Internal Validation of STRmix™ Version 2.4

using the Globalfiler™ PCR Amplification Kit and 3500/3500xL Genetic Analyzer

Part I: Estimation of STRmix™ Parameters

Part I of this validation has been technically reviewed and approved for use by:

Susan E. Welti — 121916

Susan Welti, FBU Technical Leader | Date

Part I of this validation was conducted and written by:

12-19-16

Jessica Skillman, Forensic Scientist III | Date

12-19-16

Andrew Feiter, Forensic Scientist I | Date

Technical support for this validation was also provided by Wendy Kashiwabara, Forensic Scientist II and Yoelia Perez, Forensic Scientist I.

I. STRmix™ Implementation

This document describes the estimation of the STRmix™ parameters for Globalfiler™ DNA profiling data from the District of Columbia Department of Forensic Science (DC DFS), Forensic Biology Unit (FBU) for use with STRmix™ Version 2.4.

II. STRmix™ Parameters

There are a number of parameters which are not optimised by the MCMC in a STRmix™ analysis. These parameters must be set by the user and are either determined by analysis of empirical data or modeled within STRmix™ using Model Maker. The laboratory specific parameters that are determined prior to use of STRmix™ are:

1. Stutter ratios
2. Analytical threshold (or limit of detection)
3. Saturation
4. Drop-in parameters
5. Allelic and stutter peak height variance
6. The hyper-parameter for the variance of locus specific amplification effects (LSAE)
7. Population settings including allele frequencies and theta values.

These parameters need to be defined for each STR kit, each protocol (e.g. cycle number variation), and CE platform (e.g. 3130 or 3500), and potentially each time there is a significant change of platform (e.g. a camera or laser change). All settings were determined for 29 cycle Globalfiler™ data analyzed on 3500xL A and 3500 B using GeneMapper™ *ID-X* Version 1.5. Peak height variance and locus specific amplification efficiencies were calculated using Model Maker within STRmix™ from analysis of empirical profile data. The results of these analyses are described within this report.

III. Stutter Ratios

As part of the laboratory's internal validation of the Globalfiler™ STR amplification kit and STRmix™ Version 2.4, a full stutter study was conducted to establish the parameters for both the initial data interpretation in the GeneMapper *ID-X* (GMID-X) Version 1.5 software and subsequent data interpretation by the STRmix™ Version 2.4 software. The study which established the STRmix™ parameters will be discussed in this section. The study which established the stutter filters for initial data interpretation will be discussed in Part II of this validation.

A limitation of the GMID-X software is its inability to apply stutter on a per allele basis. Previous research has shown stutter described on a per allele basis is a more accurate model for stutter behavior. The probabilistic genotyping software, STRmix™, has the ability to model stutter on a per allele basis; therefore as part of the internal validation, a stutter study focusing on the per allele stutter ratios was also performed.

The first parameter which was determined was maximum stutter which was set at .3 (30%) for back stutter and .1 (10%) for forward stutter, upon inspection of laboratory data and recommendations from the STRmix™ installation manual. The second STRmix™ portion of the studies included two separate studies for forward and back stutter which modeled allele specific stutter for all back stutter (N-4) and for the one locus for forward stutter, D22S1045.

a. Back Stutter

Back stutter is stutter that is one repeat smaller than its parent allele. In the case of the Globalfiler™ kit this means it is either N-4 or N-3 (in the case of D22S1045). This stutter is best described based upon the expected height of the parent allele. The values used to determine expected stutter heights are consider 'per allele'. Per allele stutter ratios are calculated using a linear equation and regressing stutter ratio against allele. Within STRmix™, stutter is estimated using the model *SR=m x Allele + c* where the intercept (c) and the slope (m) are determined using regression. Values for m and c were calculated using Microsoft Excel. A plot of SR (stutter ratio) versus Allele for each locus is provided in Appendix 1. A summary of the STRmix™ allelic stutter files for the DFS data is given below.

Locus	Intercept	Slope
D3S1358	-0.05058	0.007861
vWA	-0.08755	0.009079
D16S539	-0.04952	0.009205
CSF1PO	-0.04184	0.008606
TPOX	-0.02483	0.005184
D8S1179	0.013675	0.003463
D21S11	-0.02692	0.003036
D18S51	-0.03814	0.006889
D2S441	0.060762	-0.00135
D19S433	-0.06343	0.009036
TH01	0.011952	0.000787
FGA	-0.06719	0.005992
D22S1045	-0.11515	0.012739
D5S818	-0.03493	0.007995
D13S317	-0.05448	0.009109
D7S820	-0.04381	0.008549
SE33	0.03845	0.002218
D10S1248	-0.04828	0.008663
D1S1656	0.01621	0.003714
D12S391	-0.09527	0.008924
D2S1338	-0.01732	0.004419

A better explanatory variable for stutter ratio for loci with compound and complex structure has been shown to be the longest uninterrupted stretch of common repeats (LUS) within the allele and not the allele designation itself. LUS values for an allele are determined by sequencing, with a number of the common alleles used in forensic genotyping having been previously sequenced. A summary of these appear on STRBase. A plot of SR vs. LUS for compound and complex loci within the Globalfiler™ multiplex is provided within Appendix 1.

Another parameter within STRmix™ that determines expected stutter peak heights is known as a stutter exception file. This is a file based upon either LUS or an average observed stutter ratio. LUS is used where it is a good explanatory variable for the SR otherwise the average of the observed SR for each allele is used. A stutter exception file based upon laboratory data has been created and was used in this analysis. Where alleles are not present in this file (denoted by a "0") the expected stutter rates are calculated from the allele file. A summary of the source of the predicted SR for each locus is given below.

Locus	Source
D3S1358	Allele
vWA	Allele
D16S539	Allele
CSF1PO	Allele
TPOX	Allele
D8S1179	Average Observed
D21S11	Average Observed
D18S51	Allele
D2S441	Average Observed
D19S433	LUS
TH01	LUS
FGA	Allele
D22S1045	Allele
D5S818	Allele
D13S317	Allele
D7S820	Allele
SE33	Average Observed
D10S1248	Allele
D1S1656	LUS
D12S391	Allele
D2S1338	Allele

b. Forward Stutter

Within STRmix™ V2.4 forward stutter ratios (FSR) are modeled using a per allele model. DFS forward stutter peaks for 106 samples were analyzed. A summary of the numbers of observed forward stutter peaks is given in the table below. The locus demonstrating the highest rate of forward stutter was D22S1045. This was not unexpected given that this is a trinucleotide repeat marker, which are known to stutter more than tetra- or pentanucleotide repeats.

It is assumed that all loci are stuttering in the forward position (N+4) however most of these peaks are below the analytical threshold and therefore not visible. D22S1045 is the only locus where allele designation is a good predictor of FSR. As for the back stutter ratios mentioned previously, the intercept and slope within the parameters were determined by regression. For all other loci the average observed was calculated.

