MixMaSTR: a Software Package for Designing and Interpreting Forensic DNA Validation Studies

Sarah Riman
Applied Genetics Group
NIST Continuing Education Day

Acknowledgment

This work was supported by the Accelerating Forensic Innovation for Impact through the NIST Special Programs Office: *Forensic Genetics*.

****Hari Iyer (SED)**
****Sarra Chouder (ISG)**
Edgar Robitaille (JHU)
Sicen Liu (JHU)
Asmitha Sathya (JHU)
Benjamin J. Long (ISG)
Peter M. Vallone (AGG)

U.S. Forensic Laboratories:

- *Kyle Duke and* Jeanette Wallin (*CalDoJ)*
- Kristy Kadash (*JCRCL)*
- Kaitlin Huffman (*FBI)*
- Michelle Madrid (*LACSD)*
- Toni Diegoli (*ATF)*

sarah.riman@nist.gov
mixmastr@nist.gov

Disclaimer

- Points of view in this presentation are those of mine and do not necessarily represent the official position or policies of the U.S. Department of Commerce.

- Certain commercial software, 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 NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

- All work involving NIST samples has been reviewed and approved by the NIST Research Protections Office.

Forensic DNA Validation Studies

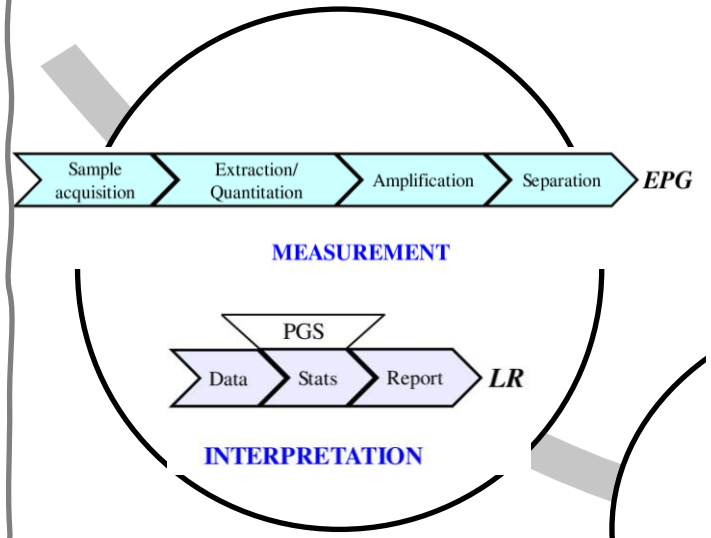
- ***Any established protocol*** that will be used by a forensic lab to measure/interpret a sample from a crime scene ***should be supported by validation studies to assess its performance***.

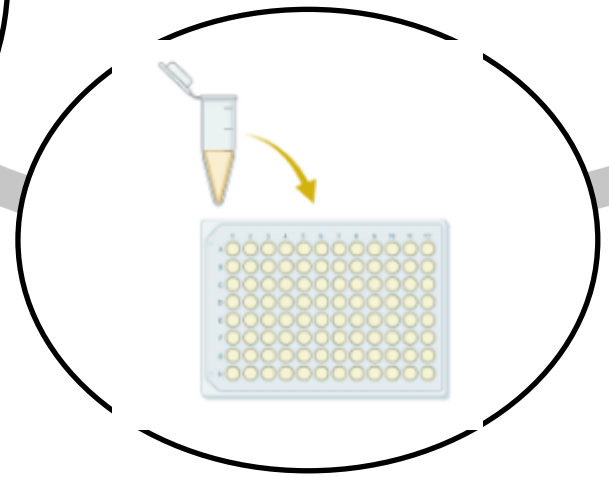
- Conducted with samples of ***known origin (ground-truth)*** that ***reasonably*** ***cover the factor space of the samples that are routinely accepted and tested by a lab***.

- ***Not a one-time process*** as laboratories continue to revalidate when changes or upgrades are introduced to their workflows.

Internal Validation Workflow

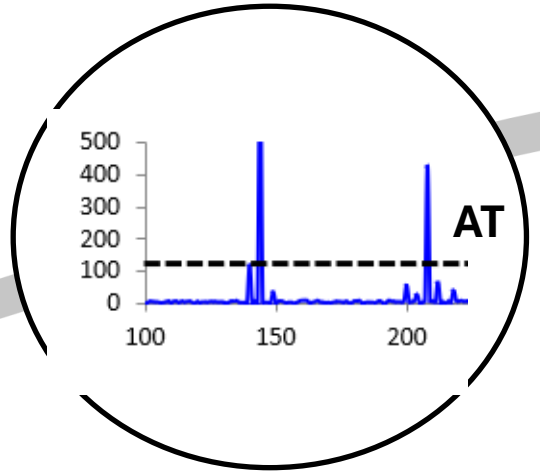
Define the entire pipeline (protocol)

