Estimation of STRmix™ parameters for OCME New York Laboratory

STRmix™ Implementation

This document describes the estimation of the STRmix™ parameters for PowerPlex Fusion 5C DNA profiling data from the Office of the Chief Medical Examiner New York Forensic Laboratory (OCME NY) for use in STRmix™ V2.4. A description of the methods used to generate this data is available within The STRmix™ V2.4 Implementation and Validation Guide.

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 modelled within STRmix™ using Model Maker. The laboratory specific parameters that are determined prior to use of STRmix™ are:

- Analytical threshold (detection threshold)
- Stutter ratios
- Drop-in parameters
- Saturation
- Allelic and stutter peak height variance
- The hyper-parameter for the variance of locus specific amplification effects (LSAE).

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). Stutter ratios and saturation were determined for OCME Fusion data analysed on a 3130 capillary electrophoresis instrument. Peak height variance and locus specific amplification efficiencies are calculated using Model Maker within STRmix™ from analysis of empirical profile data. The results of these analyses are described within this report.

Analytical Thresholds

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 a threshold, often termed the limit of detection (LOD) or analytical threshold (AT), to minimise the detection of artefacts whilst maximising the detection of allelic peaks.

Optimum AT values of 50 rfu have previously been determined by OCME NY for all the Fusion loci. These values were used for all data analysis within this report.

Stutter

Back stutter

There are three parameters within STRmix™ that assist with the calculation of expected back stutter heights and therefore require optimisation. The first is the maximum allowable stutter ratio. The maximum allowable stutter ratio reduces run time by only permitting peaks in a stutter position below a certain percentage to be considered stutter. This parameter has been set at 0.3 (30%) for back stutter and 0.1 (10%) for forward stutter based on inspection of laboratory stutter ratio data.

The second parameter is a file used to model the expected heights of the stutter peaks based on their partner allele designation. The values used to determine expected stutter heights are '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 $SR = m \times Allele + c$ where the intercept (*c*) and slope (*m*) are determined using regression. Values for *m* and *c* were calculated using regression analysis in Excel. A plot of back and forward stutter ratio versus Allele for each locus is provided in Appendix 1 and Appendix 2, respectively. A summary of the STRmix™ back and forward allelic stutter files for the OCME NY data is given in Table 1.

Table 1: OCME NY per allele Fusion back and forward stutter values for STRmix™

	Back stutter		Forward stutter	
Locus	Intercept	Slope	Intercept	Slope
D3S1358	-0.07389	0.00976	0.00813	0
D1S1656	0.01003	0.00493	0.00997	0
D2S441	0.07168	-0.00186	0.0088	0
D10S1248	-0.05564	0.0094	0.00266	0.00041
D13S317	-0.0492	0.00917	-0.00115	0.00097
Penta E	-0.01536	0.00376	0.00857	0
D16S539	-0.06043	0.01102	-0.00407	0.001
D18S51	-0.03995	0.00779	-0.00468	0.00088
D2S1338	-0.01163	0.00444	0.00677	0
CSF1PO	-0.03106	0.00858	-0.01061	0.00212
Penta D	0.01175	0.00075	0.0074	0
TH01	-0.01115	0.0045	0.00506	0
vWA	-0.0837	0.00945	0.0081	0
D21S11	-0.0886	0.00547	0.01109	0
D7S820	-0.0586	0.01111	0.00266	0.00041
D5S818	-0.06403	0.01057	-0.0076	0.00148
TPOX	-0.02138	0.00537	-0.00165	0.00087
DYS391	0	0	0	0
D8S1179	0.01684	0.00416	0.00977	0
D12S391	-0.10877	0.01019	0.00758	0
D19S433	-0.05558	0.0089	0.00037	0.00051
FGA	-0.08396	0.00684	-0.0001	0.00038
D22S1045	-0.11718	0.01391	-0.0714	0.0079

A better explanatory variable for stutter ratio for some loci with compound and complex repeat structure has been shown to be the longest uninterrupted stretch of common repeats (LUS) within the allele [1-3] and not the allele designation itself. Values for LUS are determined by sequencing alleles. A number of common alleles for forensic loci have been typed. A summary of these appear on STRBase [4, 5]. A plot of *SR* versus LUS for compound and complex loci within the Fusion multiplex is also provided within Appendix 1.

