve as a centralized repository for STR sequence alleles; it can also be used offline, with periodic updates.

3 Rebecca Just presented on using the LUS (longest uninterrupted stretch) to represent sequence alleles and stutter in existing probabilistic genotyping applications [12], and Peter Gill demonstrated the use of LUS-based allele designations in EuroForMix [13]. The designator consists of the locus name, length-based allele, and LUS (e.g., D12S391 23_13 represents an [AGAT]13 [AGAC]9 AGAT sequence/allele). Some loci regularly exhibit multiple alleles which would have the same designator, as in the aforementioned D12S391 23_13 which also describes [AGAT]13 [AGAC]10; however, by extending the designation to secondary or tertiary reference regions, nearly all known alleles can be differentiated. An example locus with rarely non-differentiable alleles under this system is D21S11, at which five subunits of the most common motif have shown variability (indicated by bolded n): [TCTA]**n** [TCTG]**n** [TCTA]**n** TA [TCTA]**n** TCA [TCTA]2 TCCATA [TCTA]**n**.

4 *Included for completeness/context,* Lisa Borsuk presented on the STRSeq BioProject [2](www.ncbi.nlm.nih.gov/bioproject/380127), which is a catalog of sequences maintained as GenBank records at NCBI, where each sequence has a unique accession number (e.g., MH167243.1). STRSeq records are created for sequences published in population studies after quality control. Many STRSeq records represent sequencing results for a single sample across multiple assays, with different ranges of flanking sequence overlap. When a flanking region polymorphism is present outside of the range of one assay, different accession numbers may be assigned to the same sequence in that assay. For example, MH167243.1 and MH167244.1 are both 205 nucleotide (nt) D16S539 sequences with repeat region [GATA]9. These records are differentiated by rs11642858, present 20 nt from the 3′ end of the reported string, included in the ForenSeq range and not in the PowerSeq range. Therefore, the 173 nt PowerSeq sequence is identical for these two accession numbers. If a designator system is recommended by the ISFG DNA Commission, the unique designators could be added and maintained within STRSeq records, connecting such parallel records for easier comparison.

2.2. Bracketed repeat

For condensing the repeat region of a sequence string into a descriptive, “human readable” format, the so-called bracketed repeat is useful for reporting and other applications (e.g., interpretation of stutter). Historically, the original publication characterizing the repeat region for forensic use defined this format, in which the repeat region of the sequence is represented by the repeated motif and the number of repeats. Efforts were made to standardize the start/stop and inclusion/exclusion of neighboring repetitive elements on a per-locus basis [14–19]; however, many exceptions exist due either to historical legacy (locus was characterized before guidance was published), or the inability of a rule set to encompass all scenarios [4,5].

Historically, the bracketed sequence encompassed the start/stop points of the “counted” repeat region. This maximizes the ability to visually discern the length-based allele from the bracketed repeat; however, this approach is not well-suited to some situations (e.g., a 10 allele at D13S317 with the common rs9546005 A > T would be bracketed as [TATC]10 TATC… rather than [TATC]11). In addition, practically speaking, this approach precludes coding programs for automatic bracketing; instead requiring a look-up database. This introduces the possibility of variable approaches among laboratories when sequences are encountered which are not present in the database, particularly at more complex loci such as D21S11 or SE33.

Jerry Hoogenboom and Kristiaan van der Gaag presented a program called STRNaming (manuscript in preparation), which standardizes and automates conversion of the STR string into a bracketed format, based on a defined set of parameters. Similar to genomic sequence alignment methods, points are assigned for desirable features (e.g., length of repetitive run) and penalties are levied for undesirable features (e.g., introduction of gaps). At the time of the meeting, the developers were evaluating settings and preparing to engage users for feedback, with an eventual goal of establishing universal parameters that yield the most coherent arrangement of the repeat region structure and overall data display regarding any locus in present or future use.

Challenges to this approach include a likely change in bracketed designation for some commonly used loci, where significant sequence data have already been published in recent years. Additionally, implementing an algorithm such as this is likely to result in apparent discrepancies between the length-based CE allele number and the bracketed repeat. While STRNaming results in a more inclusive user-friendly representation of the sequence string, the length-based allele number would still be inferred from the full sequence length and is maintained as part of the allele name.

Fig. 1 demonstrates parameterized bracketing for various D13S317 alleles. The length-based CE allele number is explicitly represented in the name, as the bracketed sequence includes additional repeats outside

```
CE11_TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3]
CE11_TATC[10]AATC[3]ATCT[3]
CE11_TATC[11]AATC[2]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]_-24G>A
CE11_TATC[12]AATC[1]ATCT[3]_-25C>T
CE11_TATC[13]ATCT[3]
CE12_TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]
CE12_TATC[12]AATC[2]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
CE12_TATC[13]AATC[1]ATCT[3]_-25C>T
CE12_TATC[13]AATC[2]ATCT[2]
CE12_TATC[14]ATCT[3]
CE13_TATC[13]AATC[2]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]_-24G>A
CE13_TATC[14]AATC[1]ATCT[3]_-25C>T
CE13_TATC[15]AATC[1]ATCT[3]_+9GTCT>-
CE13_TATC[15]ATCT[3]
```

Fig. 1. Example of automated bracketing results for a collection of alleles a the D13S317 locus.

the originally "counted" repeat region. Some length variation can be observed in this "extra" bracketed sequence. The allele name format accommodates sequence variation outside the repeat region by means of variant calls, where variations 5′ or 3′ of the repeat region have negative or positive position numbers, respectively. For example, -25C > T indicates that a T nucleotide was encountered 25 bases 5′ of the repeat region, whereas the reference sequence has a C in that position. Although this particular variant is also known as rs73250432, the nomenclature does not use rs numbers to avoid potential issues with novel variants and the dependency on database lookups.

2.3. Full string

As stated in the 2016 considerations paper [4], the unformatted, entire reported sequence and associated genomic coordinates serve as an unequivocal record of results. The way in which this information is stored (e.g., in the case report, case file, or as a database with corresponding short designators applied per case), falls under the purview of each laboratory.

At this time, forensic DNA databasing software (e.g., CODIS) is generally not equipped to store or search STR sequence strings. Such databases primarily contain convicted offender samples; therefore, enabling STR sequence storage or search capabilities may be of limited use until laboratories begin routinely sequencing this sample type. In the interim, length based (numerical allele) profiles can be developed via STR sequencing assays. Profiles generated with one such assay have recently been approved for upload to the U.S. National DNA Index System (see *CODIS and NDIS Fact Sheet* at https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet#NDIS, accessed May 30, 2019). Analysts confirming interlaboratory matches could compare sequence data, when applicable.

3. Defined coordinates

A second outcome of the meeting is the need for a recommended start and stop per locus, oriented to a reference genome. This is prerequisite to a short designator system. Four possible definitions were discussed; these are described below and applied to the D13S317 locus in Fig. 2.

3.1. Assay specific

Coordinates designed to maximize flanking region sequence per assay/software. Maximizing reported flanking region is desirable for research purposes, to detect private mutations and assess potential association of flanking region polymorphisms with repeat number alleles or a motif. For casework purposes, at some loci, it may be challenging to obtain high quality/high read depth flanking region data for larger alleles. Removing reads because they do not contain high quality flanking region sequence would likely be an undesirable trade-off in low-level samples. A recent analysis of ForenSeq SNP data showed reporting the flanking region nominally decreased read depth (> 95% of reduced region) [20]; however, the effect of these bounds has yet to be reported for the longer amplicons of STRs.

Additionally, assay-centric coordinates would require changes in concert with assay design changes, and the need to establish new coordinate sets for future assays. A key piece of information needed for such coordinates is the "analyzable range" per assay, which has been released for the three existing commercial STR sequencing assays. To facilitate the nomenclature discussion, these ranges have been compiled into Supplementary File 2, a single spreadsheet formatted similarly to the STR Sequence Guide.

3.2. Informative universal coordinates

Coordinates designed to maximize informative polymorphisms in

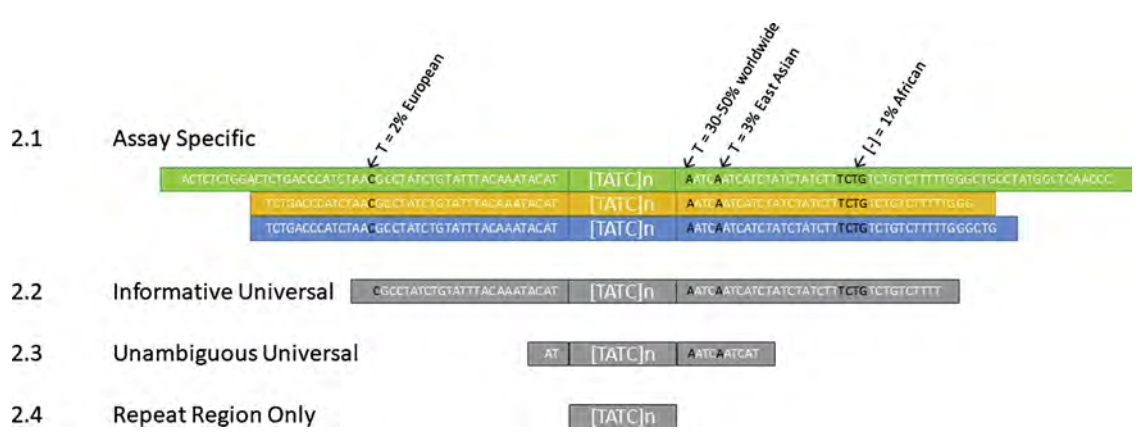


Fig. 2. Four possible range definitions applied to the D13S317 locus. Flanking region polymorphisms > 1% frequency are shown, associated rs numbers are (left to right) rs73250432, rs9546005, rs202043589, rs561167308.

flanking regions across existing assays. Maximizing informative SNPs and indels would lead to increased differentiation of alleles. The above indicated trade-off in quality would still apply. Additionally, considering information gain without regard to current assay design may result in a recommended set of coordinates requiring significant redesign of current manufactured assays (and repeated validation experiments for early adopters).

3.3. Unambiguous universal coordinates

The minimum range of coordinates, which provide unambiguous termination of the designated repeat region. For multiple loci, additional tetranucleotides similar to the repeat motif are present adjacent to the "counted" region. In such cases, a single change may create the appearance of an additional repeat, and often, this change has been observed at measurable frequencies (e.g., D13S317: rs9546005 [adjacent to the repeat in Fig. 2] and vWA: rs199970098). Ambiguous regions such as these would be included/reported under this coordinate definition; the range would terminate when at least two substitutions (not previously observed in tandem) would be needed to create the appearance of an additional repeat.