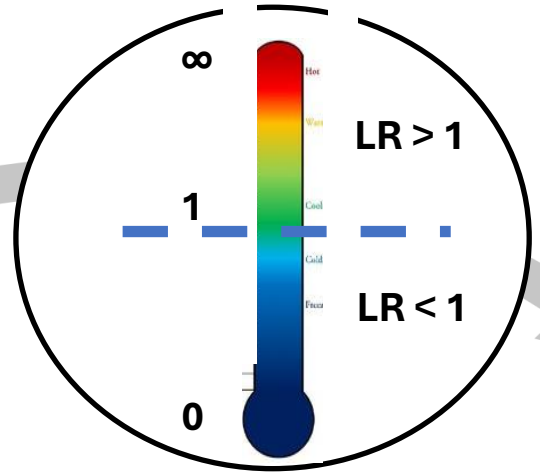




Choose and prepare validation experimental design

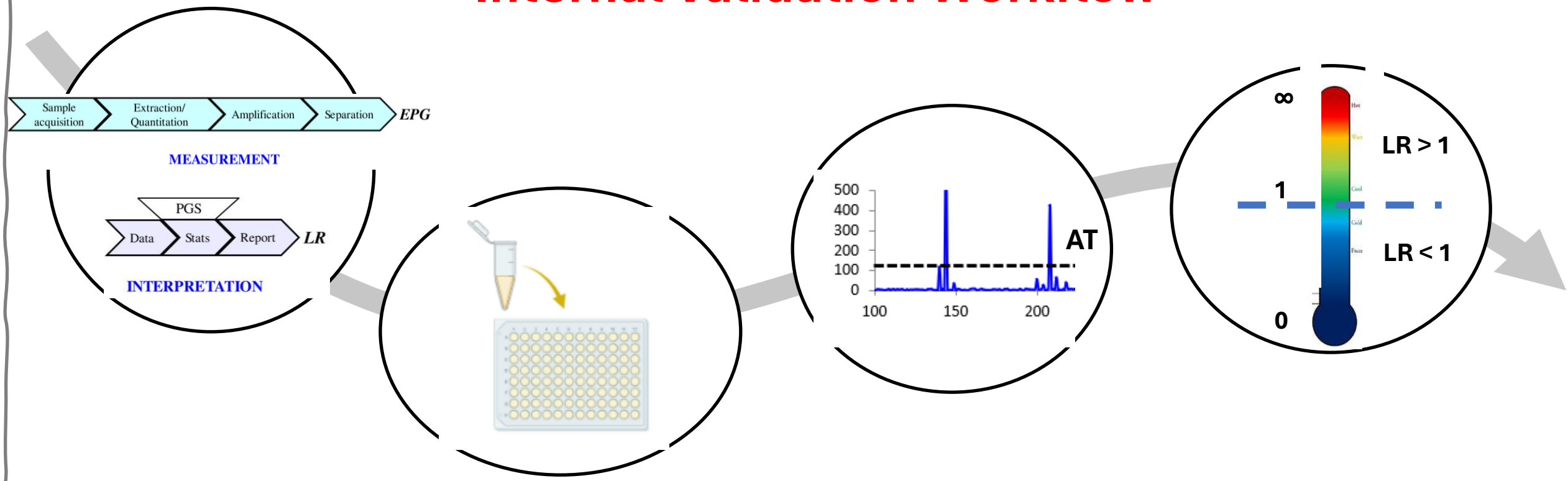
Run samples using the defined pipeline and determine an Analytical Threshold (AT)





Interpret data

Internal Validation Workflow



- This workflow generates large amounts of data.
- Forensic practitioners face the challenges of designing validation experiments, post-processing, analyzing, and understanding the validation data.
- The field lacks an ***open-source software*** that can assist in these tasks.

MixMaSTR: A Software Package for Designing and Interpreting Forensic DNA Validation Studies

Developing a standalone and easy-to-use application with a well-designed graphical user interface that will help practitioners in:

1) ***Generating all possible mixture genotype combinations from their provided ground truth single-source profiles and selected NoCs***
2) ***Computing various metrics of interest for each constructed mixture combination***
3) ***Designing validation experiments that adequately cover a user selected factor space***
4) ***Providing an efficient strategy for preparing the desired mixtures (i.e., mixture calculations)***
5) ***Providing performance-based metrics to aid a laboratory in examining different AT methods and choosing the optimal one for casework they commonly encounter***
6) ***Calculating and reporting various metrics:*** *average peak heights; % of profile recovered; drop-out events; and drop-in events*
7) ***Visualizing the resulting data.***

The scope of the software design was guided by discussions on validation challenges with forensic practitioners from different laboratories.

A Software Package for Designing and Interpreting Forensic DNA Validation Studies

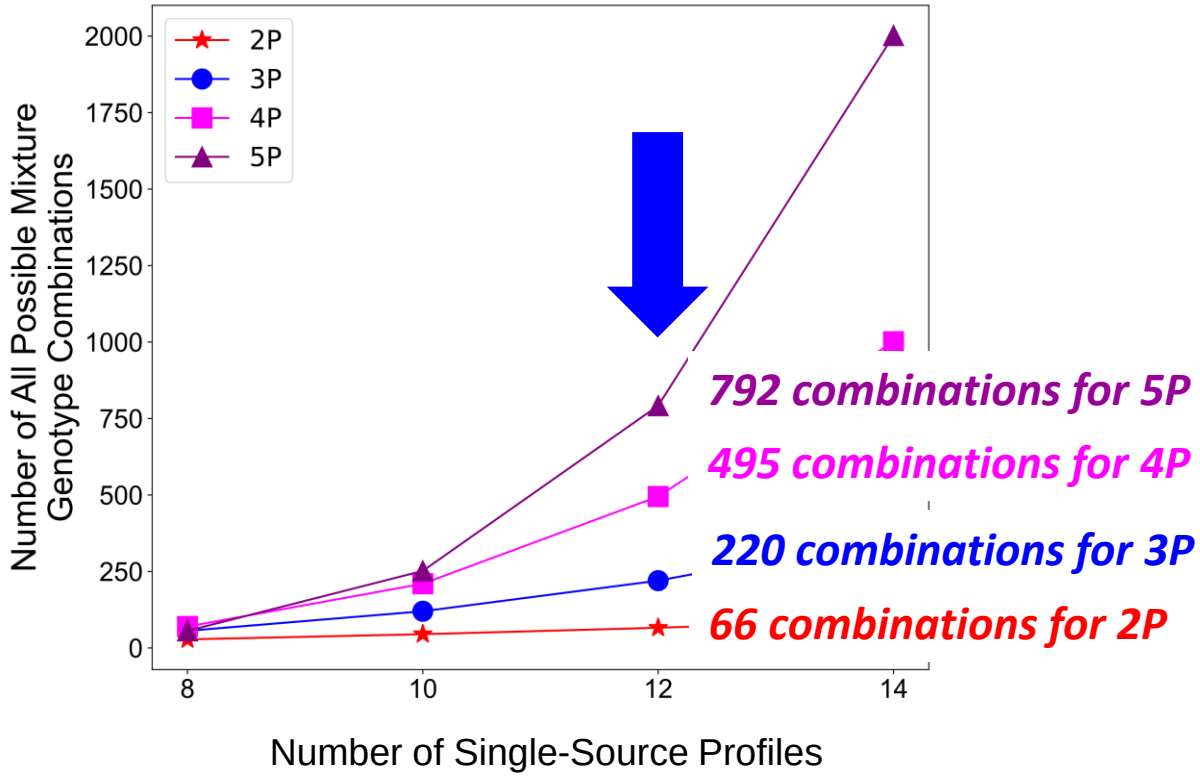
Developing a standalone and easy-to-use application with a well-designed graphical user interface that will help practitioners in:

1) ***Generating all possible mixture genotype combinations from their provided ground truth single-source profiles and selected NoCs***
2) *Computing various metrics of interest for each constructed mixture combination*
3) *Designing validation experiments that adequately cover a user selected factor space*
4) *Providing an efficient strategy for preparing the desired mixtures (i.e., mixture calculations)*
5) *Providing performance-based metrics to aid a laboratory in examining different AT methods and choosing the optimal one for casework they commonly encounter*
6) *Calculating and reporting various metrics: average peak heights; % of profile recovered; drop-out events; and drop-in events*
7) *Visualizing the resulting data.*

The scope of the software design was guided by discussions on validation challenges with forensic practitioners from different laboratories.