Forward Stutter observations and average forward stutter rates (where intercept= average)

Marker	Count of FS observations	Intercept	Slope
D3S1358	83	0.0070	0
vWA	50	0.0051	0
D16S539	114	0.0070	0
CSF1PO	73	0.0075	0
TPOX	13	0.0030	0
D8S1179	135	0.0071	0
D21S11	158	0.0080	0
D18S51	152	0.007924	0
D2S441	101	0.0076	0
D19S433	21	0.0073	0
TH01	5	0.0031	0
FGA	105	0.0068	0
D22S1045	136	-0.05934	0.006204
D5S818	113	0.0076	0
D13S317	134	0.0061	0
D7S820	87	0.0046	0
SE33	219	0.0080	0
D10S1248	39	0.0070	0
D1S1656	189	0.0077	0
D12S391	43	0.0083	0
D2S1338	41	0.0088	0

IV. Analytical Threshold

The assignment of a signal as allelic product as opposed to baseline or noise is important in DNA profile analysis. This differentiation is usually undertaken using a set threshold above which peaks are deemed to be allelic if they also meet certain morphological requirements, and below which they are ignored, regardless of morphology. The issue is to assign an analytical threshold (AT) to minimize the detection of artifacts while maximizing the detection of allelic peaks. The recommended analytical threshold for samples amplified at 29 cycles with Globalfiler™ and run on 3500xL A or 3500 B is 90rfu. Details regarding this recommendation can be found in the Globalfiler™ Validation.

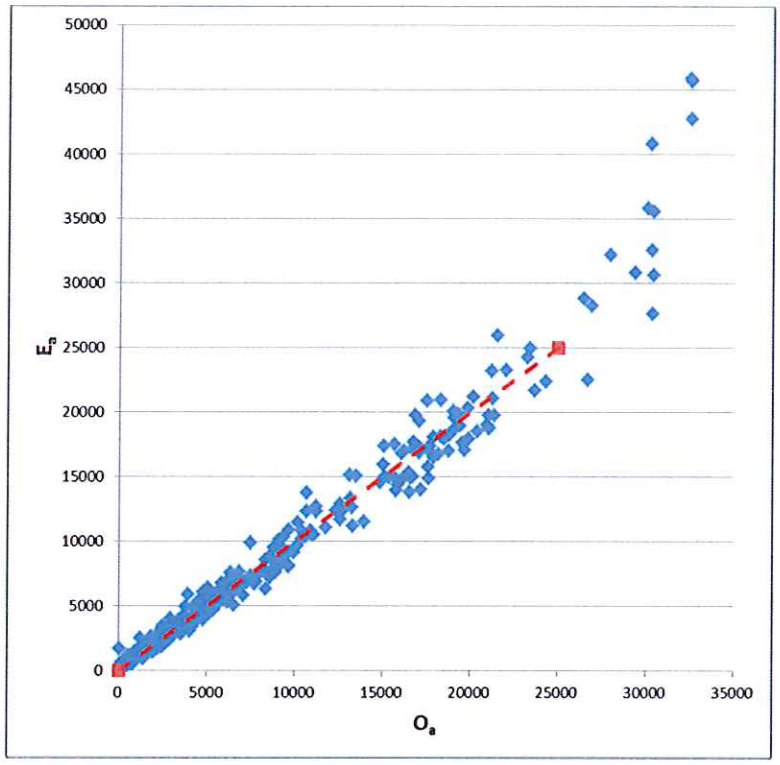
V. Saturation

The peaks in a DNA profile are measured using fluorescence. The amount of fluorescence is proportional to the amount of DNA present. This fluorescence is captured by a camera. It is expected that as more DNA is added into a PCR the resulting peak height (measured in relative fluorescent units) in an electropherogram will increase. The camera can become saturated when there is too much fluorescence detected. This means we can no longer accurately measure the height of the peaks observed or estimate how much DNA is really represented by this result. Following this we can no longer accurately model over saturated peak heights using STRmix™. The saturation setting is the upper limit for a peak's height permitted in the software, beyond which the model is no longer optimal. The software will treat peaks in the input evidence data above this value as qualitative only. Saturation, like the analytical threshold, is mostly instrument related and not kit or method dependent.

The expected height of alleles within the Globalfiler™ Validation Sensitivity Study dataset was calculated using the formula:

$$E_a = \frac{O_{a-1}}{\varepsilon SR_a}$$

Where (E_a) is the expected peak height calculated from the observed stutter height (O_{a-1}) and εSR_a is the expected stutter ratio for allele a calculated using the equation described in Section IIIa. A plot of O_a versus E_a is provided in the figure below. A vertical line at O_a = 30,000 rfu is a common saturation limit for a 3500 instrument. The points should deviate from the $x = y$ line at the saturation value. After inspection, a saturation threshold setting of 25,000 rfu is recommended.



VI. Drop-in parameters

The Globalfiler™ validation contains a full assessment of the alleles designated as drop-in for all studies. Due to the limited amount of data obtained, a successful gamma distribution with which to model drop-in (as suggested by the STRmix™ V2.4 Implementation and Validation Guide) is not expected. An alternative method is to use the values obtained directly from the validation as uninformed priors. This will cause the software to apply an equal probability to any peak considered drop-in regardless of its peak height. The following settings will be used:

- Drop-in cap: 200
- Drop-in frequency: 6/13200 = 0.0004545
- Drop-in parameters: 0,0

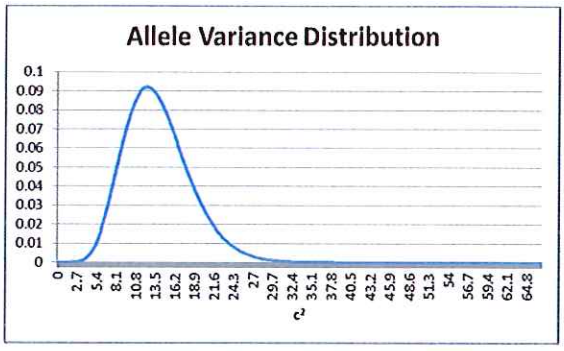
VII. Peak height variance and LSAE using Model Maker

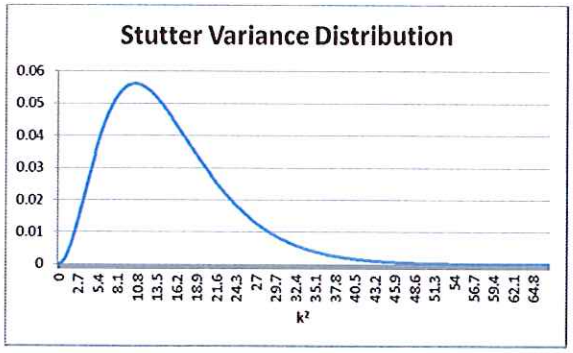
A set of 10 samples amplified at 10 different targets (0.1-1.0ng) was prepared for detection. The plate was run on both 3500xL A and 3500 B. LAM-0_9_03_D07 and ladder_07_E07 were not included in the analysis due to a sample preparation error at the amplification step. AB_0_6_08_D08, AB-0_7_03_E08, JMF-0_3_05_C05, SM-0_8_03_C03, SMM-0_5_01_F01_3500B and WK-0_2_10_E10_3500 Instrument were not included in the analysis due to possible poor injection. NCJ-1_0_04_C10 and NCJ-1_0_10_C10 were not included in the analysis due to offscale peak heights. All other samples were analyzed using GeneMapper™ *ID-X* Version 1.5 with the settings recommended by the Globalfiler™ validation and N-2 filters only.