3.4. Repeat region only

Coordinates defining the "counted" repeat region only. While this approach would work for many loci, there are examples where it would lead to ambiguous sequence reporting (as discussed in Section 3.3) and could result in increased challenges for string searching.

Several considerations regarding defined coordinates were discussed in the meeting, as follows.

For the coordinate definitions in 3.2, 3.3 and 3.4, the concept of a "recommended" range pertains to unifying results across laboratories/assays; high quality data may be present outside of this range. If the eventual recommended range lies within the extent of high quality data, it is expected that some laboratories will continue to interpret flanking region polymorphisms beyond these bounds. It would be the laboratory's own decision to determine how this information is applied. One relevant analogy may be the use of STR allele(s) below analytical threshold on an electropherogram to exclude contributors; however, it is important to distinguish that the analytical threshold is determined based on data quality whereas coordinate definitions 3.2, 3.3 and 3.4 are not directly related to data quality.

One issue pertinent to establishing ranges is that different countries have varied legislation regarding forensic applications of SNP data. As this discussion expands and progresses, it will be useful to understand existing legislation which may prohibit a laboratory from reporting SNPs in these non-coding STR flanking regions.

Any future recommended ranges will exclude the primer sequences, meaning bases reported within these ranges should reflect the genomic sequence of the sample donor rather than the primer sequence used in its amplification. For example, if the recommended range is "repeat region only", the STR sequencing assay primers must bind entirely outside of the repeat region. It is expected some current assay redesign will be required in order to meet this criterion, due to existing examples where the primer binding site appears to extend into the repeat region. Inference of genomic sequence based upon the incorporation of primers is not considered a rigorous scientific approach.

Finally, it has come to the attention of the STRAND Working Group that some researchers have considered the flanking sequence included in the Forensic STR Sequence Structure Guide [5] to be the recommended range. This is not a recommended range, but rather a neutral, arbitrary setting of currently 100 base pairs on either side of the repeat region, designed to highlight significant flanking region sequence features that may only be relevant to some forensic primer designs.

4. Forensic-specific reference

A significant point of discussion in the meeting was the possibility of designating a forensic-specific reference genome (as opposed to, e.g., GRCh38 human genome reference sequence). Three advantages of creating such a reference genome are: a) Elimination of rare SNP alleles in STR flanking regions and incorporation of known insertions; b) Stability, i.e., the forensic community would control changes/updates; c) Ability to create repeat regions most representative of worldwide populations, or representative of maximal complexity. Three arguments against creating such a reference genome are: a) Significant effort would be required for curation, maintenance, version control, and enforcement of general use within the forensic community, b) Duplication of existing effort/infrastructure, c) Impact on established bioinformatic methods.

If it is useful to have forensic-specific references for loci/regions of interest, this can be accomplished by designating STRSeq GenBank records as representative of characteristics, e.g. most common flanking region sequence or most complex repeat region. The annotated reference alleles could be provided in the "STR Seq Nomenclature" page of STRidER, where the Forensic STR Sequence Structure guide is currently made available (https://strider.online/nomenclature).

5. Resources

To ensure all interested parties have access to existing resources, we provide the following tables of population STR sequence data and STR sequence software/tools.

5.1. STR sequence population data

Table 1 contains publications which include at least 50 population samples, with citations ordered by publication date. Populations listed are as defined in the publication.

5.2. STR sequence analysis software

Table 2 contains a list of software currently available for STR sequence analysis and citations or links to additional information.

A final topic, on which a philosophical discussion focused, was that of thresholds; specifically, how thresholds may be implemented more intelligently for sequence data than has been possible for traditional CE methods. Sequencing STR loci allows users to differentiate erroneous sequences of the same length as genomic alleles. With traditional CE methods, amplification errors are incorporated into the RFU intensity of the allele. The discussion centered on the possibility of incorporating into the allele read depth a validated level of sequences determined to have originated from the parent allele, rather than attempting to exclude such sequences via thresholds. This approach could clarify when additional contributors are present in mixed DNA samples and might allow for lower analytical thresholds in general. Furthermore, the possibility of integrating a validated level of sequence-based stutter into the parent allele read depth, was raised. These forward-thinking concepts are presented to encourage discussion; as more thorough exploration of such ideas is beyond the scope of this paper.

Lack of nomenclature is often named as a roadblock to STR sequencing implementation; therefore, our ultimate goal is an official (ISFG) recommendation on STR sequence nomenclature. This follows the

Table 1
Publications containing STR sequence population data.

Citation	Year	First Author	Total Number of Samples	Populations	Sequenced STR Loci	Additional Data	Bioinformatic Method(s)
[6]	2016	Novroski	777	Caucasian Hispanic African American East Asian	27 Autosomal STR 24 Y-STR 7 X-STR	CE-STR	ForenSeq UAS STRait Razor v2.0
[21]	2016	van der Gaag	297	Netherlands Nepal Bhutan Central African Pygmy	17 Autosomal STR	CE-STR	TSSV (FDSTools)
[22,23]	2016, 2017	Wendt	62	Yavapai	27 Autosomal STR 24 Y-STR 7 X-STR	94 iiSNP 56 aiSNP 22 piSNP	STRait Razor v2s
[24]	2017	Casals	231	Spanish Roma Catalans	27 Autosomal STR 24 Y-STR 7 X-STR	94 iiSNP	ForenSeq UAS
[25]	2017	Silva	59	South Brazilian	22 Autosomal STR 23 Y-STR	CE-STR	Altius Cloud System
[26]	2018	Borsuk	1036	Caucasian African American Hispanic Asian	1 Autosomal STR (SE33)	CE-STR	STRait Razor v2.0
[7]	2018	Devesse	400	White British British Chinese	27 Autosomal STR	CE-STR	ForenSeq UAS
[9]	2018	Gettings	1036	Caucasian African American Hispanic Asian	27 Autosomal STR	CE-STR	ForenSeq UAS STRait Razor v2.0
[27]	2018	Huszar	100	African European Australian Asian Near and Middle Eastern American	23 Y-STR	CE-STR	FDSTools v1.1.1
[28]	2018	Kim	209	Korean	27 Autosomal STR 24 Y-STR 7 X-STR	CE-STR	ForenSeq UAS
[8]	2018	Phillips	944	CEPH (51 populations)	27 Autosomal STR 24 Y-STR 7 X-STR	CE-STR	ForenSeq UAS
[29]	2018	Salvador	143	Filipino	7 X-STR	CE-STR	ForenSeq UAS STRait Razor v2s
[30]	2019	Hussing	363	Danish	26 Autosomal STR 24 Y-STR 6 X-STR	CE-STR 94 iiSNP 56 aiSNP 22 piSNP	STRinNGS 1.0 ForenSeq UAS
[31]	2019	Hwa	119	Taiwanese	27 Autosomal STR 24 Y-STR 7 X-STR	CE-STR 94 iiSNP	ForenSeq UAS
[32]	2019	Wu	108	Han Chinese	27 Autosomal STR 24 Y-STR 7 X-STR	CE-STR	ForenSeq UAS
[33]	2019	Barrio	496	Spanish	31 Autosomal STR	CE-STR	Converge 2.0 STRait Razor v3.0

tradition of STR allele designation guidelines coming from the ISFG [16,17] and further evolving as the technology expanded (e.g. Y-STRs [18,19]). Such an approach encourages a rigorous, science-based system. We view this meeting as the first step towards STR nomenclature recommendations; the STRAND WG is committed to facilitating continued dialogue among practitioners, researchers, vendors, and database representatives.

With this communication, we invite the broader forensic community to actively contribute in these discussions. Individuals interested in receiving future communications and/or meeting invitations from the STRAND Working Group may register by email strandwg@gmail.com (please include a brief description of your work in STR sequencing/bioinformatics). Feedback emailed to strandwg@gmail.com will be distributed and discussed at future STRAND Working Group meetings.

Table 2
STR Sequence Analysis Software.

Name	Availability
Agnostic, freeware	
FDSTools [34]	Python Package Index; www.fdstools.nl
Seqmapper [35]	http://forensic.mc.ntu.edu.tw:9000/SEQMapperWeb/Default.aspx
STRait Razor v2s [3] STRait Razor 3.0 [36]	https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor/
STRinNGS [37]	Upon request from the University of Copenhagen
ToaSTR [38]	https://www.toastr.de/
Agnostic, for purchase	
ExactID	https://www.battelle.org/government-offerings/homeland-security-public-safety/security-law-enforcement/forensic-genomics/exactid
GeneMarkerHTS	https://softgenetics.com/GeneMarkerHTS.php
Armed Expert Mixture Ace	https://nichevision.com/mixtureace/
Assay specific, for purchase	
Converge	https://www.thermofisher.com/order/catalog/product/A35131
Universal Analysis Software	https://verogen.com/products/

Funding and Disclaimers

This work received support from the European Union grant agreement number 779485-STEFA - ISFP-2016-AG-IBA-ENFSI and the Special Programs Office of the U.S. National Institute of Standards and Technology. Identification of commercial assays and/or software is not intended to imply recommendation or endorsement by the U.S. National Institute of Standards and Technology.

Appendix A. Supplementary data

Supplementary material related to this article can be found in the online version at doi:https://doi.org/10.1016/j.fsigen.2019.102165. **Supplementary File 1.** Attendee list, meeting agenda, and presentations from the meeting (when permitted by the presenter). **Supplementary File 2.** Flanking region analysis ranges provided by assay manufacturers for 24 autosomal STR loci (includes loci reported in at least two assays). In the first tab, sequences are aligned to a simplified version of the Forensic STR Sequence Structure Guide (current version without assay tracks is available at https://strider.online/nomenclature); range shown is four bases beyond farthest manufacturer range. *PowerSeq 46GY* (tracks in blue) are the analysis ranges in GeneMarkerHTS (v2.0.4); *ForenSeq DNA Signature Prep Kit* (tracks in orange) are the analysis ranges included in the UAS (v1.3) flanking region report; *Precision ID GlobalFiler NGS STR Panel v2* (tracks in purple) are the ranges specified in the target file (Precision_ID_GlobalFiler_NGS_STR_Panel_Targets_v1.1.bed), available at https://www.thermofisher.com/us/en/home/technical-resources/software-downloads/converge-software.html. The second tab contains a .bed file of the information in the first tab.