Software Key Features

1 Construction of all possible mixture genotype combinations



Mixture Genotype Combinations Combinatorial Formula

$$Combinations, {}_nC_r = \frac{n!}{r!(n-r)!}$$

${}_nC_r$ = number of possible genotype combinations
n = number of single-source profiles
r = number of contributors

= COMBIN(# of ss profiles, NoC)

- Supports loading user-provided ***ground truth single-source profiles*** genotyped by any STR multiplex kit.
- Generates all possible mixture genotype combinations depending on the ***number of single-source profiles and NoCs*** chosen.

A Software Package for Designing and Interpreting Forensic DNA Validation Studies

Developing a standalone and easy-to-use application with a well-designed graphical user interface that will help practitioners in:

1) ***Generating all possible mixture genotype combinations from their provided ground truth single-source profiles and selected NoCs***
2) ***Computing various metrics of interest for each constructed mixture combination***
3) *Designing validation experiments that adequately cover a user selected factor space*
4) *Providing an efficient strategy for preparing the desired mixtures (i.e., mixture calculations)*
5) *Providing performance-based metrics to aid a laboratory in examining different AT methods and choosing the optimal one for casework they commonly encounter*
6) *Calculating and reporting various metrics: average peak heights; % of profile recovered; drop-out events; and drop-in events*
7) *Visualizing the resulting data.*

The scope of the software design was guided by discussions on validation challenges with forensic practitioners from different laboratories.

Software Key Features

2

Computation of various metrics of interest for each constructed mixture combination

- **Allele Sharing Ratio (ASR)**
- **Counts of homozygote genotypes**
- **Instances of rare alleles**
- **Instances of allele-allele 1-bp difference**
- **Maximum allele count (MAC) → MAC NoC**

Software Key Features

❷ Computation of various statistic metrics of interest for each constructed mixture combination

- **Allele Sharing Ratio (ASR)**
- Counts of homozygote genotypes
- Instances of rare alleles
- Instances of allele-allele 1-bp resolution
- Maximum allele count (MAC) → MAC NoC

Allele Sharing Ratio (ASR)

$$ASR = 1 - \frac{(actual\ number\ of\ observed\ peaks\ - minimum\ possible\ number\ of\ peaks)}{(maximum\ possible\ number\ of\ peaks\ - minimum\ possible\ number\ of\ peaks)}$$

- A summary of the degree to which different contributors to a DNA mixture have overlapping alleles.

- ASR ranges between 0 and 1. Higher values indicate MORE shared alleles.

- Identical twins share all alleles; ASR = 1.

- The ASR for 2P mixtures involving brothers: **0.25 for one pair of brothers**
  0.78 for a second pair of brothers

- Even randomly chosen individuals will USUALLY share SOME alleles.
  E.g., The ASR for 2P mixtures for 1036 individuals in the NIST population database ranges between **0.0526** and **0.6857**

- Theoretically, it is possible that 2 individuals do not share any alleles; ASR = 0 (low probability).

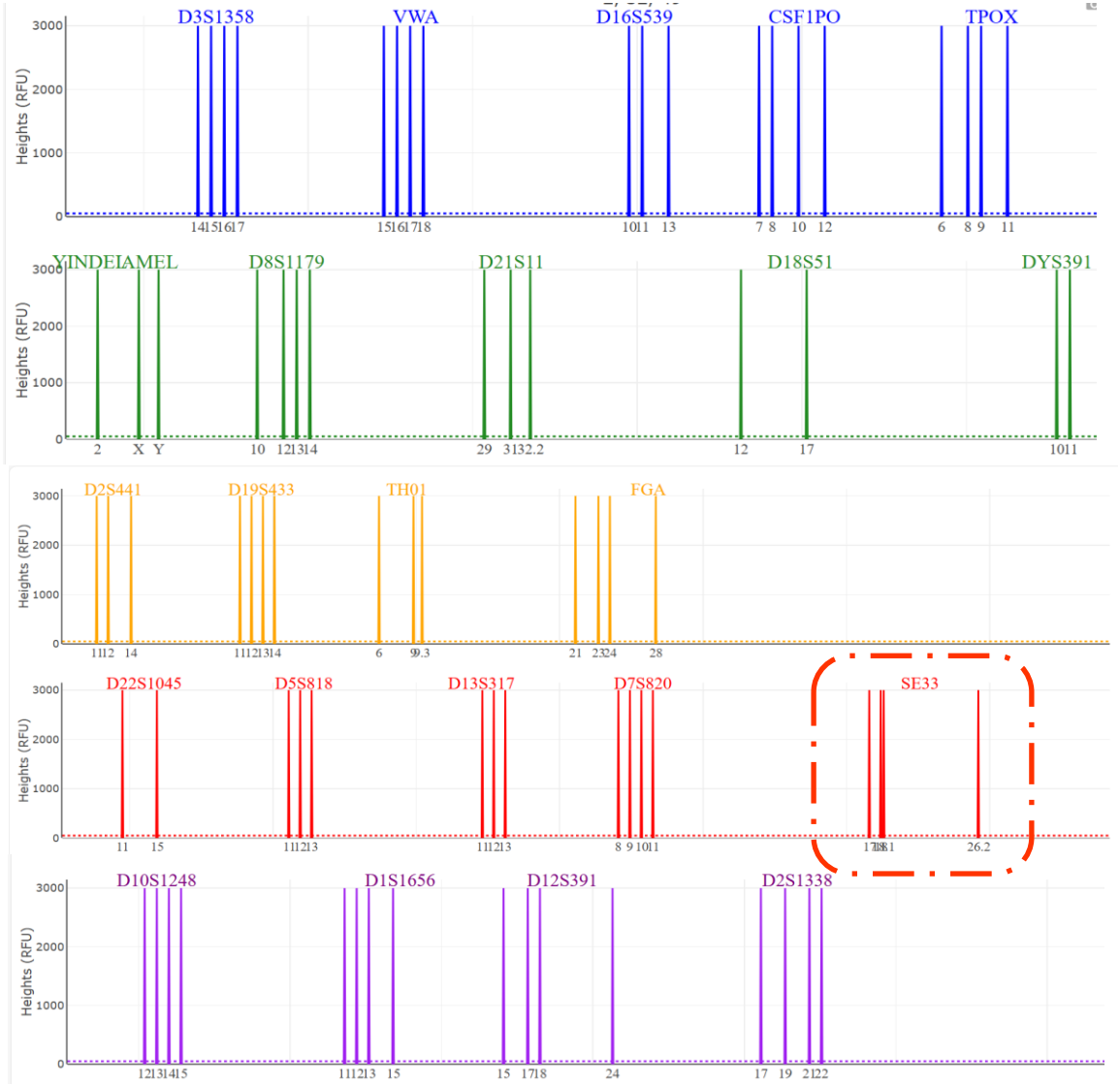
Software Key Features

❷ Computation of various statistic metrics of interest for each constructed mixture combination