The third parameter within STRmix™ that determines expected stutter peak heights is an exception file based on either LUS or an average observed stutter ratio. LUS is used where it is a good

explanatory variable for *SR* otherwise the average of the observed *SR* is used. A stutter exception file based on laboratory data has been created and was used in this analysis. Where alleles are not present in this file the expected stutter rates are calculated from the allele file (Table 1). A summary of the source of the predicted *SR* for each locus is given in Table 2.

Table 2: A summary of the explanatory variables for the predicted *SR* for each of the OCME NY loci

Locus	Explanatory variable
D3S1358	Allele
D1S1656	Average
D2S441	Average
D10S1248	Allele
D13S317	Average
Penta E	Allele
D16S539	Allele
D18S51	Average
D2S1338	Average
CSF1PO	Allele
Penta D	Allele
TH01	LUS
vWA	Average
D21S11	Average
D7S820	Allele
D5S818	Allele
TPOX	Allele
DYS391	N/A
D8S1179	Average
D12S391	Allele
D19S433	LUS
FGA	Average
D22S1045	Allele

Drop-in parameters

Drop-in is non-reproducible, unexplained peaks observed within a profile. There are four parameters used for the modelling of drop-in in STRmix™. These are:

1. Z: the detection threshold or analytical threshold
2. A cap on the maximum allowed drop-in peak height
3. The drop-in frequency
4. α,β: two parameters for the gamma model.

Drop-in rates for a laboratory platform (multiplex and instrument combination) should be monitored. This is done by recording counts and corresponding heights of drop-in peaks observed in negative controls and counts of negative controls without drop-in peaks. Within STRmix™ drop-in is modelled using a gamma distribution.

OCME NY's drop-in optimised parameters are provided in Table 3.

Table 3: OCME NY drop-in parameters for STRmix™ for the Fusion data

Drop-in cap	100
Drop-in frequency	0.0024
Drop-in parameters	0,0

The drop-in parameters (0,0) apply a uniform prior which applies the same penalty (or probability) to peaks less than the cap and considered drop-in irrespective of their height.

Saturation

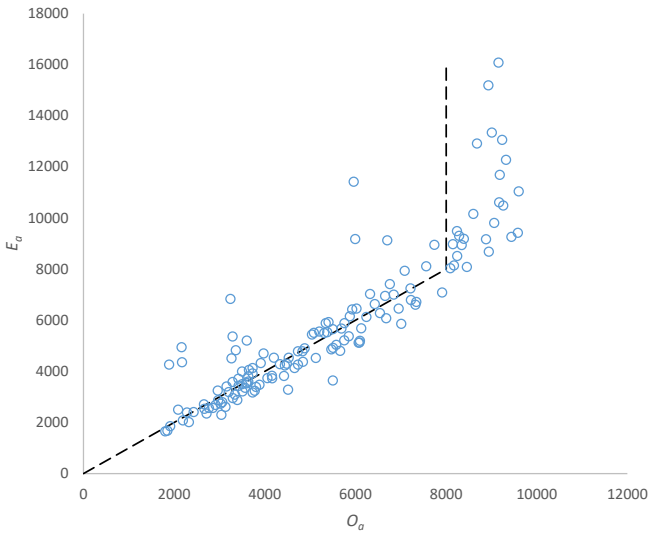
The peaks in a DNA profile are measured using fluorescence. The amount of fluorescence is proportional to the amount of DNA present. This fluorescence is captured by a camera. It is expected that as more DNA is added into a PCR the resulting peak height (measured in relative fluorescent units) in an 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 every allele within the stutter ratio dataset was calculated using the fomula:

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

Where (E_a) is the expected peak height calculated from the observed stutter height (O_{a-1}) and SR_a is the stutter ratio for allele *a* calculated using the equation described above. A plot of O_a versus E_a is provided in Figure 1. A vertical line at O_a = 8000 rfu indicates a common saturation limit for a 3130 instrument. The points should deviate from the *x* = *y* line at the saturation value. After inspection of Figure 1 we recommend a saturation threshold setting of 8000 rfu is applied.