References

[1] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci. Int. Genet. 24 (2016) 97–102.
[2] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, J. King, W. Parson, C. Phillips, P.M. Vallone, STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.
[3] J.L. King, F.R. Wendt, J. Sun, B. Budowle, STRait Razor v2s: advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci. Int. Genet. 29 (2017) 21–28.
[4] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmao, D.R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.
[5] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int. Genet. 34 (2018) 162–169.
[6] N.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.
[7] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D.S. Court, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups, Forensic Sci. Int. Genet. (2017).
[8] C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Alvarez Iglesias, A. Freire-Aradas, N. Oldroyd, C. Holt, D. Syndercombe Court, A. Carracedo, M.V. Lareu, Global patterns of STR sequence variation: sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis 39 (21) (2018) 2708–2724.
[9] K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci. Int. Genet. 37 (2018) 106–115.
[10] A. Alonso, P.A. Barrio, P. Muller, S. Kocher, B. Berger, P. Martin, M. Bodner, S. Willuweit, W. Parson, L. Roewer, B. Budowle, Current state-of-art of STR sequencing in forensic genetics, Electrophoresis 39 (21) (2018) 2655–2668.
[11] B. Young, T. Faris, L. Armogida, A nomenclature for sequence-based forensic DNA analysis, Forensic Sci. Int. Genet. 42 (2019) 14–20.
[12] R.S. Just, J.A. Irwin, Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results, Forensic Sci. Int. Genet. 34 (2018) 197–205.
[13] O. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44.
[14] A. Urquhart, C.P. Kimpton, T.J. Downes, P. Gill, Variation in short tandem repeat sequences–a survey of twelve microsatellite loci for use as forensic identification markers, Int. J. Leg. Med. 107 (1994) 13–20.
[15] C. Puers, H.A. Hammond, L. Jin, C.T. Caskey, J.W. Schumm, Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder, Am. J. Hum. Genet. 53 (1993) 953–958.
[16] W. Bar, B. Brinkmann, P. Lincoln, W.R. Mayr, U. Rossi, DNA recommendations–1994 report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems, Int. J. Leg. Med. 107 (3) (1994) 159–160.
[17] W. Bar, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W.R. Mayr, B. Olaisen, DNA recommendations: further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems, Int. J. Legal Med. 110 (4) (1997) 175–176.
[18] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M.A. Jobling, P. de Knijff, M. Kayser, M. Krawczak, W.R. Mayr, N. Morling, B. Olaisen, V. Pascali, M. Prinz, L. Roewer, P.M. Schneider, A. Sajantila, C. Tyler-Smith, DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Int. J. Legal Med. 114 (6) (2001) 305–309.
[19] L. Gusmao, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the

recommendations on the use of Y-STRs in forensic analysis, Int. J. Legal Med. (2005) 1–10.
[20] J.L. King, J.D. Churchill, N.M.M. Novroski, X. Zeng, D.H. Warshauer, L.H. Seah, B. Budowle, Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 60–76.
[21] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, P. de Knijff, Massively parallel sequencing of short tandem repeats-population data and mixture analysis results for the PowerSeq system, Forensic Sci. Int. Genet. 24 (2016) 86–96.
[22] F.R. Wendt, J.D. Churchill, N.M. Novroski, J.L. King, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system, Forensic Sci. Int. Genet. 24 (2016) 18–23.
[23] F.R. Wendt, J.L. King, N.M. Novroski, J.D. Churchill, J. Ng, R.F. Oldt, K.L. McCulloh, J.A. Weise, D.G. Smith, S. Kanthaswamy, B. Budowle, Flanking region variation of ForenSeq DNA signature prep kit STR and SNP loci in Yavapai native Americans, Forensic Sci. Int. Genet. 28 (2017) 146–154.
[24] F. Casals, R. Anglada, N. Bonet, R. Rasal, K.J. van der Gaag, J. Hoogenboom, N. Solé-Morata, D. Comas, F. Calafell, Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations, Forensic Sci. Int. Genet. 30 (2017) 66–70.
[25] D. Silva, F.R. Sawitzki, M.K.R. Scheible, S.F. Bailey, C.S. Alho, S.A. Faith, Genetic analysis of Southern Brazil subjects using the PowerSeq AUTO/Y system for short tandem repeat sequencing, Forensic Sci. Int. Genet. 33 (2018) 129–135.
[26] L. Borsuk, K.B. Gettings, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for the SE33 locus, Electrophoresis 0 (2018) 1–8.
[27] T.I. Huszar, M.A. Jobling, J.H. Wetton, A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing, Forensic Sci. Int. Genet. 35 (2018) 97–106.
[28] S.Y. Kim, H.C. Lee, U. Chung, S.K. Ham, H.Y. Lee, S.J. Park, Y.J. Roh, S.H. Lee, Massive parallel sequencing of short tandem repeats in the Korean population, Electrophoresis 39 (21) (2018) 2702–2707.
[29] J.M. Salvador, D.L.T. Apaga, F.C. Delfin, G.C. Calacal, S.E. Dennis, M.C.A. De Ungria, Filipino DNA variation at 12 X-chromosome short tandem repeat markers, Forensic Sci. Int. Genet. 36 (2018) e8–e12.
[30] C. Hussing, R. Bytyci, C. Huber, N. Morling, C. Borsting, The Danish STR sequence database: duplicate typing of 363 danes with the ForenSeq DNA signature prep kit, Int. J. Legal Med. 133 (2) (2019) 325–334.
[31] H.L. Hwa, M.Y. Wu, W.C. Chung, T.M. Ko, C.P. Lin, H.I. Yin, T.T. Lee, J.C. Lee, Massively parallel sequencing analysis of nondegraded and degraded DNA mixtures using the ForenSeq system in combination with EuroForMix software, Int. J. Legal Med. 133 (1) (2019) 25–37.
[32] J. Wu, J.L. Li, M.L. Wang, J.P. Li, Z.C. Zhao, Q. Wang, S.D. Yang, X. Xiong, J.L. Yang, Y.J. Deng, Evaluation of the MiSeq FGx system for use in forensic casework, Int. J. Legal Med. 133 (3) (2019) 689–697.
[33] P.A. Barrio, P. Martin, A. Alonso, P. Muller, M. Bodner, B. Berger, W. Parson, B. Budowle, D. Consortium, Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power, Forensic Sci. Int. Genet. 42 (2019) 49–55.
[34] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40.
[35] J.C. Lee, B. Tseng, L.K. Chang, A. Linacre, S.E.Q. Mapper, A DNA sequence searching tool for massively parallel sequencing data, Forensic Sci. Int. Genet. 26 (2017) 66–69.
[36] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait razor 3.0, Forensic Sci. Int. Genet. (2017).
[37] S.L. Friis, A. Buchard, E. Rockenbauer, C. Borsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci. Int. Genet. 21 (2016) 68–75.
[38] S. Ganschow, J. Silvery, J. Kalinowski, C. Tiemann, toaSTR: a web application for forensic STR genotyping by massively parallel sequencing, Forensic Sci. Int. Genet. 37 (2018) 21–28.

Forensic Science International: Genetics 52 (2021) 102473

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

STRNaming: Generating simple, informative names for sequenced STR alleles in a standardised and automated manner

Jerry Hoogenboom [a,*], Titia Sijen [a,b], Kristiaan J. van der Gaag [a]

[a] Division of Biological Traces, Netherlands Forensic Institute, The Hague, The Netherlands
[b] University of Amsterdam, Swammerdam Institute for Life Sciences, Science Park 904, 1098XH Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Forensic science
STR
MPS
NGS
Nomenclature
Allele names

ABSTRACT

The introduction of Massively Parallel Sequencing in the forensic domain has exposed the need for comprehensive nomenclature of sequenced Short Tandem Repeat (STR) alleles. In general, three strategies are at hand: 1) the full sequence mapped to the human genome reference sequence, which ensures exact data exchange; 2) shortened, human-readable formats for forensic reporting and data presentation and 3) very short codes that enable compact figures and tables but do not convey any sequence information. Here, we describe an algorithm of the second type: STRNaming, which generates human-readable names for sequenced STR alleles. STRNaming is guided by a reference sequence at each locus and then functions independently to automatically assign a unique, sequence-descriptive name that also includes the capillary electrophoresis allele number. STRNaming settings were established based on preferences that were surveyed internationally in the forensic community. These settings ensure that a small change in the sequence corresponds to a small change in the allele name, which is helpful for recognising for instance stutter products. Sequence variants outside of the repeat units are indicated as simple variant calls. Since the STR name is sequence-descriptive, the sequence can be traced back from the allele name. Because STRNaming is fully guided by an assignable reference sequence, no central coordination or configuration is required and the method will work for any STR locus, be it autosomal, Y-, X-chromosomal in current or future use. The algorithm is publicly available online and offline.

1. Introduction

Short Tandem Repeats (STRs) represent the main forensic marker type, as typically STR profiling data are stored in (criminal) DNA databases. Traditionally, STR profiles are generated through Capillary Electrophoresis (CE), although Massively Parallel Sequencing (MPS) is an upcoming method in various molecular fields including forensics [1]. MPS has two main advantages over CE-based STR analysis: 1) a higher discriminatory power because of the inclusion of sequence variation which can assist the interpretation of complex mixtures and 2) the high multiplexing capacity for amplicons of similar size which can assist in the analysis of degraded DNA as all amplicons can be short. The output of MPS analyses are sequence reads and read coverage numbers for all the different DNA sequences that pass filtering. Specialised computer software can readily compare such DNA sequences for identicalness but in forensic practise, forensic scientists favour a more intuitive representation to assist when using DNA profile comparison and evaluation tools [2], facilitate discussions with colleagues and ease presentation of the results in reports and to court.

In general, two types of shortened naming schemes for MPS alleles can be envisioned. The simpler is a short code name that uniquely identifies a sequence, but conveys little information about that sequence and how it relates to other sequences. Such code names minimally include the repeat length of the allele, allowing them to be used in comparisons to CE-based DNA profiles, but are otherwise kept as short as possible. Examples of this type of naming scheme are NomAut [3] and FLAD [4], which use a central online database to store the allele name corresponding to each sequence, and SID [5], which uses a one-way hash function to generate a fixed name for each sequence. For the second type of naming scheme, a more intricate nomenclature is developed that seeks to assign a similar name to similar sequences while preserving important sequence characteristics, such as the identity and arrangement of the repeated element(s) which is especially informative for STRs with a complex structure of multiple or interrupted repeats (a.k.a. complex STRs).

Previous proposals of sequence-descriptive STR allele nomenclature

* Corresponding author.
E-mail address: j.hoogenboom@nfi.nl (J. Hoogenboom).

https://doi.org/10.1016/j.fsigen.2021.102473
Received 27 November 2020; Received in revised form 11 January 2021; Accepted 18 January 2021
Available online 29 January 2021

share the common idea that the repeated stretches of sequence can be written in shortened form by indicating the number of repeats and that sequence variation outside the repeat region can be briefly communicated as variant descriptions or dbSNP identifiers [6–8]. However, no consensus existed about the definitions of a 'repeated stretch of sequence' or the 'repeat region' of a locus. In recent years, efforts have been made by the DNA Commission of the ISFG [9,10] and others [7] to establish consensus about these aspects of STR nomenclature, leading to the STR Sequence Structure Guide [10] and the STRSeq initiative [11]. While these efforts have introduced guidance into how STR alleles can be named in a uniform way, the manual application of these guidelines can be laborious and this process is difficult to automate because each locus is addressed individually [12].