All N+4 and other artifacts were manually removed. All controls produced expected results with the exception of the ladder identified above.

Empirical observations and experience suggests that profiles differ in variance (hereafter "quality"). Within STRmix™ the variability of peaks within profiles is described using a model containing a variance constant. Within V2.4 allele and stutter peaks have separate variances, c^2 and k^2, respectively. The c^2 and k^2 terms are variables which are determined through the MCMC process. The starting position for these values within the MCMC is the mode of a gamma distribution based on empirical values from the DFS laboratory. 190 single source profiles of varying quality were analysed using the Model Maker function within STRmix™. A summary of the results and plots of the allele and stutter gamma distributions are provided below. In this table, the values for allele and stutter variance are provided as rounded values from Microsoft Excel calculations. Variation in the STRmix™ calculated modes was observed at the thousandths place when parameters were input with the rounded alpha and beta values. No effect on data interpretation is expected based on the nominal value of this variation.

Summary of Model Maker Results				
Multiplex	Number of Profiles Analyzed	Allele Variance Parameters	Stutter Variance Parameters	Mean LSAE variance
Globalfiler™ (29 cycle)	190	gamma (9.277,1.492) Mode 12.350	gamma (3.318, 4.502) Mode 10.437	0.01881





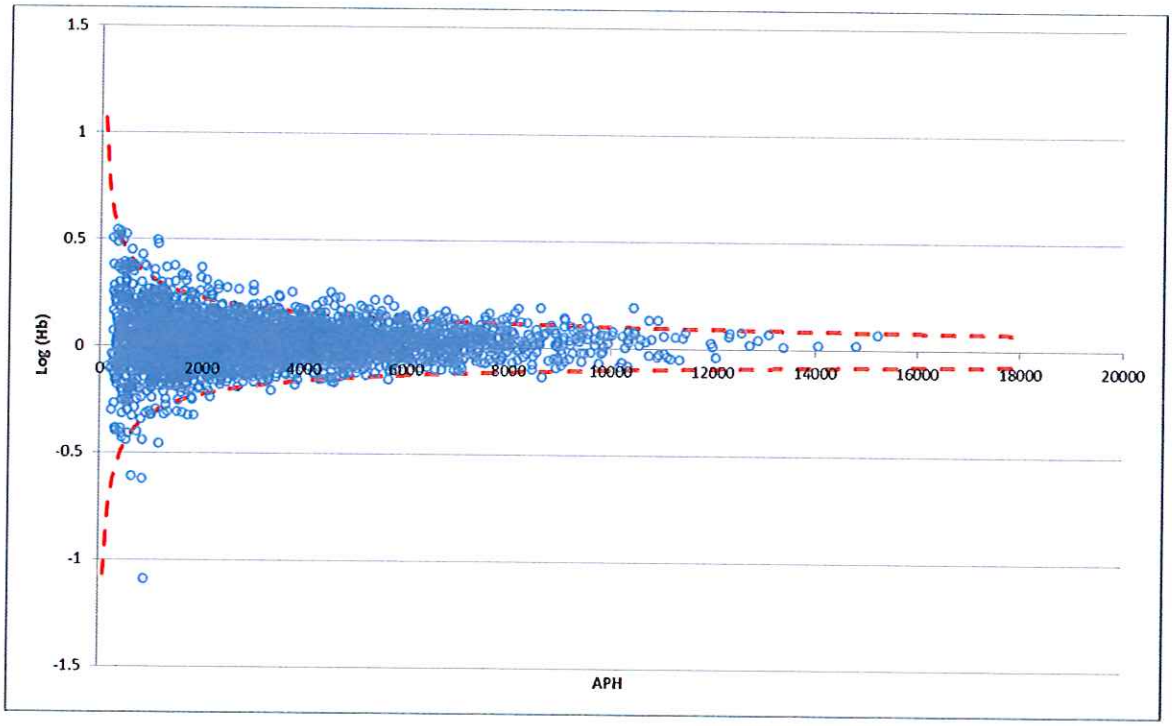
Heterozygote balance was calculated for all heterozygote loci for the Model Maker profiles. Heterozygote balance (*Hb*) was calculated as:

$$Hb = \frac{O_{HMW}}{O_{LMW}}$$

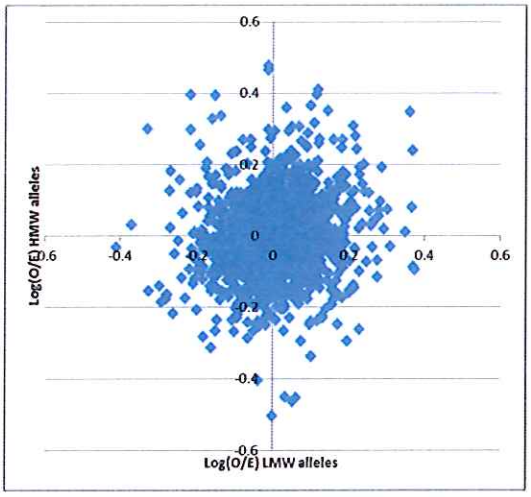
Where O_{HMW} refers to the observed height of the high molecular weight allele and O_{LMW} the observed height of the low molecular weight allele. Previous work has suggested that there is a

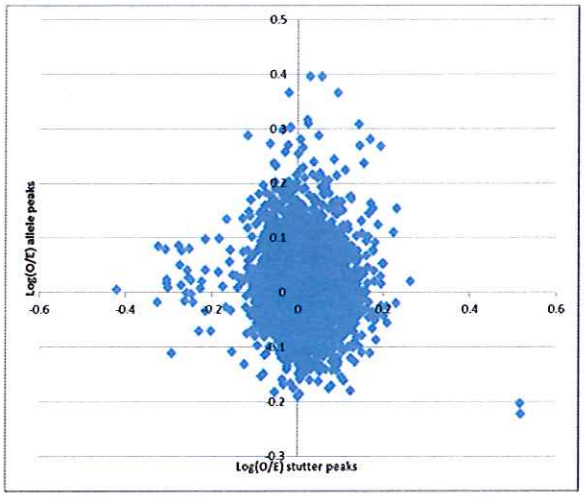
relationship between the variation in peak height and the variation in *Hb* [6,7]. In single source profiles, variability in *Hb* reduces as the average peak height (APH) at a locus increases. The variance of *Hb* is expected to be twice the variance of the individual allelic peaks assuming the variance of each peak is the same. This allows an approximate comparison between the variance from the STRmix™ MCMC approach and a readily determined variable from empirical data.

The plot of log*Hb* versus APH for each of the datasets described above and the expected 95% bounds (plotted as dotted lines) calculated at $\pm\sqrt{2}\times 1.96\times\sqrt{\frac{c^2}{APH}}$ where c^2 = 13.35, the 50th percentile from the gamma distribution from the combination data set. The 95% bounds encapsulate sufficient data as demonstrated in the graphs (coverage = 95.5%) demonstrating that the values for variance are sufficiently optimised. The plot below is an approximate check of Model Maker.