Figure 1: Observed versus expected peak heights



Peak height variance and LSAE using Model Maker

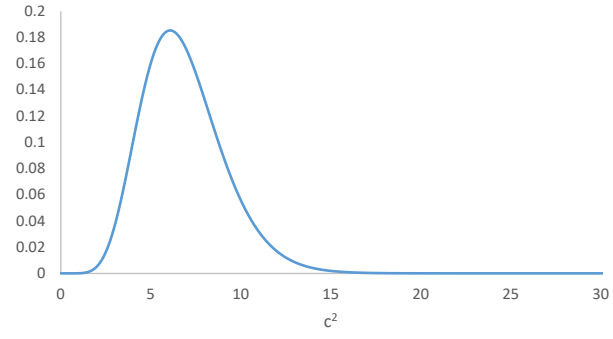
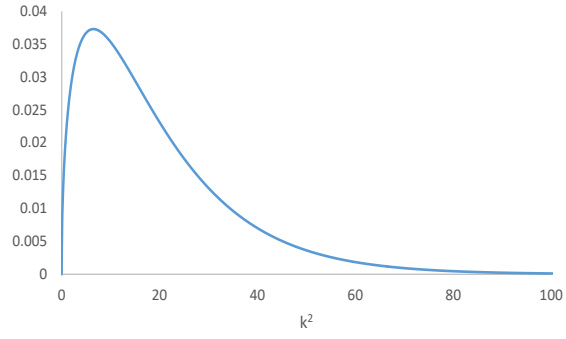
Empirical observations and experience suggests that profiles differ in variance (hereafter "quality"). Within STRmix™ the variability of peaks within profiles is described using a model containing a variance constant. Within V2.4 allele and stutter peaks have separate variances, c^2 and k^2, respectively. 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 OCME NY laboratory.

Single source profiles of varying quality run were analysed using the Model Maker function within STRmix™. A summary of the results for both c^2 and k^2 for the dataset is provided in Table 4. A plot of the allele and stutter gamma distributions are provided in Figure 2.

Table 4: Summary of Model Maker results for OCME NY dataset

Number profiles analysed	Allele variance parameters (Mode)	Stutter variance parameters (Mode)	Mean LSAE variance
147	gamma(9.1374,0.7472) (6.0799)	gamma(1.5007,12.9748) (6.4960)	0.0065