Here we introduce the STRNaming algorithm that automatically produces unique, short, human-readable allele names descriptive of the variation in the repeat structure as well as in repeat-flanking or intervening regions. Special care is taken to ensure that names of common artefacts, such as PCR stutter or hybrids (a type of artefact produced by template switching in the PCR [13,14]), are similar to their parental allele(s) so that they are easily recognised as artefacts. Also, in familial studies where mutations may accumulate, it is useful if the name informs whether alleles differ by a one-step mutation (such as a one repeat unit insertion or deletion).

2. Materials and methods

2.1. Definitions

Developing an STR naming algorithm requires definition of the sequence components within an STR allele (Fig. 1). The *repeat structure* extends from the first to the last repeat. *Repeat stretches* represent the sequences within this repeat structure that consist entirely of repeated copies of a same short sequence motif. This motif is called the *repeat unit*. A repeat structure can have repeat stretches consisting of different repeat units. *Interruptions* may intervene the repeat stretches; these are non-repetitive sequences (otherwise it would be another repeat unit). The 5′ and 3′ sequences flanking the repeat structure (up to the primer sequences) are denoted the *prefix* and *suffix* respectively. The combined region of prefix, repeat structure and suffix is denoted *target region*. One can choose to limit the length of the prefix and suffix in reporting; this region is denoted *reporting region*. A target region may contain multiple structures.

2.2. Finding repeats in a reference sequence

The STRNaming algorithm, outlined in Fig. 2, uses the human genome reference sequence (GRCh38, forward orientation) as the basis for allele naming [15]. To establish the start and end position of the repeat structure, the repeat stretches need to be identified. First, STRNaming searches the longest uninterrupted repeat stretch in the entire sequence. When two stretches are of equal length, the most 5′ option takes precedence. Then, the next-longest non-overlapping stretch in the remaining sequence is marked and this process is repeated until no repeat stretches remain that pass the criteria in Table 1A. The repeat unit of each stretch is noted, along with the repeat units that would be obtained by shifting the starting nucleotide of the motif to each nucleotide in the repeat (e.g., when a stretch of GATC-repeats is found, the motifs ATCG, TCGA and CGAT are also examined).

Then, an optimal combination of repeat stretches in the reference sequence is determined using the procedure outlined in Section 2.3. This analysis is repeated when not all repeat units found previously are used in the optimal combination of stretches; this time starting with only the selected repeat units. When this results in a structure containing a repeat stretch of at least four repeats, STRNaming recognises the structure as an STR locus and the genomic locations of its repeat stretches are stored. To maintain compatibility with CE-based data, the length of the repeat structure of the reference allele and its corresponding CE allele number are also saved.

When a sufficiently large amount of flanking sequence is provided as input, STRNaming will repeat this procedure on the 5′ and 3′ flanking sequences to find additional nearby STR structures of at least 20 nucleotides each. This is useful for STR loci which are located very close to one another, such as DYS460 and DYS461.

2.3. Optimal shortening of repeat stretches in an STR structure

STRNaming scans the sequence for all occurrences of each of the repeat units found in the reference sequence. In this step the criteria in Table 1A are ignored, so that also single occurrences of the repeat units are recorded, with a minimum repeat stretch length of 4 nt; an unrepeated occurrence of a trinucleotide repeat unit is also recorded when it is exactly three nucleotides away from a longer repeat of the same unit. Each repeat unit is considered separately and stretches of different repeat units may therefore overlap.

For each of these repeats, STRNaming repeatedly finds the longest non-overlapping repeat stretch in the remaining sequence similar to how this was done for the reference sequence as outlined in Section 2.2 (using the criteria presented in Table 1A). Again, a list of repeat units is constructed, but the motif is not shifted to start at each of the nucleotides in the unit anymore. This way, repeat units may be discovered that were not repeated in the reference sequence. For these units, all additional repeat stretches of at least two repeats (at least four for mononucleotide repeats) are also recorded.

Next, STRNaming proceeds to a combined analysis of all recorded repeat stretches in which repeat stretches must not overlap. In this merging process, the criteria in Table 1B are taken into account plus the limitation that when a repeat unit is used, it must also be used in the repeat stretch where it was first detected. The first and last repeat stretch must use one of the repeat units found in the reference sequence. To select the most suitable name, a scoring process takes place according to the criteria outlined in Table 2. These scoring criteria were established using a large set of different STRs (see Section 2.7) and taking the results of a questionnaire into account (see Section 2.8). Finally, the highest-scoring repeat structure is selected. When multiple repeat structures

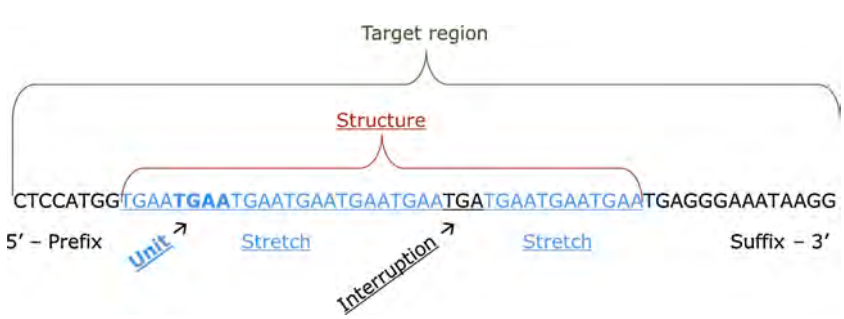


Fig. 1. Sequence components within an STR allele identified by STRNaming.

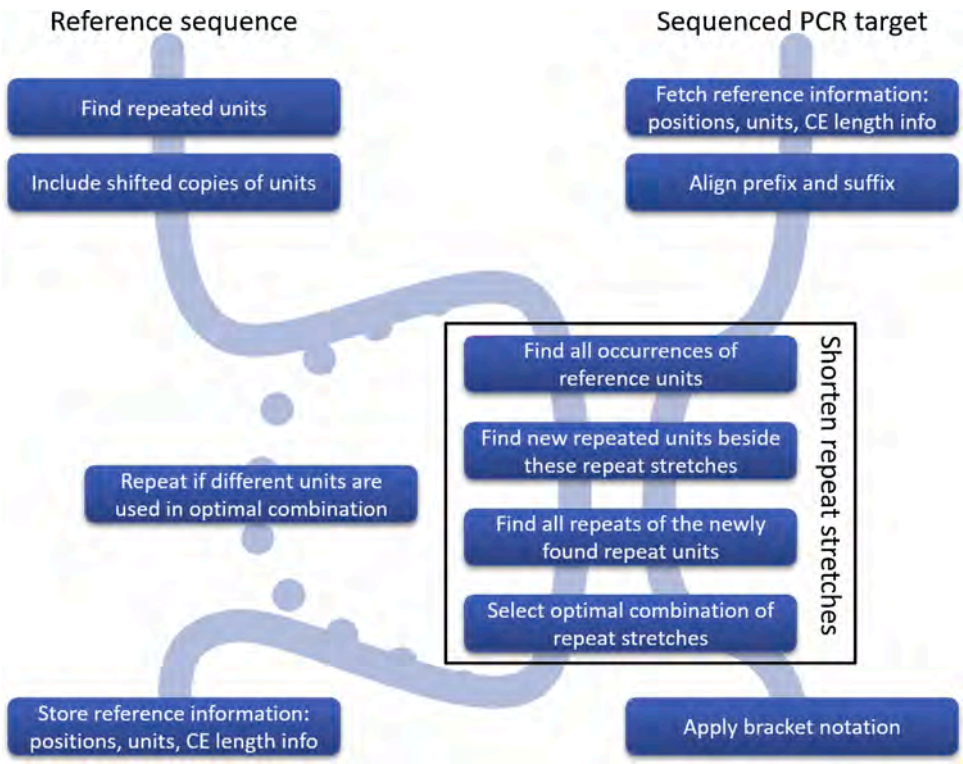


Fig. 2. Outline of the STRNaming algorithm. Two paths extend from the top to the bottom. The left path depicts the analysis steps used for the reference sequence (Sections 2.2 and 2.3). The strand orientation of the repeat is thus determined by the reference sequence. The right path depicts the analysis steps used for naming sequenced STR alleles (Sections 2.3–2.5).

result in the same score, the structure in which the repeat stretches are shifted furthest to the 5′ end takes precedence. As a result, STRNaming will always output the same name for a given sequence and is therefore deterministic.

2.4. Alignment and extent of prefix and suffix

The prefix 5′ end and suffix 3′ end correspond to the ends of the amplified fragment excluding the primers (i.e., the target region). Alternatively, a shorter reporting region can be defined, for instance in case of a shorter amplified region (or when there are legal constraints) for which the 3′ ends of the primers need to be known. Information from the reference sequence analysis (Section 2.2) is used to determine which STR loci reside within the target region (or reporting region). The 3′ end of the prefix is identified by aligning the reference sequence 5′ of the first in-range repeat stretch to the 5′ end of the target sequence (alignment parameters: match score +1, mismatch penalty -1, linear gap penalty -1). Likewise, the 5′ end of the suffix is identified by aligning the reference sequence 3′ of the last in-range repeat stretch to the 3′ end of the target sequence. Then, the optimal combination of repeat stretches is obtained as outlined in Section 2.3.

2.5. Allele name with bracket notation

To convert the optimal repeat structure into a human-readable allele name, multiple mutually compatible notations are possible (Section 3.9). Using the default notation, the allele name starts with the CE allele number followed by an underscore. Then, the repeat units in the repeat structure are listed in 5′ to 3′ order of appearance, each time followed by the number of repeats between brackets. Small repeat interruptions are presented too, but interruptions longer than 8 nt are represented with brackets only (see Section 3.6 for an example). The prefix and suffix that were saved from the reference sequence are omitted. Any sequence variation with respect to the reference sequence in the omitted regions is included in the name in the form of variant calls. The nucleotide position and the type of variant are indicated and separated from the repeat stretches by underscores. For the suffix and long interruptions STRNaming counts 5′ to 3′; for the prefix 3′ to 5′. Variants in the prefix are marked by a '-', variants in the suffix by a '+'. The variants in the long interruption are placed between the brackets. Substitutions are marked > (e.g., C>G); insertions .1-> (e.g, 345.1->G, in which the '.1′ indicates that the insertion occurred between positions 345 and 346) and deletions >- (e.g., C>-). A substitution at the fourth base in the suffix will thus be indicated as '+4C>G'. This way, a short, human-readable, unique allele name is obtained from which the sequence can be traced back. Naming is such that artefacts (stutter products, PCR hybrids) and one-step mutations can be recognised readily from the name. STRNaming can automatically colourize the repeats, using the same colour for the same repeat unit at different locations within an STR locus.