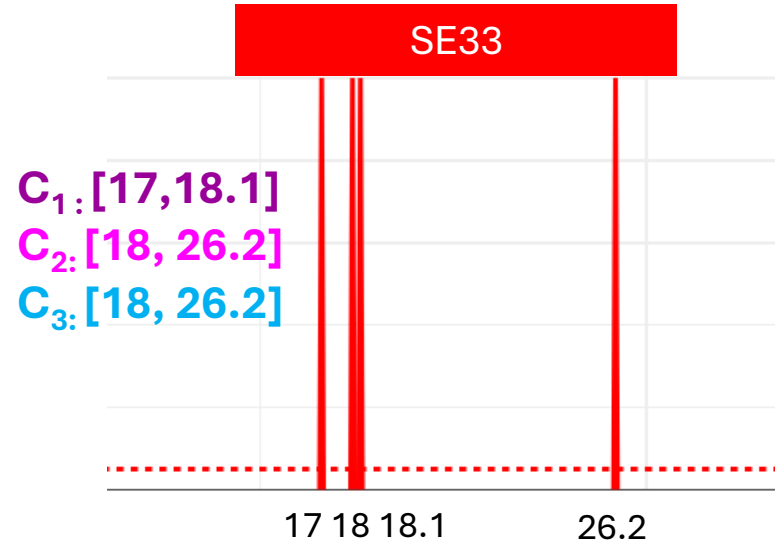
- **Allele Sharing Ratio (ASR)**
- **Counts of homozygote genotypes**
- **Instances of rare alleles**
- **Instances of allele-allele 1-bp resolution**
- **Maximum allele count (MAC) → MAC NoC**

Computation of various metrics

As an illustration, the mixture genotype combination simulated from the genotypes of three PROVEDIt single-source samples



Locus metrics

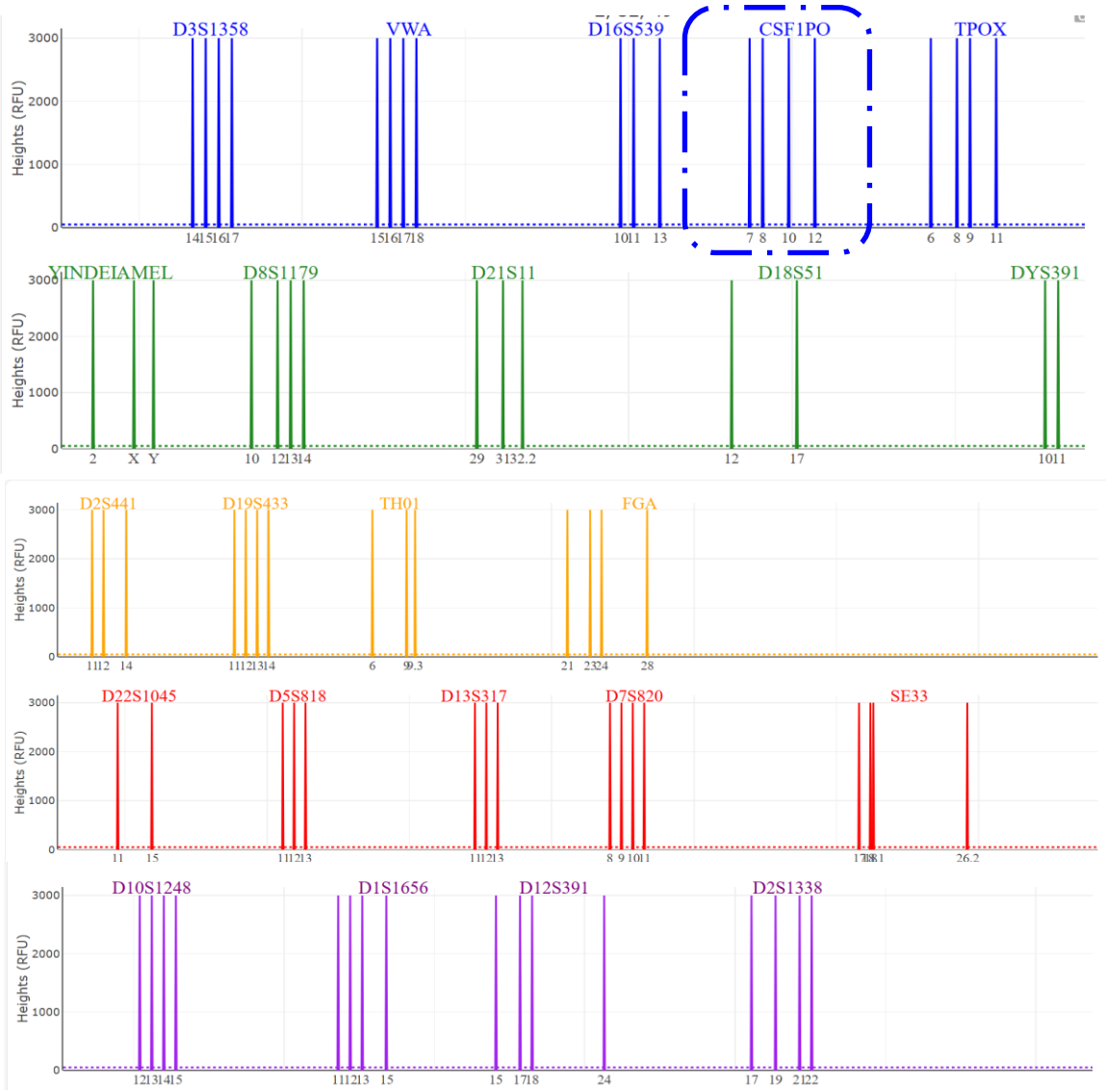


Expected SE33 Alleles :	[17, 18, 18.1, 26.2]
# of Alleles	4
Homozygosity	[] | 0
A-A 1 bp	C2 [18], C3 [18], C1 [18.1] | 1
Rare Alleles	[] | 0