The graphs below are correlation plots for LMW versus HMW allele and allele versus stutter peaks for the Model Maker dataset. The final correlation for the LMW versus HMW alleles was 0.1159 and the final correlation for the allele versus stutter peaks was -0.03184. The distribution of the points within the figures is as expected, with no observed correlation. There are some outliers observed in the logarithm of the observed over expected stutter peak height versus log(O/E) allelic peak height plot. These are larger than expected stutter peaks that were labeled at analysis however they do not affect the results.





VIII. Population Frequency Data Files and Settings

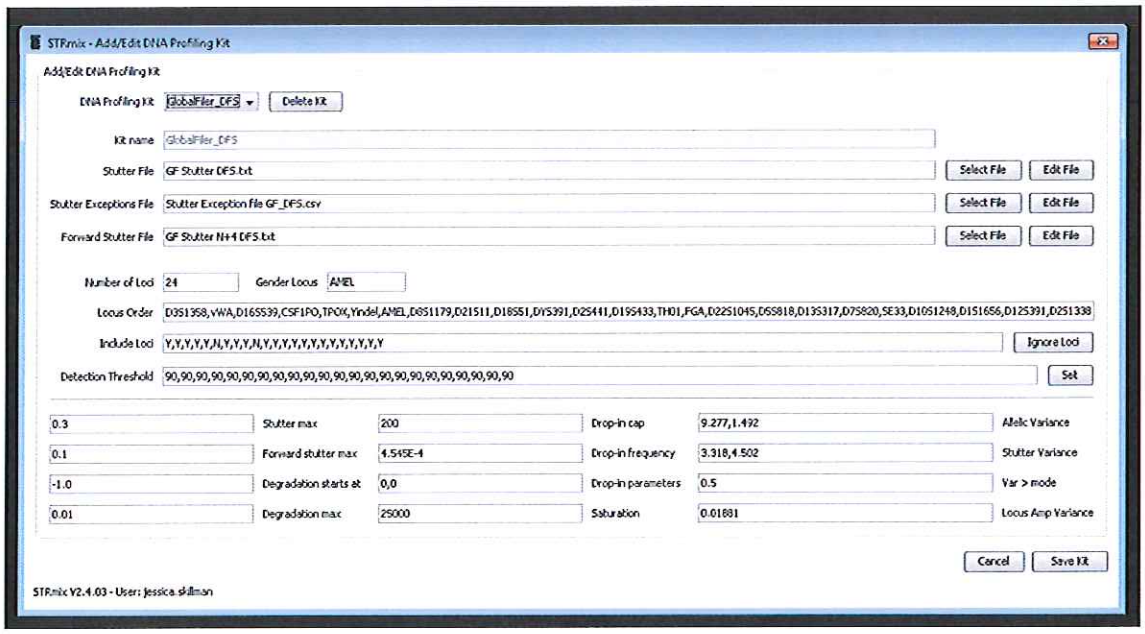
Population frequency files from the African American (Combined), Caucasian, Southeast Hispanic and Southwest Hispanic databases for the 2015 Expanded FBI STR Population Data were provided by STRmix™. Each file was checked for concordance to the published population data by three separate individuals. The following discrepencies were observed, however determined to be insignificant:

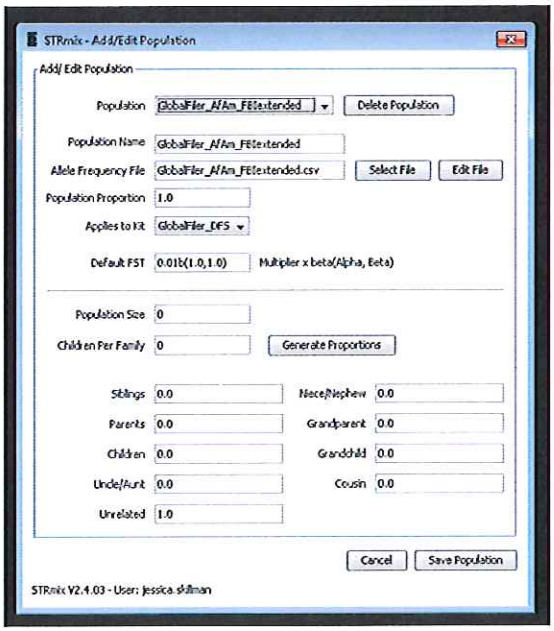
- The published databases contained some frequencies for the Yindel locus and the STRmix™ files did not. The Yindel locus will not be used for statistical calculations at the DC DFS FBU.
- For the published databases, N is equal to the number of individuals tested. For the STRmix™ files, N is equal to the number of alleles tested. The STRmix™ files therefore list N as double the published FBI data. Discrepancies were observed at the Yindel and DYS391 loci, however, these locations will not be used by DC DFS FBU for statistical calculations.
- In the Southeast Hispanic published database, an allele frequency is listed for an allele ">27" at D18S51. Because STRmix™ does not accept this allele designation and requires an analyst to assign a specific value for all alleles, it is not listed in the STRmix™ file. The total frequency for this locus accurately reflects a value less than one.

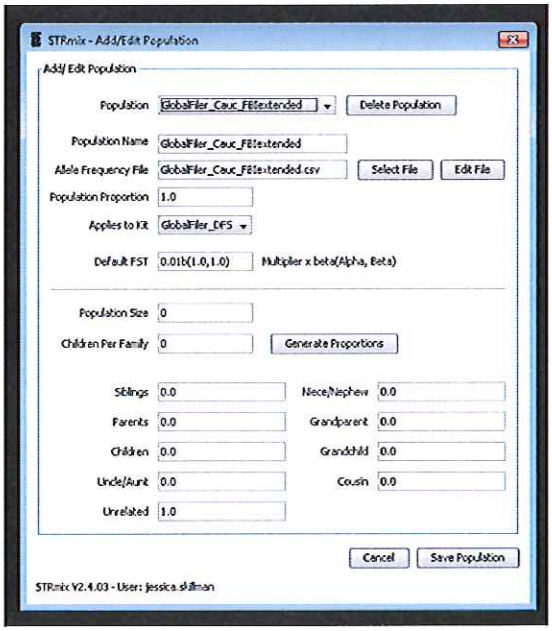
Additionally, settings files for all four populations were provided by STRmix™. Each was verified to ensure appropriate values were set to provide a non-stratified, non-unified statistic using a point estimate of 0.01 for theta.