2.6. Naming multiple STR loci in a single target region

When the target region includes multiple or duplicated STR loci, such as DYS389I/II or DYS460/DYS461, STRNaming calculates the optimal combination of repeat stretches for each locus separately. The non-repetitive sequence intervening the loci is treated as a long interruption. A single CE allele number is calculated using the length of the entire target region sequence. These markers are further explored in Sections 3.6 and 3.7.

2.7. Sequence data used for testing

To optimize the parameters of the algorithm, data from 450 samples sequenced with the ForenSeq™ DNA Signature Prep Kit (Verogen) and analysed through FDSTools [16] were used. The ForenSeq kit analyses 58 STRs as indicated by Verogen: 27 autosomal STRs, 24 Y-STRs and 7 X-STRs resulting in 1239 unique alleles named in the 450 samples. To

Table 2
Criteria for the scoring of repeat structures. Positive scores indicate desirable properties, negative scores undesirable ones. For some criteria the score is multiplied by a factor for every subsequent occurrence. For example, the score of one interruption (criterion number 4 in Table 2A) has the rounded value of −9.6, the score of two interruptions is calculated (using the unrounded values) as $-9.6 + (1.4 \times -9.6) = -23.1$, and the score of three interruptions as $-9.6 + (1.4 \times -9.6) + (1.4^2 \times -9.6) = -42.3$. **A.** Scores used for analysing the reference allele and subsequent alleles. **B.** Scores used only for naming subsequent alleles, not the reference allele.

		Criterion	Score	Multiplier
A	1	For every nucleotide covered by a repeat	+ 0.15874379	
	2	For every distinct repeat unit used	− 10.17864730	× 1.09052798
	3	For every repeat of a unit	+ 4.14278510	
	4	For every interruption between repeat stretches	− 9.56645361	× 1.41646677
	5	For every interruption that is exactly one repeat unit in length	+ 7.27483601	
	6	For every nucleotide in an interruption between repeat stretches	− 0.56939138	
B	7	For every repeat of a unit that was not used in the reference sequence	+ 3.49237881	
	8	For every nucleotide inserted or deleted in the prefix, suffix or long interruption	− 1.78595700	

cover ancestry-specific alleles, samples from three geographic origins were included: 80 samples involved donors originating from the Himalayan region [17], 80 from a population of African Pygmies [18] and 290 samples were selected from a large dataset of Dutch males [19].

In addition, to independently validate the STRNaming algorithm, a dataset consisting of 6479 unique sequences originating from 260 different individuals from various South African populations was used. These samples were sequenced with the ForenSeq™ DNA Signature Prep Kit (Verogen) for a population study (Heathfield et al., article in preparation). Sequence data was initially analysed with the Universal Analysis Software (UAS, Verogen) using the default settings and exported as a Flanking Region Report [20]. The raw data were not heavily curated nor cleaned prior to sharing, and as a result, some of the apparent alleles in the dataset used for testing the algorithm may represent commonly recurring artefacts such as PCR stutter.

2.8. Questionnaire

The optimal values for the numbers in Tables 1 and 2 were determined by maximising congruence of STRNaming output with the preference expressed by 26 participants of a questionnaire (including members of the STRAND Working Group) from 16 institutions worldwide. In the questionnaire, ten representative examples for fundamental choices in allele naming were presented. These questions relate to: 1) the preferred length of an interruption in context of the length of the repeat unit; 2) including a repeat with an interruption in the repeat structure or leaving it in the prefix; 3) less or smaller interruptions at the cost of more different repeat units; 4) maximum coverage of the sequence but more different repeat units or a long interruption. Participants were asked to rate their preference on a six-step (in case of two alternatives) or seven-step scale (in one question with three ordered alternatives). In addition, two questions were devoted to choosing between writing the bases or the repeat numbers in brackets and including the length of interruptions larger than 8 nt in the name or not. Finally, one last question was included to determine the most intuitive position numbering to use for insertions in the prefix. The entire questionnaire including the results can be found in Supplementary File 1. A weight of preference for each choice was calculated from the answers and the STRNaming settings in Tables 1 and 2 were subsequently tuned to match these preferences as closely as possible using a particle swarm optimisation algorithm [21] set up to maximise congruence of STRNaming output with the participant's answers. Effects of scoring values distinct from those presented in Table 2, resulting in suboptimal allele names, are provided in Table 3.

Table 1
A. Criteria for initial detection of repeat units and stretches. **B.** Criteria for repeat structures.

	Criterion	Value
A	Minimum number of consecutive repeats	2
	Minimum length of repeat stretches	8 nt
	Maximum length of repeat units	6 nt
B	Maximum length of repeat stretch interruptions[1]	8 nt
	Maximum number of repeat stretch interruptions	5

[1] One 'long interruption' of at most 20 nucleotides is permitted. Longer interruptions lead to separate definitions of repeat structures.

3. Results and discussion

3.1. Reference sequence results

To name STR alleles, STRNaming uses a reference sequence for each locus. This sequence is taken from build 38 of the human genome reference sequence (in the forward orientation) and its name is fundamental for the naming of all other alleles of that locus. The repeat units used in the name of the reference sequences can therefore be embedded in the algorithm together with the genomic coordinates of the prefix 3′ end and suffix 5′ end. As a result, only the coordinates of the prefix 5′ end and suffix 3′ end (i.e., the reporting region) need to be provided to use the STRNaming algorithm. The names and corresponding details for the reference sequences of 60 STR loci are provided in Supplementary Table 1 and an example for D1S1656 is given in Fig. 3. Interestingly, the CE allele number of the reference sequence for this locus is 17, despite having only 16 tetranucleotide repeats. The Forensic STR Sequence Structure Guide defines this marker's structure as 'CCTA [TCTA]n' [10].

3.2. Naming ForenSeq STRs in 450 samples

To optimize the performance of STRNaming, 450 samples from three geographical origins sequenced through the ForenSeq system were analysed and the outcomes are summarised in Table 4. The full list of obtained allele names is provided in Supplementary File 2. Examples of naming issues are detailed in Sections 3.3 to 3.7. As expected, for almost all markers the CE allele numbers at a locus exhibit a logical relationship to the repeat structure. As can be seen in Table 4, this relationship is not the same for all markers, which is partly due to the fact that repeats of some markers have been counted differently in the past. In most cases the CE allele number is congruent to the length of the entire structure or to the longest uninterrupted stretch (LUS). For some markers the CE allele number appears to reflect the length of all repeat stretches excluding the interruptions, or only a part of the structure. The latter occurs when STRNaming includes an additional repeat stretch that was historically not included in the structure. Similar discrepancies have previously been addressed by defining 'counted' and 'uncounted' repeats [7,10].

3.3. Naming a locus with flanking-site variation: D1S1656

Table 5 lists a selection of allele names obtained for D1S1656, showing how sequence variation for three CE15, four CE15.3 and a CE16 allele translates to distinct allele names. All CE15 and CE15.3 alleles have 14 tetranucleotide repeats; the CE16 allele has 15 tetranucleotide repeats. The 15.3 alleles carry in addition an interruption of 3 nt; when

Table 3
Examples of suboptimal names that could be obtained with scores different from those in Table 2. The numbers in the first column correspond to the criteria in Table 2. Grey shaded names correspond to the reference allele, which is unaffected in these examples.

Criterion in Table 2	Locus	Preferred (obtained with scores from Table 2)	Suboptimal
A1: Repeat coverage	D18S51	CE18_GAAA[18]A[1]AG[4]GAAA[2]	CE18_GAAA[18]A[1]AG[4]GAAA[2]
		CE16_GAAA[16]G[1]AG[4]GAAA[2]	CE16_GAAA[15]GAA[1]AG[5]GAAA[2]
A2: Distinct repeat units	D3S1358	CE16_TATC[2]TGTC[1]TATC[13]	CE16_TATC[2]TGTC[1]TATC[13]
		CE17_TATC[2]TGTC[1]TATC[11]CATC[1]TATC[2]	CE17_TATC[2]TG[1]TCTA[11]TCCATC[1]TATC[2]
A3: Reference unit repeats	D1S1656	CE17_AC[6]CTAT[16]	CE17_AC[6]CTAT[16]
		CE12_AC[5]AT[1]CTAT[11]	CE12_AC[5]ATCT[12]_+1CT>-
A4: Interruptions	CSF1PO	CE13_TCTA[13]ATCT[3]	CE13_TCTA[13]ATCT[3]
		CE10.3_TCTA[5]TCA[1]TCTA[5]ATCT[3]	CE10.3_TCTA[5]TC[1]ATCT[5]A[1]ATCT[3]
A5: 'Nice' interruptions	D18S51	CE18_GAAA[18]A[1]AG[4]GAAA[2]	CE18_GAAA[18]A[1]AG[4]GAAA[2]
		CE16_GAAA[13]GATA[1]GAAA[2]A[1]AG[4]GAAA[2]	CE16_GAAA[13]GAT[1]AGAA[2]AA[1]AG[4]GAAA[2]
A6: Interruption coverage	DXS10135	CE23_GAAA[20]	CE23_GAAA[20]
		CE28_GAAA[18]GGAA[2]GAAA[3]GGAA[1]GAAA[1]	CE28_GAAA[18]GGAAGG[1]AAGA[3]AAGGAA[1]GAAA[1]
B7: Non-ref. unit repeats	D3S1358	CE16_TATC[2]TGTC[1]TATC[13]	CE16_TATC[2]TGTC[1]TATC[13]
		CE16.2_TATC[2]TGTC[3]TC[1]TATC[11]	CE16.2_TATC[2]TGTCTGTCTGTCTC[1]TATC[11]
B8: Flanking indels	DYS570	CE17_TTTC[17]	CE17_TTTC[17]
		CE21_TTCC[1]TTTC[20]	CE21_TTTC[1]C[1]TTTC[20]_-1T>-

polymerase slippage occurs at the longest uninterrupted repeat stretches (CTAT[11]/[12]/[13]) the stutter products will also carry a 3 nt interruption and be recognised readily as a stutter. For two alleles variation occurs in the suffix; the suffix sequence is CTACATCA**T**ACAGTT, which means that the C at the ninth position in the reference (counting 5′ to 3′ in the suffix; bold in Fig. 3) has been replaced with a T. For brevity, the sequences of the prefix and suffix are not included in the names; the variant is marked as '+9C>T' (see Section 2.5).