Figure 2: A plot of the allele and stutter gamma distributions for each dataset

Allele	Stutter
	

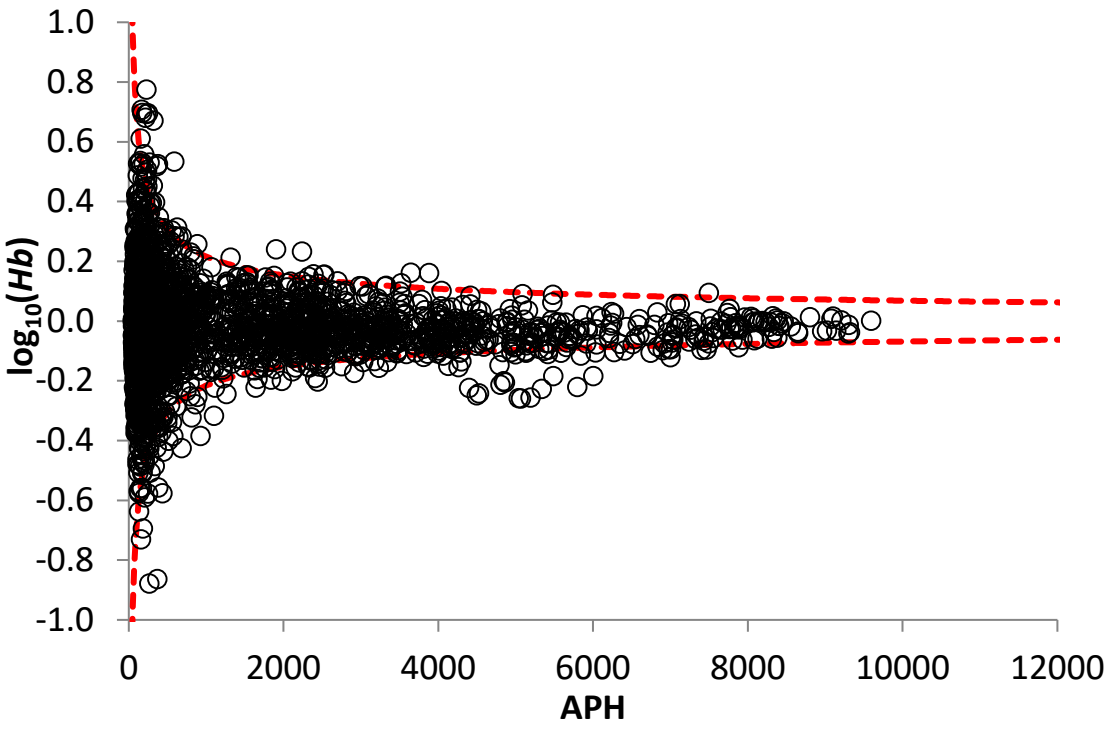
Heterozygote balance was calculated for all heterozygote loci for the Model Maker profiles. Heterozygote balance (*Hb*) was calculated as:

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

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.

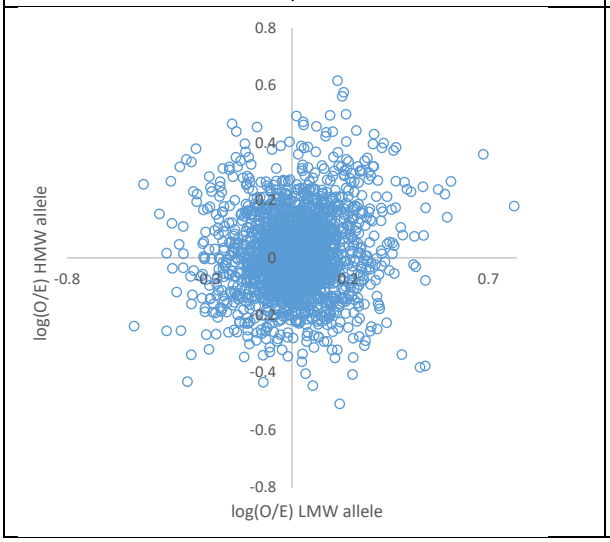
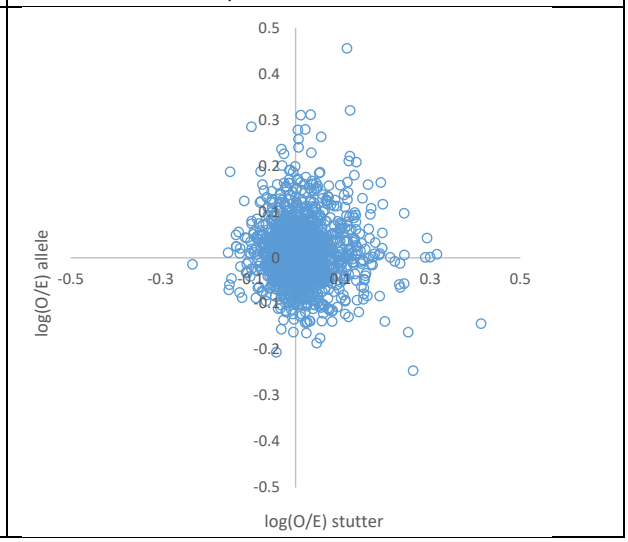
The plot of log*Hb* versus APH for the dataset described above and the expected 95% bounds (plotted as dotted lines) calculated at $\pm\sqrt{2}\times 1.96\times\sqrt{\frac{c^2}{APH}}$ where c^2 = 6.08, the 95th percentile from the gamma distribution from the combination data set. The 95% bounds encapsulate sufficient data as demonstrated in the graphs (coverage = 93.5%) demonstrating that the values for variance are sufficiently optimised. The plot in Figure 3 is an approximate check of Model Maker.