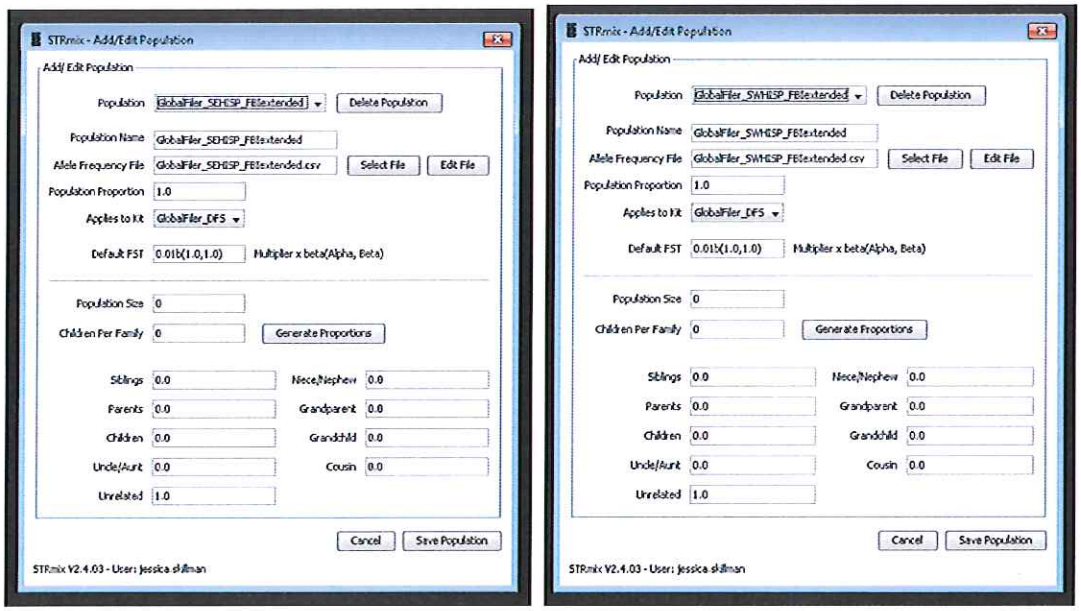
IX. Conclusions

The recommended STRmix™ V2.4 default parameters for the interpretation of the DFS 29 cycle Globalfiler™ profiles run on 3500xL A or 3500 B are shown below.


STRmix - Add/Edit DNA Profiling Kit
Add/Edit DNA Profiling Kit
DNA Profiling Kit
GlobalFiler_DFS
Delete Kit
Kit name
GlobalFiler_DFS
Stutter File
GF Stutter DFS.txt
Select File
Edit File
Stutter Exceptions File
Stutter Exception file GF_DFS.csv
Forward Stutter File
GF Stutter N+4 DFS.txt
Number of Loci
24
Gender Locus
AMEL
Locus Order
D3S1358,vWA,D16S539,CSF1PO,TPOX,Yindel,AMEL,D8S1179,D21S11,D18S51,DYS391,D2S441,D19S433,TH01,FGA,D22S1045,D5S818,D13S317,D7S820,SE33,D10S1248,D1S1656,D12S391,D2S1338
Include Loci
Y,Y,Y,Y,Y,N,Y,Y,N,Y,Y,Y,Y,Y,Y,Y,Y,Y,Y,Y,Y
Ignore Loci
Detection Threshold
90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90,90
Set
0.3
Stutter max
200
Drop-in cap
9.277,1.492
Allelic Variance
0.1
Forward stutter max
4.545E-4
Drop-in frequency
3.318,4.502
Stutter Variance
-1.0
Degradation starts at
0,0
Drop-in parameters
0.5
Var > mode
0.01
Degradation max
25000
Saturation
0.01881
Locus Amp Variance
Cancel
Save Kit
STRmix V2.4.03 - User: jessica.shillman


STRmix - Add/Edit Population
Add/ Edit Population
Population
GlobalFiler_AfAm_FEtextended
Delete Population
Population Name
GlobalFiler_AfAm_FEtextended
Allele Frequency File
GlobalFiler_AfAm_FEtextended.csv
Select File
Edit File
Population Proportion
1.0
Applies to Kit
GlobalFiler_DFS
Default FST
0.01b(1.0,1.0)
Multiplier x beta(Alpha, Beta)
Population Size
0
Children Per Family
0
Generate Proportions
Siblings
0.0
Niece/Nephew
0.0
Parents
0.0
Grandparent
0.0
Children
0.0
Grandchild
0.0
Uncle/Aunt
0.0
Cousin
0.0
Unrelated
1.0
Cancel
Save Population
STRmix V2.4.03 - User: jessica.shillman


STRmix - Add/Edit Population
Add/ Edit Population
Population
GlobalFiler_Cauc_FEtextended
Delete Population
Population Name
GlobalFiler_Cauc_FEtextended
Allele Frequency File
GlobalFiler_Cauc_FEtextended.csv
Select File
Edit File
Population Proportion
1.0
Applies to Kit
GlobalFiler_DFS
Default FST
0.01b(1.0,1.0)
Multiplier x beta(Alpha, Beta)
Population Size
0
Children Per Family
0
Generate Proportions
Siblings
0.0
Niece/Nephew
0.0
Parents
0.0
Grandparent
0.0
Children
0.0
Grandchild
0.0
Uncle/Aunt
0.0
Cousin
0.0
Unrelated
1.0
Cancel
Save Population
STRmix V2.4.03 - User: jessica.shillman

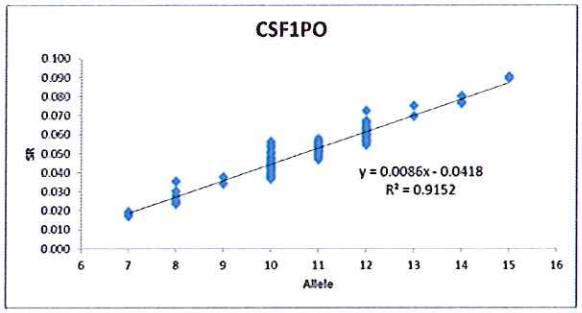

STRmix - Add/Edit Population
Add/ Edit Population
Population
GlobalFiler_SEHESP_FEtextended
Delete Population
Population Name
GlobalFiler_SEHESP_FEtextended
Allele Frequency File
GlobalFiler_SEHESP_FEtextended.csv
Select File
Edit File
Population Proportion
1.0
Applies to Kit
GlobalFiler_DFS
Default FST
0.01b(1.0,1.0)
Multiplier x beta(Alpha, Beta)
Population Size
0
Children Per Family
0
Generate Proportions
Siblings
0.0
Niece/Nephew
0.0
Parents
0.0
Grandparent
0.0
Children
0.0
Grandchild
0.0
Uncle/Aunt
0.0
Cousin
0.0
Unrelated
1.0
Cancel
Save Population
STRmix V2.4.03 - User: jessica.skillman
STRmix - Add/Edit Population
Add/ Edit Population
Population
GlobalFiler_SWHESP_FEtextended
Delete Population
Population Name
GlobalFiler_SWHESP_FEtextended
Allele Frequency File
GlobalFiler_SWHESP_FEtextended.csv
Select File
Edit File
Population Proportion
1.0
Applies to Kit
GlobalFiler_DFS
Default FST
0.01(1.0,1.0)
Multiplier x beta(Alpha, Beta)
Population Size
0
Children Per Family
0
Generate Proportions
Siblings
0.0
Niece/Nephew
0.0
Parents
0.0
Grandparent
0.0
Children
0.0
Grandchild
0.0
Uncle/Aunt
0.0
Cousin
0.0
Unrelated
1.0
Cancel
Save Population
STRmix V2.4.03 - User: jessica.skillman