3.4. Naming a locus with multiple tetranucleotide repeat units: D13S317

STRNaming may use multiple distinct repeat units within the same allele name, even when the first repeat unit could also be used for a later repeated sequence. An example is D13S317, where the longest uninterrupted repeat stretch is made with TATC units. TATC can also be used for an ATCTATCTATCT stretch, but STRNaming prefers to use a second repeat motif namely ATCT (Table 6). The use of the ATCT repeat unit prevents the occurrence of interruptions ('ATCT[3]' instead of 'ATC[1]TATC[2]*T*[1]'). Most alleles have in addition a short AATC repeat (blue) between the TATC (red) and ATCT (green) repeats (Table 6). For D13S317 there is no correlation between the CE allele number and the length of the longest uninterrupted repeat, due to length variation in the additional repeat stretches. Both in the prefix and in the suffix, variation occurs that is recurring for several STR lengths (-24G>A which corresponds to rs146621667 with a minor allele frequency (MAF) of 0.0044; -25C>T which corresponds to rs73250432, MAF = 0.0070, +9GTCT>- which corresponds to rs561167308, MAF = 0.0038). Length variation in the AATC stretch (reference: AATC[2]) has previously been reported as rs9546005 (AATC[1], MAF = 0.4215) and rs202043589 (no AATC, MAF = 0.0619 and always together with rs9546005) [11]. In the Forensic STR Sequence Structure Guide, D13S317 is defined as '[TATC]n' [10], which coincides with the major repeat identified by STRNaming.

3.5. Naming a locus with a complex repeat structure: D21S11

Consistent use of the same repeat units for different alleles of the same locus is the most effective way to achieve consistent allele naming. In the STRNaming algorithm, this is achieved by giving higher scores when the same repeat units as in the reference sequence are used for alleles. STRNaming does not strictly enforce using any particular repeat unit; when differences between sequences grow larger, the scoring system allows switching to a different repeat unit if that results in an STR structure with a higher score (and thus having more desirable properties). The D21S11 repeat structure is among the most complex of the autosomal loci in current use. It features alternating repeat units and multiple interruptions that vary in sequence and length. As can be seen in Table 7, STRNaming switches to a slightly different repeat structure for a large group of x.2 alleles, which differ from most other alleles only by an extra TA 5′ of the penultimate TATC repeat unit. This alternative structure consistently obtains a higher score because it avoids introducing the heavily penalised 'TA' interruption. Since polymerase slippage generally occurs at the longest uninterrupted stretches, stutter products will have the same structure as the parent alleles. D21S11 is the only marker for which multiple structure groups have been observed (Table 4, last column).

3.6. Naming two loci in a single target region: DYS460 and DYS461

DYS460 and DYS461 are located close together on the Y chromosome, with only 101 nucleotides between the STR structures defined by STRNaming. As a result, the DYS460 fragment targeted by the ForenSeq DNA Signature Prep Kit includes the DYS461 locus. When analysed with FDSTools, both loci are visible in one sequence. When STRNaming is configured for this target region, a combined allele name is generated that includes both the DYS460 and DYS461 STR structures. The region between both loci is treated as a (large) interruption. As can be seen in

…GAAATAGAATCACTAGGGAACCAAATATATATACATACAATTAA AC[6] CTAT[16]
CTACATCA**C**ACAGTTGACCCTTGA…

Fig. 3. Analysis results for the reference sequence of D1S1656; displayed coordinates Chr1:230.769.561-704. Prefix and suffix are presented in black, repeat structure in red and blue. Note that AT[4] (underlined) is not included as a repeat because that would introduce a third repeat unit and a 12 nt interruption between AT[4] and AC[6], both of which are heavily penalised by the scores in Table 2. The AT[4] repeat is too short to overcome these penalties.

Table 4
Summary of STRNaming results for 1241 unique alleles of 58 STRs in 450 samples.

Locus		Reference sequence				# alleles			Naming (ignoring prefix/suffix)					
Chr[1]	Locus	# distinct repeat units	# interruptions	Largest interruption (nt)	CE allele number corresponds to	Length-based (CE)	Repeat structure	Extra alleles due to prefix/ suffix variation	# alleles with other repeat units than ref	# alleles with more interruptions than ref	# alleles with interruption > 8 nt	# alleles with different relation to CE number	# alleles with unrepeated units	# allele groups with different STR structures
1	D1S1656	2				17	33			24			2	
2	D2S1338	2	1	4	structure	14	54							
2	D2S441	1			structure	13	19	2		12			3	
2	TPOX	1			structure	7	7							
3	D3S1358	1	1	4	structure	9	21	1	17	1			1	
4	D4S2408	1			structure	7	9			2			2	
4	FGA	2	1	3		22	34	2	2	14			36	
5	CSF1PO	2			LUS[2]	10	10			1		1		
5	D5S818	1			structure	8	9	6		1				
6	D6S1043	1			structure	20	29	2		20			1	
7	D7S820	2	1	8	LUS	9	17	5		1		2	22	
8	D8S1179	1	1	4	structure	11	26		3					
9	D9S1122	1	1	4	LUS	8	17			9		10	16	
10	D10S1248	1			structure	11	13	1		1			1	
11	TH01	1			structure	8	8	1		1				
12	D12S391	2			structure	19	56	23		8		1	7	
12	vWA	3			part of structure	10	29						28	
13	D13S317	3			LUS	9	25	3		2		19	15	
15	PentaE	1			structure	16	17	4		1			1	
16	D16S539	1			structure	8	9	4		1				
17	D17S1301	1			structure	9	10			1				
18	D18S51	2	1	1	LUS	16	20			3		3	2	
19	D19S433	1				18	22	3		3				
20	D20S482	1			structure	8	8	7						
21	D21S11	3	2	4		26	72	1	3	4				3
21	PentaD	2	2	8	part of structure	19	21			2	2	3	1	
22	D22S1045	1	1	3	structure	9	9	1						
X	DXS10074	3	1	5	LUS	16	29	5		24		24	34	
X	DXS10103	2	1	4	part of structure	14	32					14	32	
X	DXS10135	1				32	81	6	26	36			54	
X	DXS7132	1			structure	7	15			1		2		
X	DXS7423	1	1	8	stretches	7	7					1		
X	DXS8378	1			structure	7	9	5		2			1	
X	HPRTB	1				9	9							
Y	DYF387S1	4	3	4	part of structure	17	49			8		1	49	
Y	DYS19	2	4	6	stretches	7	7	2						
Y	DYS385a-b	4	3	14	LUS	14	18			1	18	4	18	
Y	DYS389I	3	1	4	part of structure	6	7						7	
Y	DYS389II-I	2				6	15			1			2	
Y	DYS389II					8	39	1		2	40		40	
Y	DYS390	3			part of structure	7	18						15	
Y	DYS391	3	1	4	LUS	6	9					3	9	

(*continued on next page*)

Table 4 (*continued*)

Locus		Reference sequence				# alleles			Naming (ignoring prefix/suffix)			
Y	DYS392	1			structure	8	8	1			1	
Y	DYS437	2			structure	4	8	2				10
Y	DYS438	1			structure	8	11	4		2		1
Y	DYS439	1			structure	7	7				1	
Y	DYS448	3	1	6		9	22	5		2		27
Y	DYS460	1			structure	7	7					
Y	DYS461	1				7	7					
Y	DYS461+DYS460					7	22		22			
Y	DYS481	1			structure	12	20	3		1	1	
Y	DYS505	1			structure	7	7					
Y	DYS522	1			structure	6	7	1			1	
Y	DYS533	1			structure	6	7			1		
Y	DYS549	1			structure	6	6					
Y	DYS570	1			structure	11	15	1		1		
Y	DYS576	1			structure	9	9	3				
Y	DYS612	2	2	3	part of structure	11	15	3				17
Y	DYS635	2			structure	11	22	1		1	1	1
Y	DYS643	1			structure	7	7					
Y	Y-GATA-H4	2	1	4	LUS	4	4					

[1]Chr, Chromosome. [2]LUS, Longest Uninterrupted Stretch.

Table 8, this intervening sequence is abbreviated to a pair of brackets. Sequence variation is represented as variant calls (similar to how this is done for variants in the prefix and suffix sequences), which are placed within these brackets. In allele names of loci that contain an interruption of more than 8 nucleotides in the reference sequence, the largest interruption (if more than 8 nucleotides) is abbreviated similarly. A single CE allele number is included in the name, but it will be based on the length of the entire target region as it would be in a CE-based profile. The CE allele numbers of the reference sequences of DYS460 and DYS461 are 10 and 12, respectively. Therefore, the reference CE allele number for the combined target is 22. Note that the CE allele numbers of DYS461 alleles are LUS plus one because it was historically defined as 'TCTG [TCTA]n'. As a result, the CE allele numbers in the combined names are one higher than the combined number of repeats.

When a smaller reporting region which excludes the DYS461 locus is configured, such as when using the Flanking Region Report from UAS, sequence variation in the region between the loci will be reported by STRNaming as prefix variants instead (DYS461 is located 5′ of DYS460).

Similarly, the range used for DXS10103 in the UAS flanking region report includes an unnamed AC repeat 29 nucleotides 5′ of the main repeat structure of this marker, which is recognised by STRNaming as a separate repeat structure. Another example is a small structure of CTAT repeats residing 21 nucleotides 5′ of the main Y-GATA-H4 repeat structure.

3.7. Naming a tandemly duplicated locus: DYS389

Locus DYS389 is interesting because it carries a duplication. Two different fragments are PCR-amplified with the same primer set (the reverse primer binds at two locations; the forward primer at one), with the shorter of the two (DYS389I) being a part of the longer (DYS389II) fragment. The reference sequence of the two duplicates of this locus are analysed separately by STRNaming. The two fragments are treated as different target regions. The DYS389II fragment has an interruption of 47 nucleotides; it is in this interruption that the reverse primer binds for the amplification of the DYS389I fragment. As can be seen in Table 9, the 3′ end of the STR structure of the DYS389I fragment is identical to the 5′ end of the reported DYS389II fragment. Note that the UAS flanking region report does not include the 5′ end of the STR structure of DYS389I in the reported range of the longer fragment. In fact, the reported range of the DYS389II fragment starts at the first 'T' in the 'ATAG' repeat of DYS389I. To keep naming consistent, STRNaming automatically filled in the missing 'A' in this example.

3.8. Interpretation of artefacts with STRNaming

STRNaming does not discern between genuine alleles and artefacts such as stutter and hybrid PCR products [13,14]. In general, artefact sequences are similar to the ancestral allelic sequences be it upon polymerase slippage (stutter product formation) or template switching (hybrid formation). Fig. 4 shows that as a result, allele names for artefacts are similar to the names of ancestral alleles as well.