Figure 3: Log(*Hb*) versus APH for single source profiles



In Figure 4 we plot the correlation plots for LMW versus HMW allele and allele versus stutter peaks. 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 labelled at analysis however they do not affect the results.

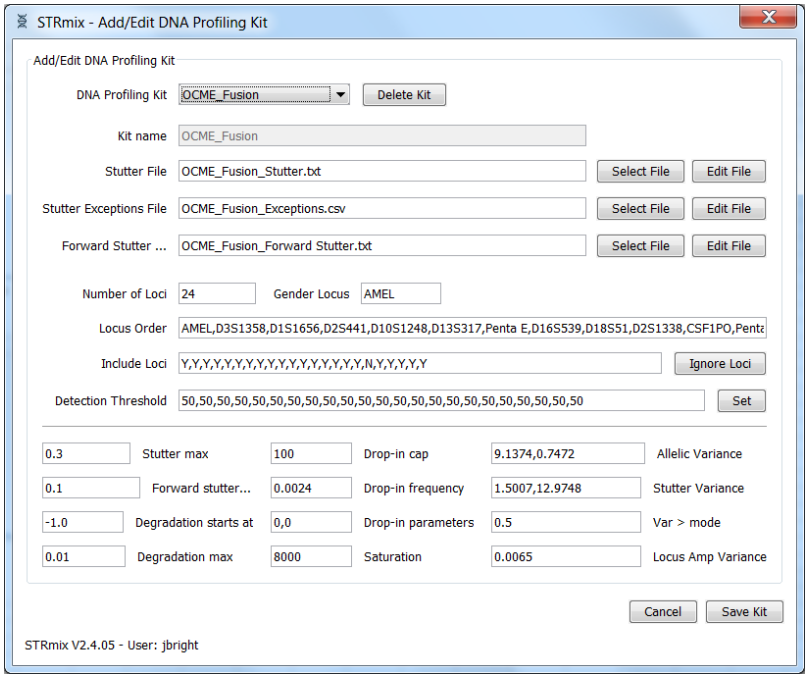
Figure 4: OCME NY Model Maker dataset correlation plots

HMW and LMW alleles, correlation = 0.0791	Allele and stutter, correlation = -0.0303
	

Conclusions

The recommended STRmix™ V2.4 default parameters for the interpretation of the OCME NY Fusion profiles run on a 3130 CE instrument are given in Figure 5.

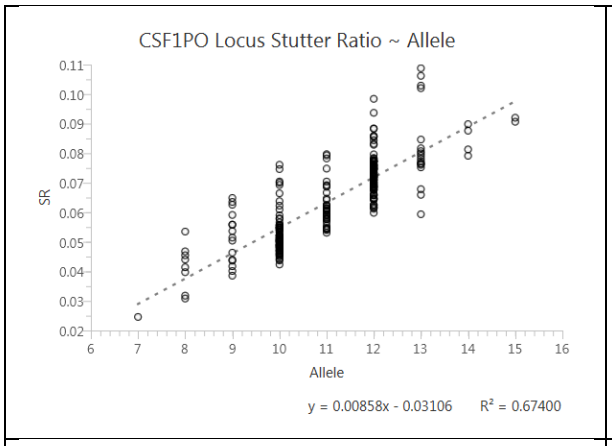
Figure 5: STRmix™ recommended default parameters for OCME NY Fusion profile interpretation