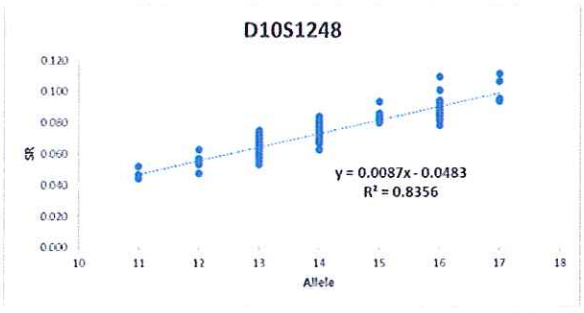
X. References

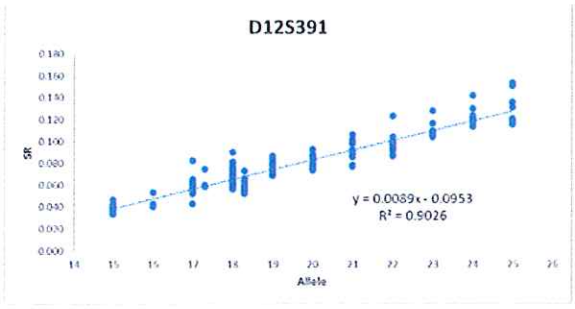
1. Bright J-A, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Science International: Genetics. 2013;7:296-304.
2. Brookes C, Bright J-A, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Science International: Genetics. 2012;6:58-63.
3. Walsh PS, Fildes NJ, Reynolds R. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Research. 1996;24:2807-12.
4. Butler JM, Reeder DJ. Short Tandem Repeat DNA Internet DataBase.
5. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Research. 2001;29:320 - 2.
6. Bright J-A, Huizing E, Melia L, Buckleton J. Determination of the variables affecting mixed MiniFiler™ DNA profiles. Forensic Science International: Genetics. 2011;5:381-5.
7. Bright J-A, Turkington J, Buckleton J. Examination of the variability in mixed DNA profile parameters for the Identifiler multiplex. Forensic Science International: Genetics. 2009;4:111-4.