This output is not intended to reflect peak height variation, just the presence of ground truth genotypes

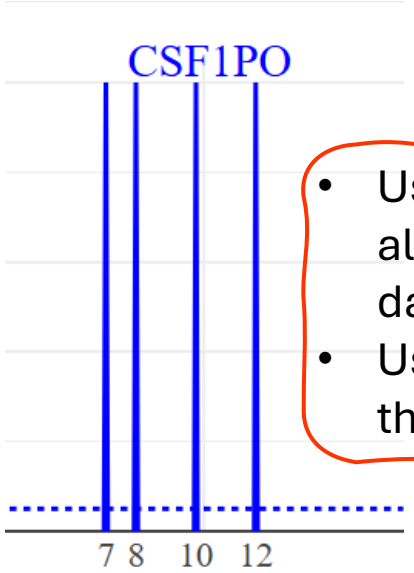
Computation of various metrics

As an illustration, the mixture genotype combination simulated from the genotypes of three PROVEDIt single-source samples



Locus metrics

C_1 : [7, 8]
C_2: [10, 12]
C_3: [10, 10]

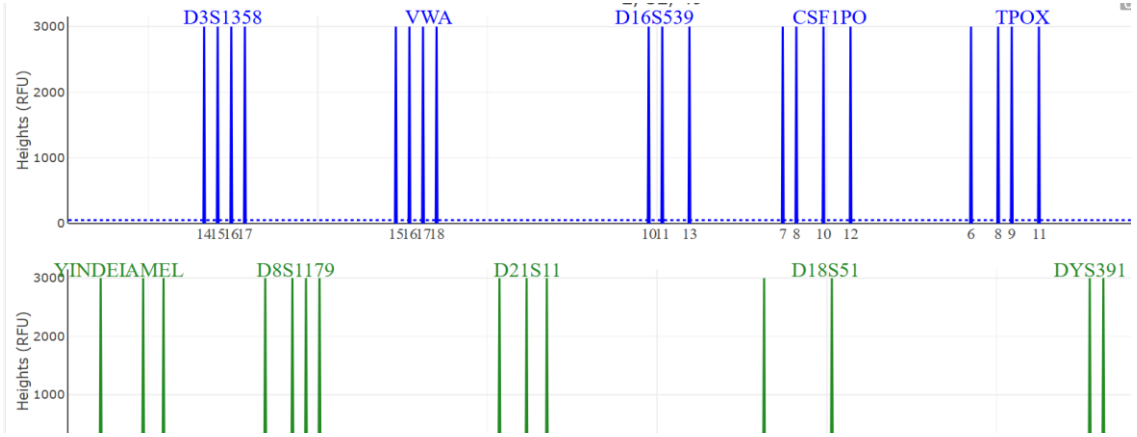


- User selected or uploaded allele frequency/population database.
- User selected rare allele threshold.

Expected CSF1PO Alleles :	[7, 8, 10, 12]
# of alleles	4
Homozygosity	C3 [10,10] | 1
A-A 1 bp	[] | 0
Rare Alleles	C1 [7] | 1

Computation of various metrics

As an illustration, the mixture genotype combination simulated from the genotypes of three PROVEDIt single-source samples



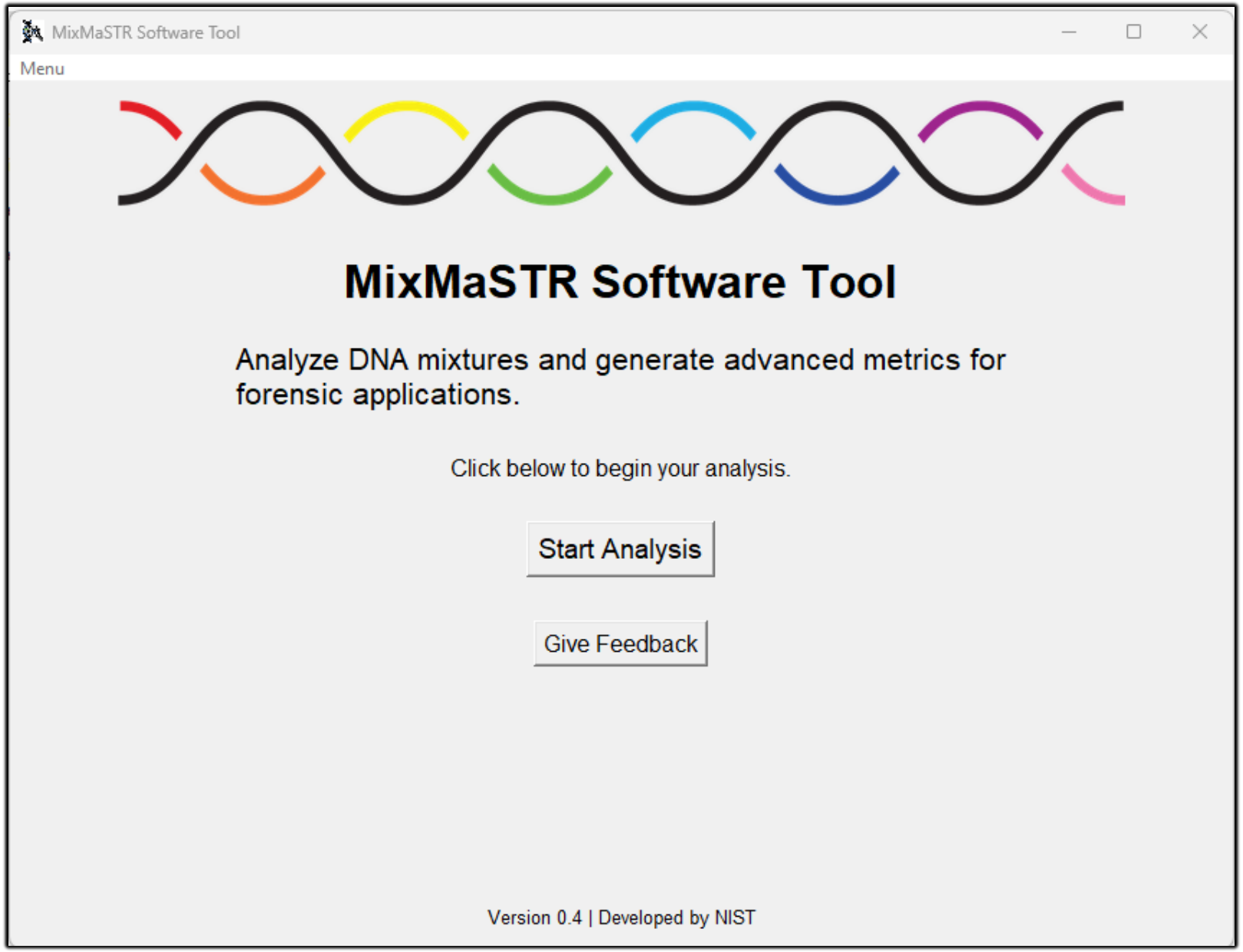
**** Profile metrics***

Summary statistics across each simulated mixture combination (N=21 loci)