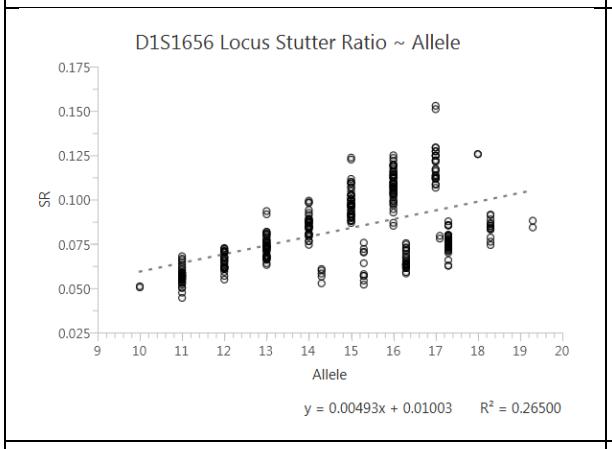


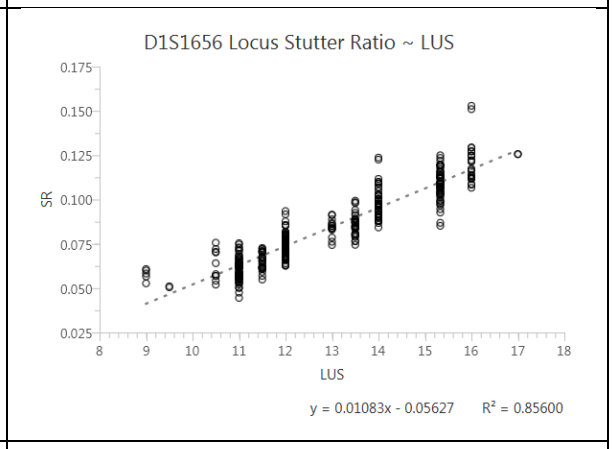
References

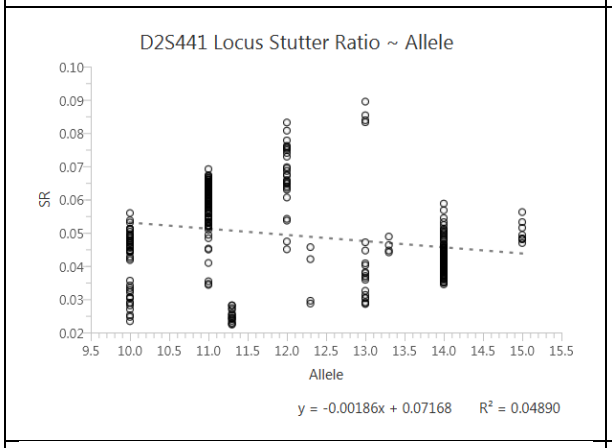
[1] Bright J-A, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Science International: Genetics. 2013;7:296-304.
[2] Brookes C, Bright J-A, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. 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. www.cstl.nist.gov/biotech/strbase Accessed 2014.
[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.

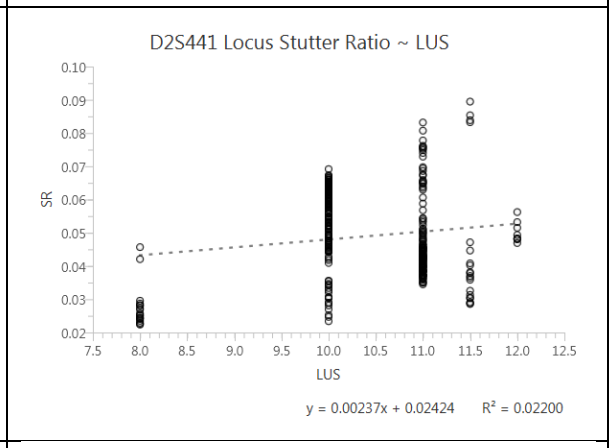
Appendix 1 Back stutter versus Allele and LUS