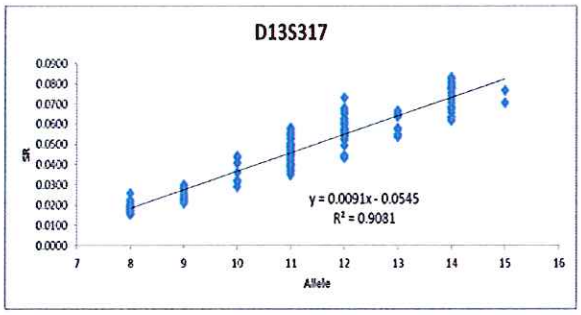
XI. Appendix 1

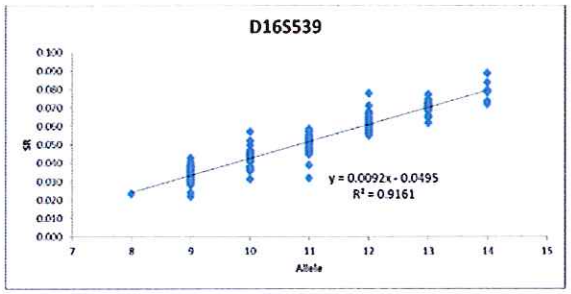
SR versus Allele

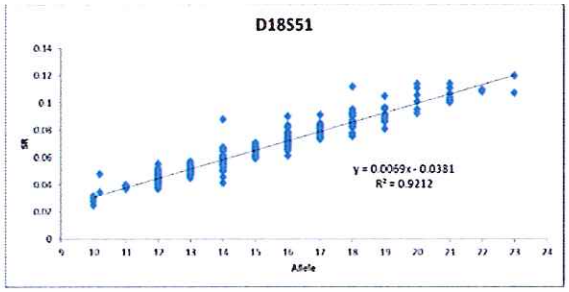


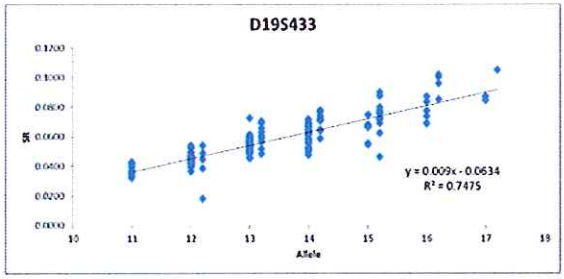


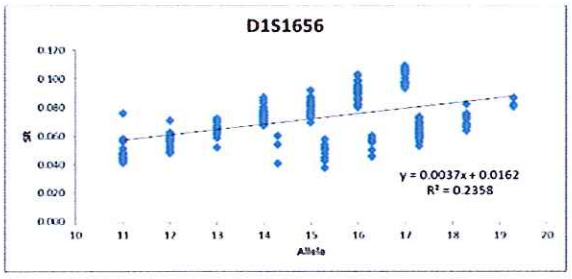


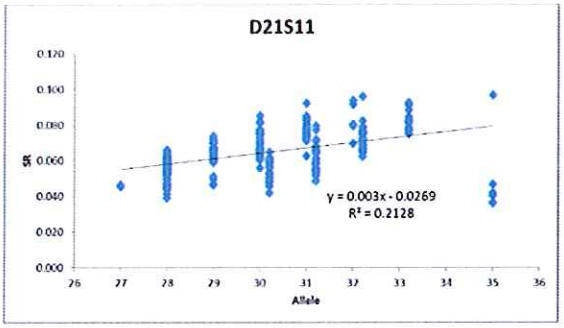


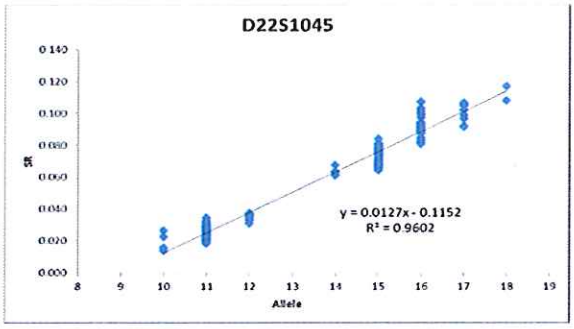


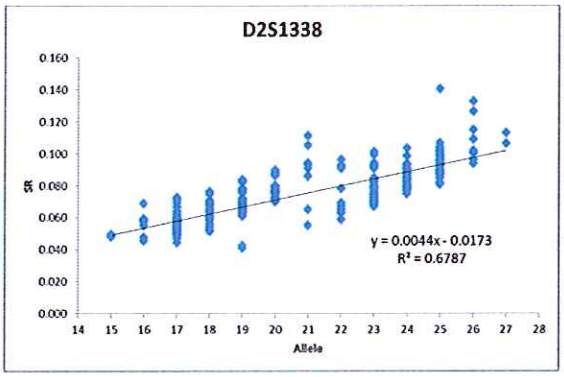


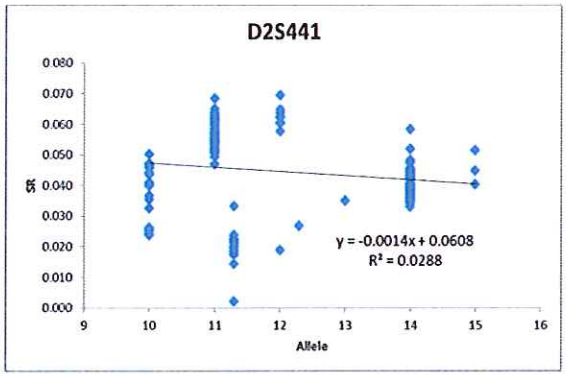


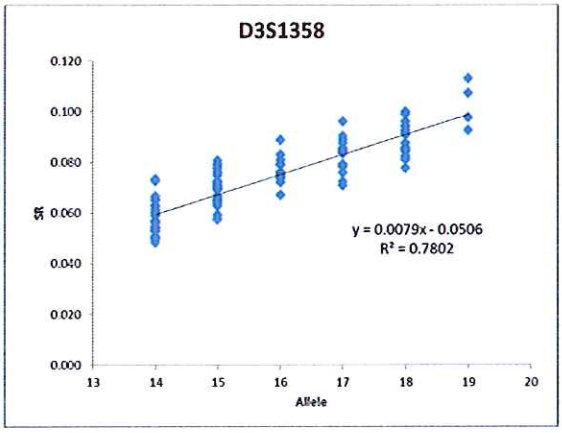


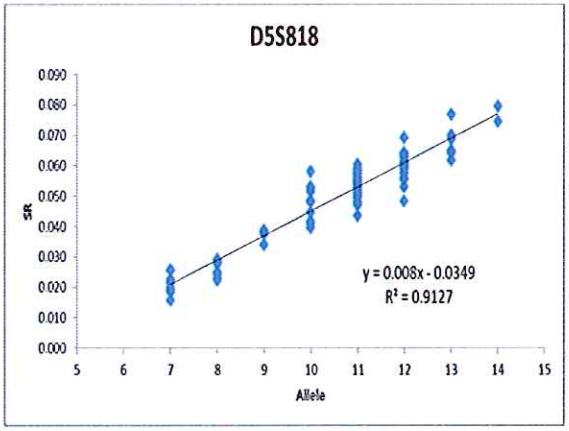

D21S11
y = 0.003x - 0.0269
R² = 0.2128
SR
Allele

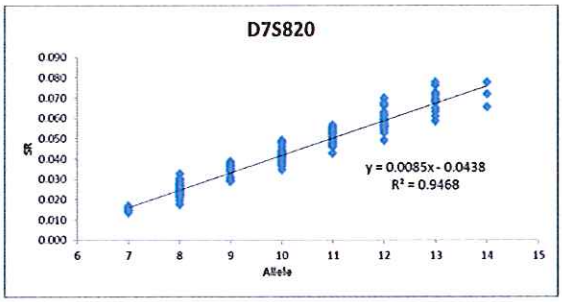

D22S1045
y = 0.0127x - 0.1152
R² = 0.9602
SR
Allele

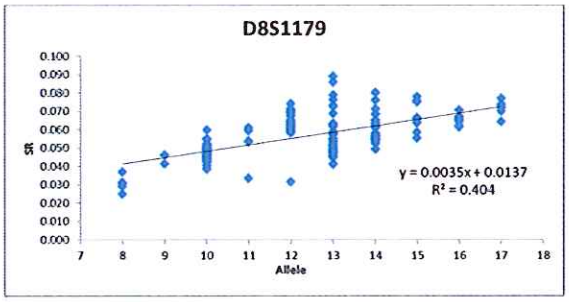

D2S1338
y = 0.0044x - 0.0173
R² = 0.6787
SR
Allele

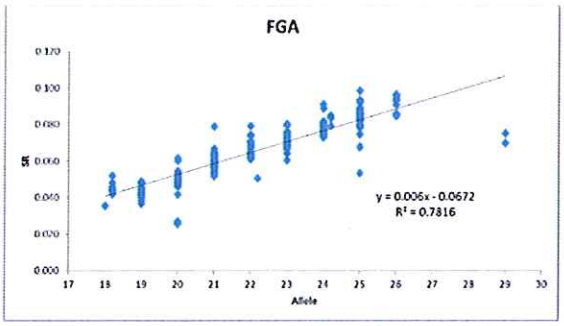

D2S441
y = -0.0014x + 0.0608
R² = 0.0288
SR
Allele

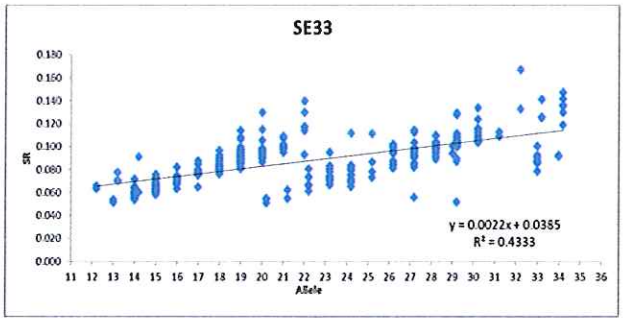

D3S1358
y = 0.0079x - 0.0506
R² = 0.7802
SR
Allele

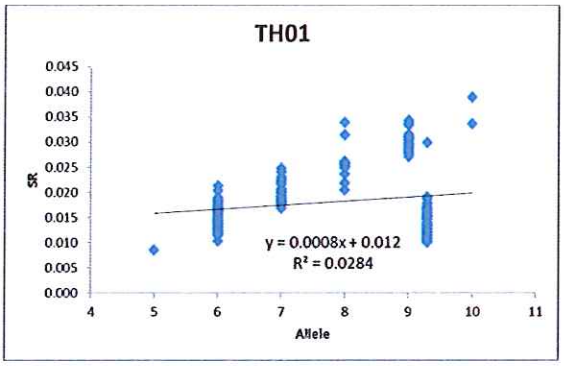

D5S818
y = 0.008x - 0.0349
R² = 0.9127
SR
Allele

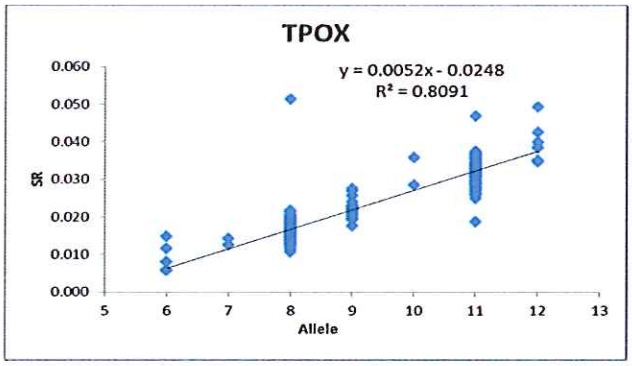