This property enables intuitive interpretation of the obtained profile. It becomes clear from the number of reads that two names correspond to the alleles of a major contributor (with CE numbers 8 and 10). The sequence-informative naming allows to deduce that the other sequences are likely a variety of artefacts. From top to bottom in Fig. 4: stutter in the CTAT repeat (CE 8 to 7); allele (CE 8); stutter in the A stretch (CE 8 to 8.1); stutter in the CTAT repeat (CE 10 to 9); allele (CE 10); hybrid PCR product of the two alleles (CE 10 allele with the suffix of the CE 8 allele).

3.9. Bracket notation

The allele name representation presented here, with brackets around the repeat counts, is just one of several options to write STR alleles in a concise format. Results from the questionnaire (Supplementary File 1,

Table 5
Allele names on locus D1S1656. Target region coordinates: Chr1:230.769.561-695.

CE15_AC[6]CTAT[14]
CE15_AC[5]AT[1]CTAT[14]
CE15_AC[6]TTAT[1]CTAT[13]
CE15.3_AC[6]CTAT[11]CAT[1]CTAT[3]_+9C>T
CE15.3_AC[6]CTAT[12]CAT[1]CTAT[2]_+9C>T
CE15.3_AC[6]CTAT[13]CAT[1]CTAT[1]
CE15.3_AC[6]CTAT[9]CAT[1]CTAT[5]
CE16_AC[6]CTAT[11]CTAC[1]CTAT[3]

Table 6
Allele names on locus D13S317. Coordinates: Chr13:82.147.986–82.148.107.

CE11_TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3]
CE11_TATC[10]AATC[3]ATCT[3]
CE11_TATC[11]AATC[2]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]_-24G>A
CE11_TATC[12]AATC[1]ATCT[3]_-25C>T
CE11_TATC[13]ATCT[3]
CE12_TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]
CE12_TATC[12]AATC[2]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
CE12_TATC[13]AATC[1]ATCT[3]_-25C>T
CE12_TATC[13]AATC[2]ATCT[2]
CE12_TATC[14]ATCT[3]
CE13_TATC[13]AATC[2]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]_-24G>A
CE13_TATC[14]AATC[1]ATCT[3]_-25C>T
CE13_TATC[15]AATC[1]ATCT[3]_+9GTCT>-
CE13_TATC[15]ATCT[3]

Question 13) indicate a preference among respondents toward placing the brackets around the repeat unit sequence instead. However, when presented with six different ways of abbreviating long interruptions (Question 14), a majority of respondents voted for one of the three options with the brackets around the repeat counts. Clearly the preferred notation is subject to personal taste. To our opinion the different notations may continue to coexist, as they are fully compatible when the STRNaming algorithm is used to determine the repeat structure.

3.10. Algorithm implementation details

To find the optimal combination of repeat stretches, the scores in Table 2 need to be applied to a potentially very large number of possibilities, which would make the algorithm impractically slow for the more complex STR loci. Therefore, STRNaming was coded in a way that minimizes the number of combinations actually assessed. A number of optimizations have been implemented.

- Short repeats that are embedded within longer repeats are discarded. For example, in the hexanucleotide repeat AGAGAT of DYS448, all the AG[2] repeats are immediately discarded. This optimization affects all loci that include repeat units of different lengths. During initial analysis of the reference sequence, almost all loci are affected because short dinucleotide repeats appear very commonly in the reference sequence and unrepeated instances of units repeated elsewhere are also included in the analysis.
- The reference sequences of the prefix and suffix are aligned to the 5′ and 3′ ends of the reported allele before evaluating combinations of repeats. When the alignment results in a positive score, the STR structure is required to start/end exactly at the aligned position. All other starting and/or ending positions are discarded.
- Consecutive repeat stretches are combined into longer uninterrupted structures. Two lists are included with each structure. The first list, ‘anchors’, contains repeat units used in their original location (or used when naming the reference sequence). The second list, ‘orphans’, contains repeat units used in a different location.
- Lookup tables are used to quickly identify whether a given range of sequence can be spanned while allowing only a limited number of interruptions. The tables include a list of repeat units that can be

Table 7
Allele names on locus D21S11. While most allele names share the same structure (**A**), STRNaming shifts to a different structure for most x.2 alleles due to an extra TA 5′ of the penultimate TATC repeat unit (**C**, extra TA underlined). A small number of alleles do not perfectly fit either category (**B**). Coordinates: Chr21:19.181.939–19.182.111.

A	CE24.2_TCTA[5]TCTG[6]TCTA[3]TCA[1]TCTA[2]TCCA[1]TATC[10]
	CE26_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[9]
	CE28_TCTA[5]TCTG[5]TCTA[3]TATC[3]A[1]TCTA[2]TCCA[1]TATC[12]
	CE29_TCTA[4]TCTG[7]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]
	CE29.3_TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[8]ATC[1]TATC[3]
	CE30_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]
	CE30_TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[2]TACC[1]TATC[10]
	CE33_TCTA[6]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[15]
	CE34_TCTA[11]TCTG[5]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[11]
	CE35_TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]
	CE35.1_TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]T[1]
	CE36_TCTA[11]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]
	CE36.1_TCTA[10]TCTG[7]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[13]T[1]
	CE38_TCTA[10]TCTG[10]TCTA[2]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]
B	CE30.3_TCTA[4]TCTG[6]TCTA[3]TATC[4]ATCT[2]ATCA[1]TCTA[2]TCCA[1]TATC[11]
	CE35_TCTA[5]TCTG[6]TCTA[3]TATC[4]ATCT[2]ATCCA[1]TATC[9]ATCT[3]ATCA[1]TCTA[2]TATC[2]
	CE36_TCTA[5]TCTG[6]TCTA[3]TATC[4]ATCT[2]ATCCA[1]TATC[10]ATCT[3]ATCA[1]TCTA[2]TATC[2]
	CE34.1_TCTA[5]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[6]A[1]TCTA[8]TATC[2]
C	CE28.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]ATC[1]TATC[2]CATA[1]TCTA[8]TATC[2]
	CE30.2_TCTA[4]TCTG[6]TCTA[3]TATC[4]ATC[1]TATC[2]CATA[1]TCTA[11]TATC[2]
	CE32.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]ATC[1]TATC[2]CATA[1]TCTA[12]TATC[2]_+1G>A
	CE33.2_TCTA[5]TCTG[6]TCTA[3]TATC[4]ATC[1]TATC[2]CATA[1]TCTA[3]CCTA[1]TCTA[9]TATC[2]
	CE34.2_TCTA[5]TCTG[7]TCTA[3]TATC[4]ATC[1]TATC[2]CATA[1]TCTA[13]TATC[2]

Table 8
Allele names on loci DYS460 and DYS461. When the two loci are sequenced in a single fragment, STRNaming is able to generate a combined name. Coordinates of the combined target region used here are chrY:18.888.802–18.889.046.

DYS461+DYS460	DYS461	DYS460
CE19_TCTA[11][]TATC[7]	CE12_TCTA[11]	CE7_TATC[7]
CE20_TCTA[8][]TATC[11]	CE9_TCTA[8]	CE11_TATC[11]
CE20_TCTA[9][]TATC[10]	CE10_TCTA[9]	CE10_TATC[10]
CE20_TCTA[10][]TATC[9]	CE11_TCTA[10]	CE9_TATC[9]
CE20_TCTA[11][]TATC[8]	CE12_TCTA[11]	CE8_TATC[8]
CE21_TCTA[9][]TATC[11]	CE10_TCTA[9]	CE11_TATC[11]
CE21_TCTA[10][]TATC[10]	CE11_TCTA[10]	CE10_TATC[10]
CE21_TCTA[11][]TATC[9]	CE12_TCTA[11]	CE9_TATC[9]
CE21_TCTA[12][]TATC[8]	CE13_TCTA[12]	CE8_TATC[8]
CE22_TCTA[10][]TATC[11]	CE11_TCTA[10]	CE11_TATC[11]
CE22_TCTA[11][]TATC[10]	CE12_TCTA[11]	CE10_TATC[10]
CE22_TCTA[12][]TATC[9]	CE13_TCTA[12]	CE9_TATC[9]
CE23_TCTA[10][]TATC[12]	CE11_TCTA[10]	CE12_TATC[12]
CE23_TCTA[11][]TATC[11]	CE12_TCTA[11]	CE11_TATC[11]
CE23_TCTA[12][]TATC[10]	CE13_TCTA[12]	CE10_TATC[10]
CE24_TCTA[11][]TATC[12]	CE12_TCTA[11]	CE12_TATC[12]
CE24_TCTA[12][]TATC[11]	CE13_TCTA[12]	CE11_TATC[11]
CE24_TCTA[13][]TATC[10]	CE14_TCTA[13]	CE10_TATC[10]
CE25_TCTA[11][]TATC[13]	CE12_TCTA[11]	CE13_TATC[13]
CE25_TCTA[12][]TATC[12]	CE13_TCTA[12]	CE12_TATC[12]
CE25_TCTA[13][]TATC[11]	CE14_TCTA[13]	CE11_TATC[11]
CE25_TCTA[14][]TATC[10]	CE15_TCTA[14]	CE10_TATC[10]

Table 9
Allele names on locus DYS389, using UAS flanking region report ranges. The large, 47-nucleotide interruption between the two STR structures is represented by [] in DYS389II. The 3′ part of the shorter fragment (DYS389I; the 3′ part starts after the AGGG repeats indicated in green) overlaps with the 5′ part of the longer fragment (DYS389II; the 5′ part is the region before the interruption indicated as []), and the shared region is named the exact same. Coordinates: DYS389I chrY:12.500.387–513; DYS389II chrY:12.500.448–633.