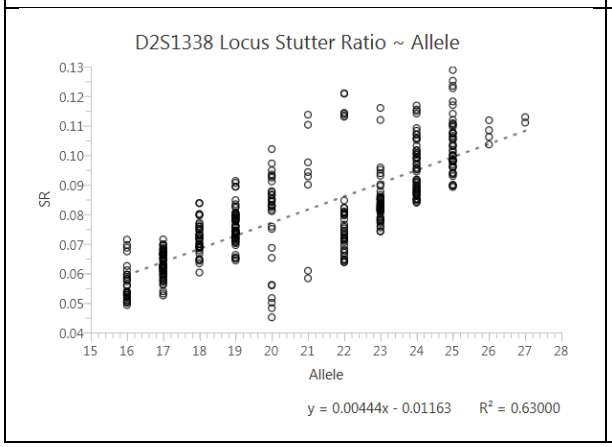


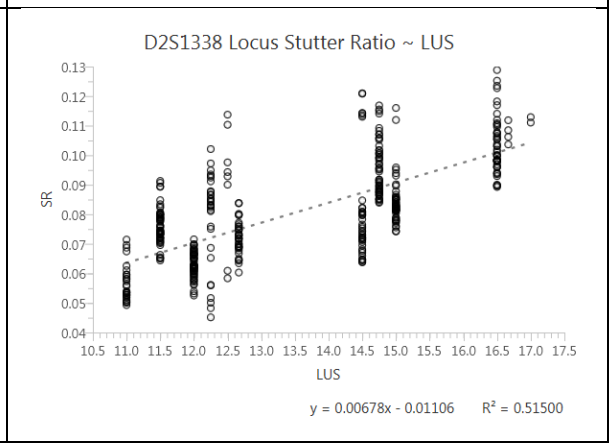


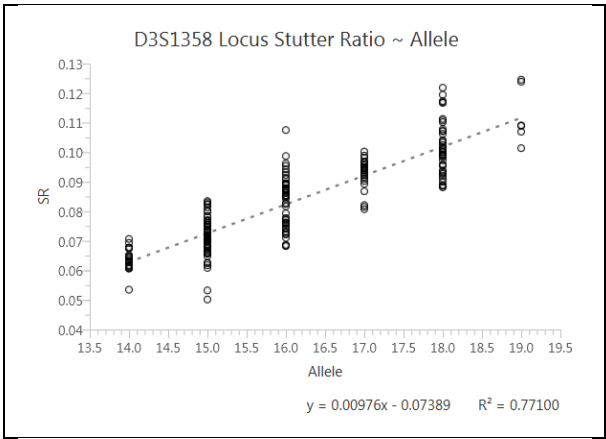


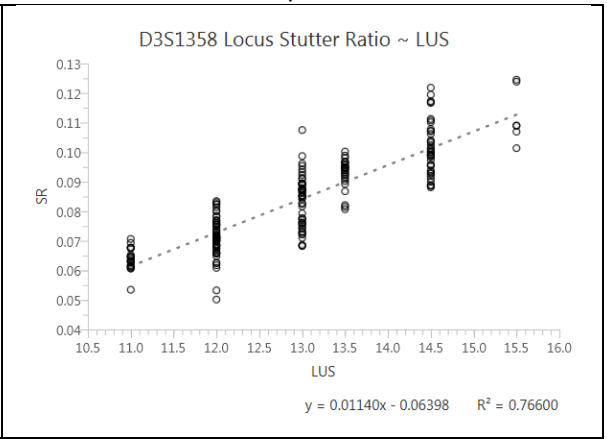


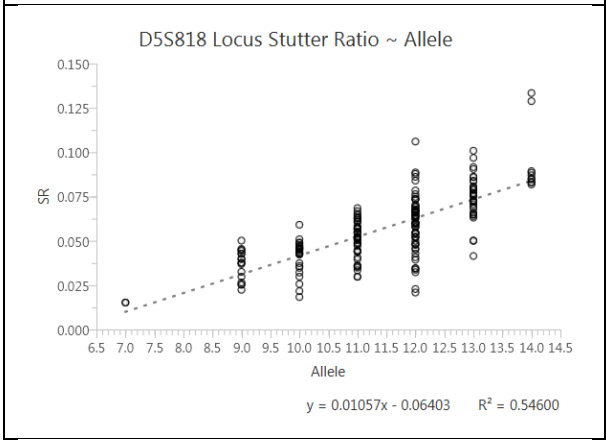


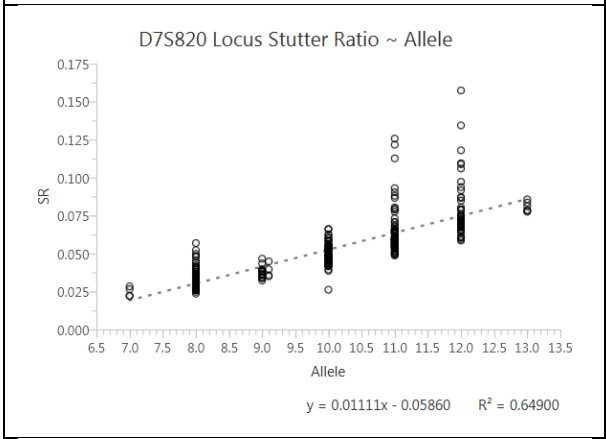


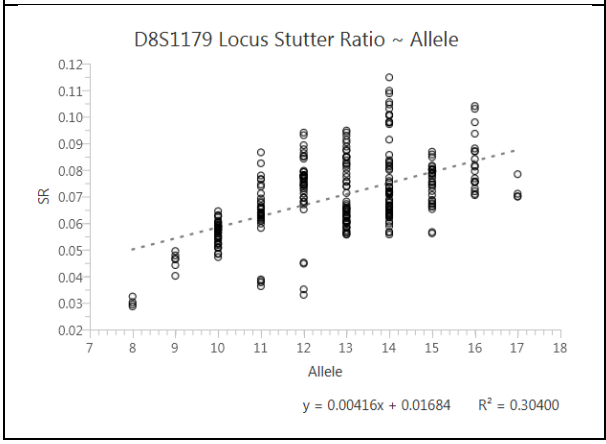