D7S820
y = 0.0085x - 0.0438
R² = 0.9468
SR
Allele

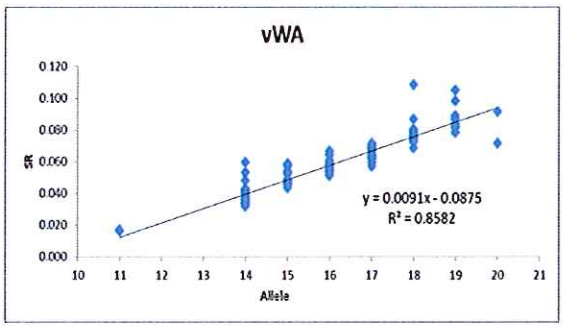

D8S1179
y = 0.0035x + 0.0137
R² = 0.404
SR
Allele

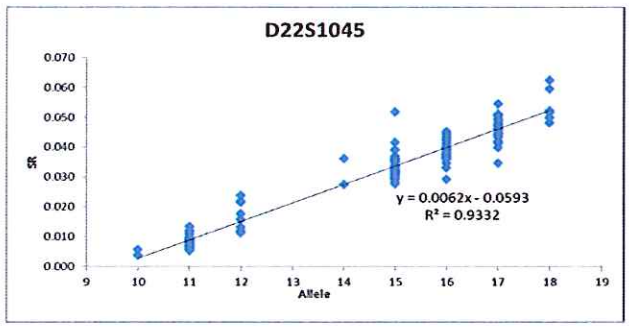






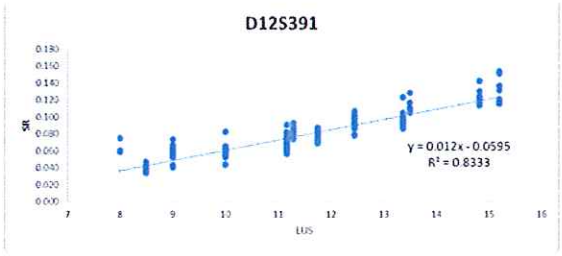


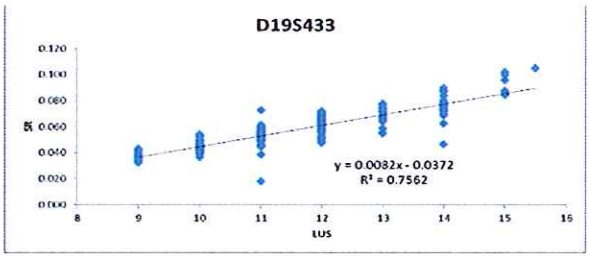


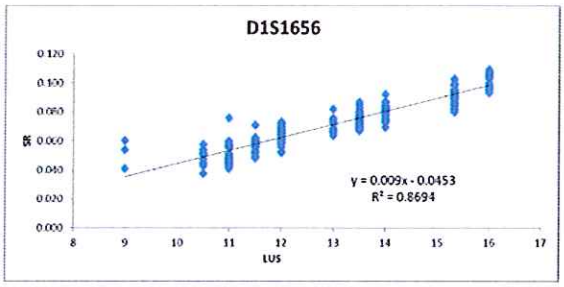


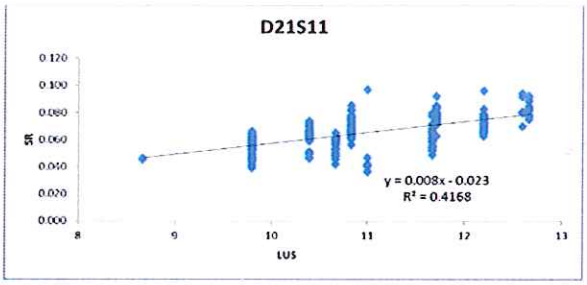
Note: The above D22S1045 Plot is of FSR

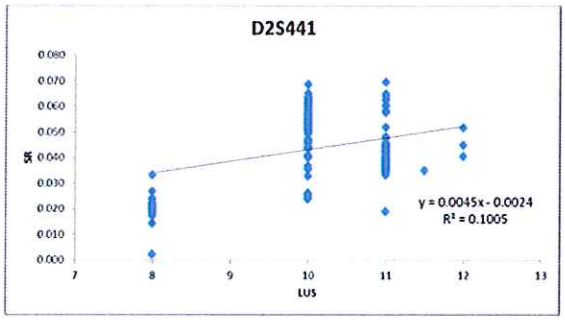
SR versus LUS

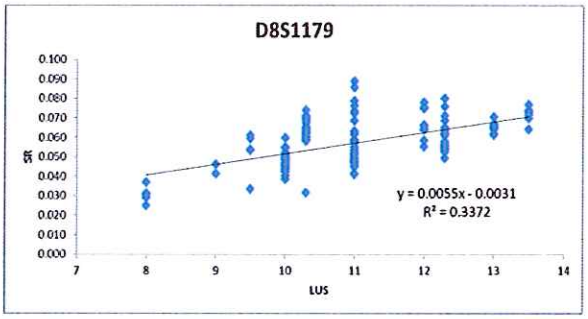


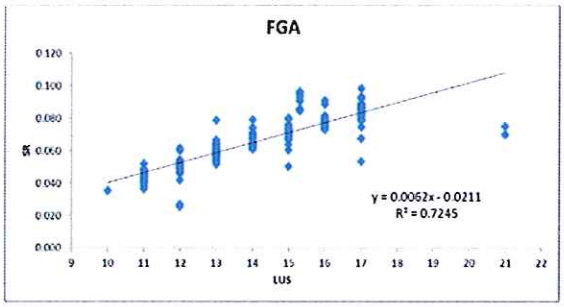


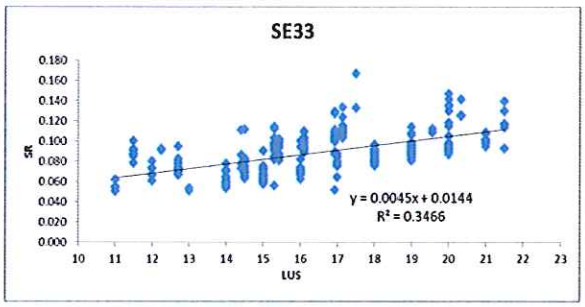

D1S1656
y = 0.009x - 0.0453
R² = 0.8694
SR
LUS

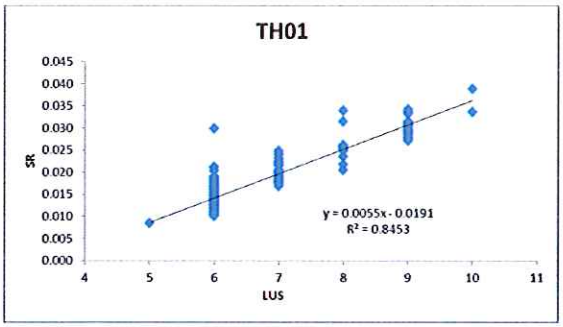

D21S11
y = 0.008x - 0.023
R² = 0.4168
SR
LUS

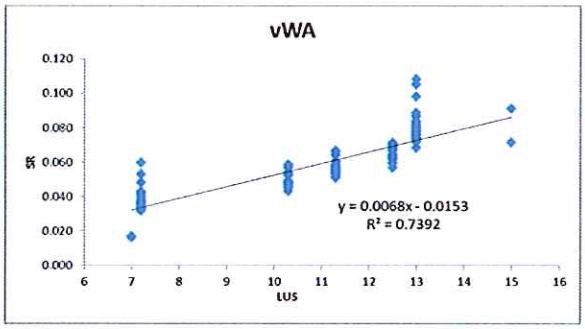

D2S441
y = 0.0045x - 0.0024
R² = 0.1005
SR
LUS


D8S1179
y = 0.0055x - 0.0031
R² = 0.3372
SR
LUS


FGA
y = 0.0062x - 0.0211
R² = 0.7245
SR
LUS


SE33
y = 0.0045x + 0.0144
R² = 0.3466
SR
LUS


TH01
y = 0.0055x - 0.0191
R² = 0.8453
SR
LUS


vWA
y = 0.0068x - 0.0153
R² = 0.7392
SR
LUS