Mixture Combinations	eNoC	MAC NoC	∑ Homozygote Genotypes	∑A-A 1 bp Difference	∑ Rare Alleles	ASR
C_1,C_2, C_3	3	2	13	1	1	0.56

ASR range
0.25 – 0.58

Software Demo



A Software Package for Designing and Interpreting Forensic DNA Validation Studies

Developing a standalone and easy-to-use application with a well-designed graphical user interface that will help practitioners in:

1) ***Generating all possible mixture genotype combinations from a user-provided ground truth single-source profiles and selected NoCs***
2) ***Computing various metrics of interest for each constructed mixture combination***
3) ***Designing validation experiments that adequately cover a user selected factor space***
4) ***Providing an efficient strategy for preparing the desired mixtures (i.e., mixture calculations)***
5) ***Providing performance-based metrics to aid a laboratory in examining different AT methods and choosing the optimal one for casework they commonly encounter***
6) ***Calculating and reporting various metrics:*** *average peak heights; % of profile recovered; drop-out events; and drop-in events*
7) ***Visualizing the resulting data.***

The scope of the software design was guided by discussions on validation challenges with forensic practitioners from different laboratories.

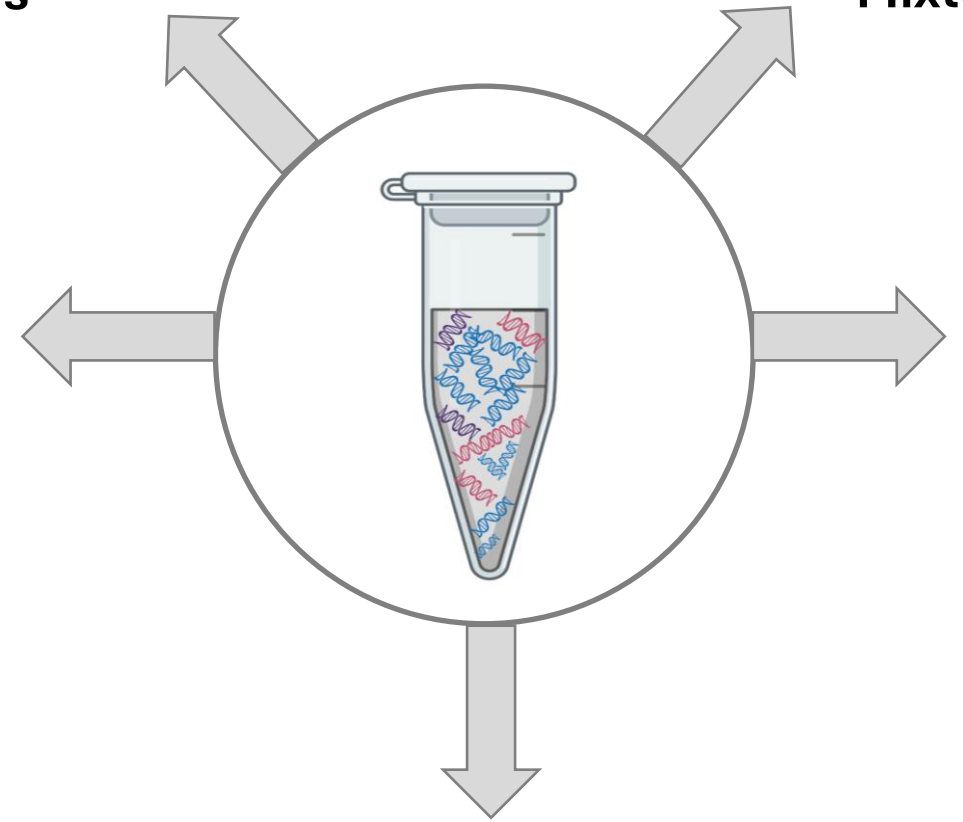
Factor Space Coverage of the Variables that Could Constitute DNA Samples of the Internal Validation

Number of Contributors

- 1, 2, 3, 4, 5……..

Mixture Genotype Combinations

- Level of Allele-Allele 1bp
- Level of Allele Sharing



DNA Quality

- Pristine
- Degraded
- PCR inhibited

Total DNA Template Amount

- 2 ng – 8 pg

Contributor's mixture ratios

- 2-person mixture: 1:1, 1:2, 1:4, 1:9……
- 3-person mixture: 1:1:1, 1:2:1, 1:4:1, 1:9:1, 1:2:2, 1:4:4, 1:9:9…..
- 4-person mixture: 1:1:1:1; 1:1:2:1; 1:1:4:1; 1:1:9:1; 1:2:2:1; 1:4:4:1; 1:9:9:1; 1:4:4:4…

E.g., Factor Space Coverage when Designing a Two Person Mixture

Number of Contributors (NOC)