DYS389I	DYS389II
CE12_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[9]ACAG[3]	CE27_ATAG[9]ACAG[3][]GATA[10]GACA[6]
CE12_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[9]ACAG[3]	CE28_ATAG[9]ACAG[3][]GATA[11]GACA[6]
CE12_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[9]ACAG[3]	CE29_ATAG[9]ACAG[3][]GATA[11]GACA[7]_+3C>T
CE12_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[9]ACAG[3]	CE29_ATAG[9]ACAG[3][]GATA[12]GACA[6]
CE12_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[9]ACAG[3]	CE30_ATAG[9]ACAG[3][]GATA[13]GACA[6]
CE13_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[10]ACAG[3]	CE28_ATAG[10]ACAG[3][]GATA[10]GACA[6]
CE13_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[10]ACAG[3]	CE29_ATAG[10]ACAG[3][]GATA[11]GACA[6]
CE13_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[10]ACAG[3]	CE30_ATAG[10]ACAG[3][]GATA[11]GACA[7]
CE13_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[10]ACAG[3]	CE30_ATAG[10]ACAG[3][]GATA[12]GACA[6]
CE13_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[10]ACAG[3]	CE31_ATAG[10]ACAG[3][]GATA[13]GACA[6]
CE14_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[11]ACAG[3]	CE30_ATAG[11]ACAG[3][]GATA[11]GACA[6]
CE14_ATAG[2]ATTG[1]ATAG[1]AGGG[2]ATAG[11]ACAG[3]	CE31_ATAG[11]ACAG[3][]GATA[12]GACA[6]

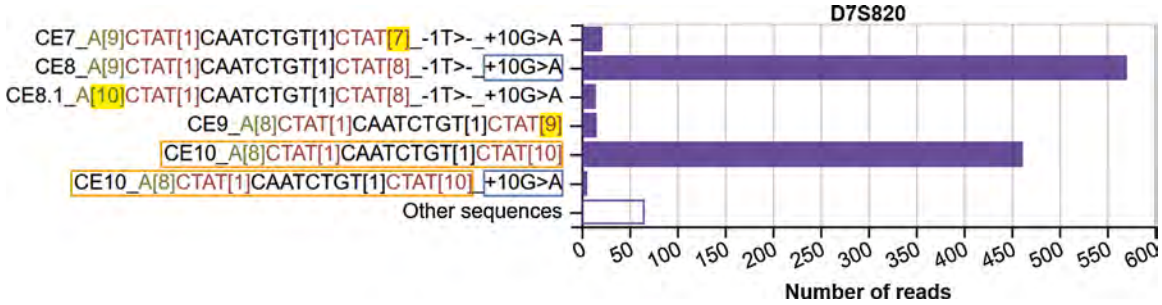


Fig. 4. Example locus of a singular MPS-STR profile with allele names obtained from STRNaming. Besides the two genuine alleles (long vertical bars; CE numbers 8 and 10), four PCR artefacts are visible: three stutter products (variation highlighted in yellow) and one PCR hybrid (CE 10 repeat structure with suffix of the CE 8 allele; as outlined). Reported range: Chr7:84160191–84160297.

Table 10
Examples of D13S317 allele names and corresponding identifiers as generated by STRNaming and SID. SID identifiers truncated to three characters are given for the entire target region (Chr13:82.147.986–82.148.107) and for only the region of the repeat structure, as defined by STRNaming (Chr13:82.148.025–82.148.088). Grey shaded alleles have the same SID name when the region of repeat structure is regarded but a different name when the full target region is used.

SID (target region)	SID (region of repeat structure)	Allele name
JGT	HBY	CE12_TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]
BQR	GBN	CE12_TATC[12]AATC[2]ATCT[3]
XDQ	LJH	CE12_TATC[13]AATC[1]ATCT[3]
RBC	LJH	CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
EZQ	LJH	CE12_TATC[13]AATC[1]ATCT[3]_-25C>T
SXN	UHE	CE12_TATC[13]AATC[2]ATCT[2]
FGW	WZA	CE12_TATC[14]ATCT[3]
LUX	TTK	CE13_TATC[13]AATC[2]ATCT[3]
CCL	DNN	CE13_TATC[14]AATC[1]ATCT[3]
NVW	DNN	CE13_TATC[14]AATC[1]ATCT[3]_-24G>A
OFG	DNN	CE13_TATC[14]AATC[1]ATCT[3]_-25C>T
BTC	AGD	CE13_TATC[15]ATCT[3]
LSM	YIR	CE13_TATC[15]AATC[1]ATCT[3]_+9GTCT>-
OBF	YIR	CE14_TATC[15]AATC[1]ATCT[3]
QIR	BKT	CE14_TATC[14]AATC[2]ATCT[3]

'anchored' within that range. During the construction of combinations of repeat stretches, these lookup tables allow to immediately recognize situations in which it is not possible to reach the 5′ end of the suffix whilst 'anchoring' all 'orphan' repeat units.

- For many markers, the full list of combinations of repeat stretches would not fit in computer memory. Construction of a list is avoided by combining the programming techniques generators and recursion. This means that only two full combinations of repeat stretches 'exist' in computer memory at any time – the one for which the score is being calculated, and the one corresponding to the highest score thus far.

Together, these optimizations enable nearly instant naming. For most markers and most alleles, only one or two possible names are examined. This includes markers with relatively complex names, such as D13S317 and vWA.

3.11. Comparison to existing nomenclature

In Supplementary Table 2, the repeat stretches as used by STRNaming are compared to the most recent version (v5) of the manually-curated Forensic STR Sequence Structure Guide available from STRidER [10]. While all reference sequences in the STR Structure Guide are in the forward orientation of the GRCh38 reference genome [15], which is also used by STRNaming, many of the STR structure definitions were originally based on reference sequences orientated in the reverse complementary direction. Unsurprisingly, many of these STRs shift to a different repeat unit when using the STRNaming definitions. For 18 of the 60 STRs compared here, STRNaming includes additional repeat stretches besides those defined by the Structure Guide. In three markers, some repeat stretches defined by the Structure Guide were excluded by STRNaming.

4. Concluding remarks

From the reference sequence for an STR locus, STRNaming derives a unique, sequence-informative name for any given sequence for that locus in a fully automated manner. The name can be read and interpreted by human eye.

The reference sequence includes relatively large flanking regions to maintain consistent results while accommodating different primer placements, but in the process of naming alleles (Section 2.4) the target region is extracted. This target region (5′ end of prefix to 3′ end of suffix) represents the amplified fragment excluding the primer sequences. Since primers of different PCR kits can bind at slightly different positions (especially when different manufacturers are involved) the target region for a locus may vary with the PCR kit. STRNaming accepts flexibility concerning the target region, but precise definition of the target region is important as it determines which flanking-site SNPs are included in the prefix and suffix regions of the name given by STRNaming. For example in D12S391, the rs138635218 variant is represented by '+85C>G', but it would be undetectable and thus omitted from the name if the target region extends fewer than 85 nucleotides past the 3′ end of the repeat structure. Therefore, the (genomic) positions of the target region should always be communicated if the raw sequences are not provided, especially when data from different PCR kits is combined. When naming and comparing sequences obtained by different kits, we recommend to trim the sequences down to the range both kits have in common. Regarding storage of sequences in allele frequency databases and national DNA databases, we recommend to store untrimmed sequences and corresponding genomic positions and proceed to trimming of the sequences when queried for a kit with a shorter range. This approach achieves maximum compatibility between sequencing data obtained through various PCR kits. The name given by STRNaming also includes the CE allele number to provide compatibility with existing STR profiling. Note that when sequence variants with the same CE allele number are merged, a CE frequency database is derived from an MPS frequency database.

In the online version of STRNaming, the SID [5] algorithm was introduced successfully to achieve the option to also generate a short code identifier. Because the SID is highly dependent on the target region, obtaining a 'universal' SID requires defining a (smaller) common target region. As exemplified in Table 10, the repeat structure defined by STRNaming (i.e., exclusion of the prefix and suffix) provides an intuitive common target region so that functional SID and STRNaming names are achieved.

STRNaming will be included in the next version of FDSTools [16], leading to great simplification of the library file needed for configuring allele naming, as only the target regions will need to be provided. It will also be included in a future update to DNAxs [2]. An on-line version of the tool is available at fdstools.nl. In addition, the source code of the algorithm (in Python 3 and ECMAScript 9) is available under an open-source license, enabling other bioinformatics software to implement the same method. STRNaming was shown to work well with various types of STR loci in current use, including X-STRs and (rapidly mutating) Y-STRs [22,23], and it is prepared to be readily applied to

new loci introduced in the future.

Acknowledgements

We thank the Forensic Laboratory for DNA Research (FLDO, LUMC, Leiden), King's College London, the STRAND Working Group and the respondents of the questionnaire for their valuable input. We are very grateful to Prof Peter de Knijff (FLDO) and Dr Laura Heathfield (University of Cape Town) for providing datasets from diverse populations, which were used to train and evaluate the performance of STRNaming. Brian Young (NicheVision) is thanked for sharing the code to generate SID nomenclature.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2021.102473.

References

[1] B. Bruijns, R. Tiggelaar, H. Gardeniers, Massively parallel sequencing techniques for forensics: a review, Electrophoresis 39 (2018) 2642–2654.
[2] C.C.G. Benschop, J. Hoogenboom, et al., DNAxs/DNAStatistX: development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles, Forensic Sci. Int. Genet. 42 (2019) 81–89.
[3] S. Willuweit, in: ChaLlenges and Paradigm Shifts by the Adoption of MPS in Forensic Casework. Conference, HIDS, 2017 researchgate.net/publication/318421173.
[4] C. van Neste, et al., Forensic Loci Allele Database (FLAD): automatically generated, permanent identifiers for sequenced forensic alleles, Forensic Sci. Int. Genet. 20 (2015) E1–E3.
[5] B. Young, T. Faris, L. Armogida, A nomenclature for sequence-based forensic DNA analysis, Forensic Sci. Int. Genet. 42 (2019) 14–20.
[6] C. Gelardi, et al., Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.
[7] K.B. Gettings, et al., STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.
[8] K. van der Gaag, P. de Knijff, Forensic nomenclature for short tandem repeats updated for sequencing, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e542–e544.
[9] W. Parson, et al., Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63.
[10] C. Phillips, et al., "The devil's in the detail": release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int. Genet. 34 (2018) 162–169.
[11] K.B. Gettings, et al., STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.
[12] C. Phillips, et al., Global patterns of STR sequence variation: sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit, Electrophoresis 39 (2018) 2708–2724.
[13] A.R. Shuldiner, A. Nirula, J. Roth, Hybrid DNA artifact from PCR of closely related target sequences, Nucl. Acids Res. 17 (1989) 4409.
[14] H. Kim, H.A. Erlich, C.D. Calloway, Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions, Croat. Med. J. 56 (2015) 208–217.
[15] V.A. Schneider, et al., Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, Genome Res. 27 (2017) 849–864.
[16] J. Hoogenboom, K.J. van der Gaag, et al., FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci. Int. Genet. 27 (2017) 27–40.
[17] T. Kraaijenbrink, et al., A linguistically informed autosomal STR survey of human populations residing in the greater himalayan region, PLoS One 9 (3) (2014), e91534.
[18] P. de Knijff, J. Pijpe, Population Genetics of African Pygmies, 2015. . Unpublished work.
[19] A. Westen, et al., Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.
[20] Illumina, ForenSeq™ Universal Analysis Software Guide, Document #VD2018007, June, 2018.
[21] A. Theophilus, et al., A novel exoplanetary habitability score via particle swarm optimization of CES production functions, IEEE Symposium Series on Computational Intelligence (2018) 2139–2147.
[22] K.N. Ballantyne, et al., Toward male individualization with rapidly mutating Y-Chromosomal short tandem repeats, Hum. Mutat. 35 (2014) 1021–1032.
[23] R. Alghafri, et al., A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-STRs, Forensic Sci. Int. Genet. 17 (2015) 91–98.