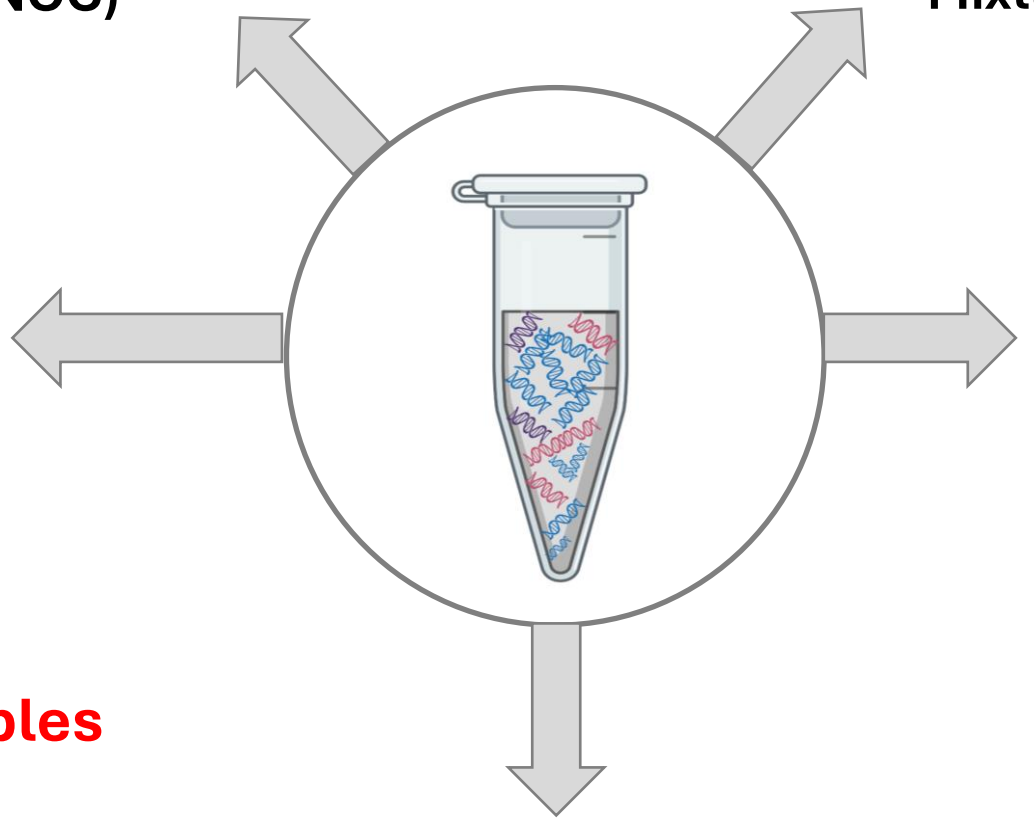
2-person mixture

Mixture Genotype Combinations

- Level of Allele-Allele Sharing

(Low/Medium/High)

(3 choices)



DNA Quality

- Pristine
- Differential degradation

(One or both Cs are degraded)

(3 choices)

Total DNA Template Amount

8 pg; 16 pg; 30 pg; 60 pg; 125 pg; 250 pg; 750 pg; 500 pg; 1 ng; 2 ng

(10 choices)

3 x10 x6 x 3 = 540 samples

Contributor's mixture ratios

1:1; 1:2; 1:4; 1:6; 1:8; and 1:9 (6 choices)

How do you get a good coverage of factor space in the validation studies without running 540 samples?

Software Key Features

3 Choosing a Validation Experimental Design

Using ***Statistical Theory of Experimental Design such as Space-Filling Design and Fractional Factorial Design*** and considering a ***laboratory available resources***, the software will output candidate experimental plans to ensure reasonable coverage of the factor space based on user specifications.

The software will ask the user to input the factor space desired to be covered by specifying:

* Experimental NoC (eNoC)
* Total number of PCR reactions per eNoC
* Total DNA template amounts
* Mixture ratios
* DNA quality

Mixture Genotype Combination:

* Level of allele sharing
* Level of A-A 1bp
* Level of homozygote genotypes
* Level of rare alleles
* MAC NoC

Software Key Features

3 Illustration of 3-Person Mixture Experimental Design

eNoC

- ☐ *1*
- ☐ *2*
- ☒ *3*
- ☐ *4*
- ☐ *5*

Number of PCR reactions

240

Serial dilution of DNA

- 8 pg
- 16 pg
- 30 pg
- 60 pg
- 120 pg
- 250 pg
- 500 pg
- 750 pg
- 1 ng
- 2 ng

10

Unique Mixture Genotype Combinations (MGCs)

24

- 20 single-source samples
- 3P = 1,140 unique mixture genotype combinations

Which MGCs should a lab choose?

Software Key Features

3 Illustration of 3-Person Mixture Experimental Design

eNoC

- ☐ 1
- ☐ 2
- ☒ 3
- ☐ 4
- ☐ 5

Number of PCR reactions

240

Serial dilution of DNA

10

Unique Mixture Genotype Combinations (MGCs)

24

	Min	Max
ASR	0.3	0.7
Homozygosity	15	33
Rare Alleles	2	5
A-A 1bp	0	1
MAC NoC	1	3

Software Key Features

3 Illustration of 4-Person Mixture Experimental Design

eNoC

- ☐ *1*
- ☐ *2*
- ☐ *3*
- ☒ *4*
- ☐ *5*

Number of PCR reactions

240

Serial dilution of DNA

- 16 pg
- 30 pg
- 60 pg
- 120 pg
- 250 pg
- 500 pg
- 750 pg
- 1 ng

8

Unique Mixture Genotype Combinations (MGCs)

30

	Min	Max
ASR	*0.4*	*0.6*
Homozygosity	*9*	*21*
Rare Alleles	*0*	*5*
A-A 1bp	*0*	*1*
MAC NoC	*3*	*4*

Software Key Features

3 Illustration of 3-Person Mixture Experimental Design

eNoC

- ☐ 1
- ☐ 2
- ☒ 3
- ☐ 4
- ☐ 5

Number of PCR reactions

240

Serial dilution of DNA

10

Unique Mixture Genotype Combinations (MGCs)

24

Min	Max	
0.3	0.7	***ASR***
15	33	***Homozygosity***
2	5	***Rare Alleles***
0	1	***A-A 1bp***
1	3	***MAC NoC***

DNA Quality (e.g., pristine/degraded)

- ☒ **0** (Pristine)
- ☒ **1** (Only 1 C is degraded)
- ☒ **2** (Only 2 Cs are degraded)
- ☒ **3** (All 3 Cs are degraded)
- ☐ 3P Mixture prepared and then degraded

Ratios

24

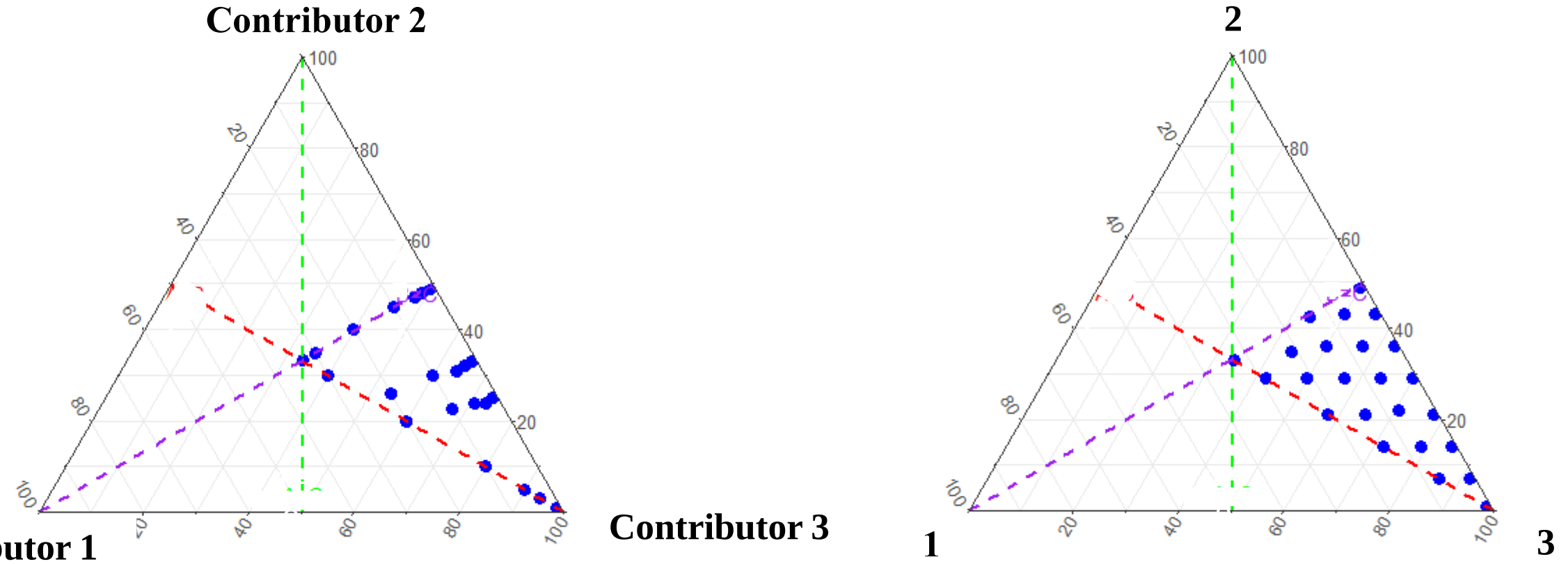
Software Key Features

Systematic Approach for Examining Coverage of Mixture Ratios

Illustration using 3P Mixtures and Ternary Diagrams

Each point represents a unique mixture ratio

SPACE FILLING (24 points)



By Hari Iyer

Software's Output: A 3-Person Experimental Design

MGC	C1	C2	C3	R1 (%)	R2 (%)	R3 (%)	Degradation
1	1	11	20d	7	7	86	1
2	2d	11d	20d	7	36	57	3
3	1	11	13	21	29	50	0
4	1	9d	11d	1	7	92	2
5	11d	12d	13d	1	1	98	3
6	1	11	15d	1	36	63	1
7	10	11	13d	21	35	44	1
8	4	12	20d	14	21	65	1
9	3d	5d	16d	1	14	85	3
10	5	10	12	14	43	44	0
11	5	9	19	7	29	64	0
12	1d	6d	7d	7	22	71	3
13	6d	13d	16d	1	49	50	3
14	4d	7d	14d	1	21	78	3
15	1	6	10	21	21	58	0
16	9	15d	18d	7	14	79	2
17	3	15d	17d	1	43	56	2
18	4	15	17	14	14	72	0
19	3	5d	15d	14	36	50	2
20	3	15	16d	14	29	57	1
21	3	7d	15d	29	29	42	2
22	4	14	19d	1	29	70	1
23	2	3	15	33	33	34	0
24	1	8d	12d	7	43	50	2

MGC = Mixture genotype combination

C = Contributor

R = Ratio

d = degradation

Software Key Features

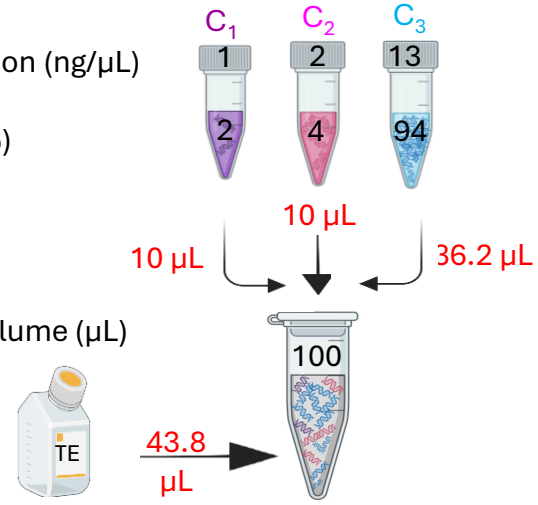
4 Mixture Calculations

The software will take user's requirements (e.g., C's concentration, desired mixture ratios) and constraints (minimum pipetting amounts, DNA mass in PCR reaction, minimum mixture stock solution) and provide an efficient strategy for making the desired mixtures.

E.g.,3P mixture preparation

Prepare Mixture Stock Concentration

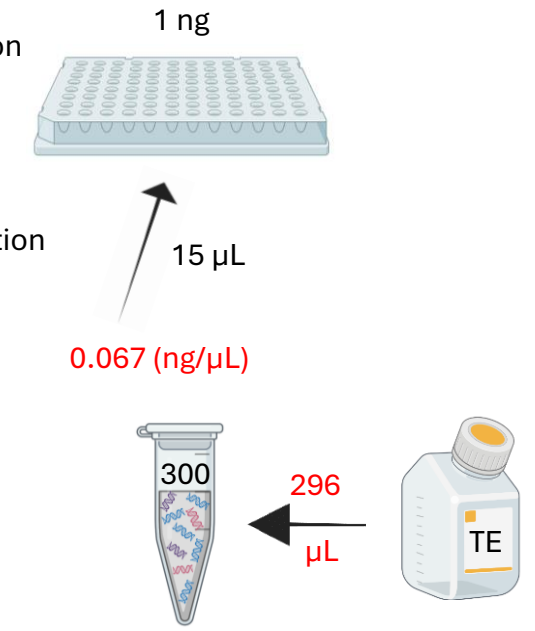
1. Contributor's Concentration (ng/μL)
2. Desired Mixture Ratios (%)
3. Desired Mixture Stock Volume (μL)
4. Desired Mixture Stock Concentration



4 μL

Prepare Mixture Working Solution

1. DNA Mass in PCR Reaction
2. DNA Volume in PCR Reaction
Mixture Final Concentration
3. Mixture Final Volume (μL)



Questions