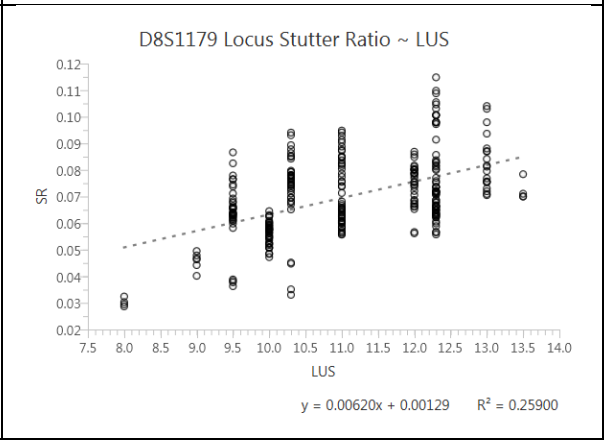

D3S1358 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00976x - 0.07389
R² = 0.77100

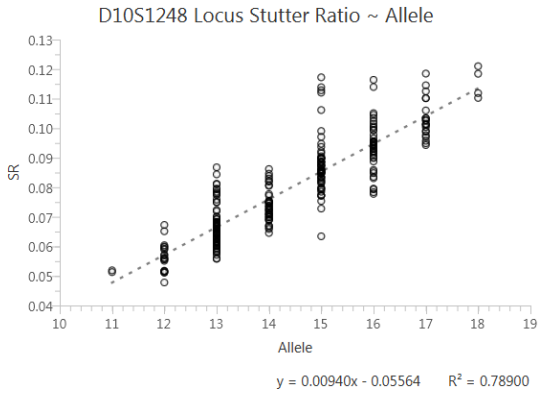

D3S1358 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.01140x - 0.06398
R² = 0.76600

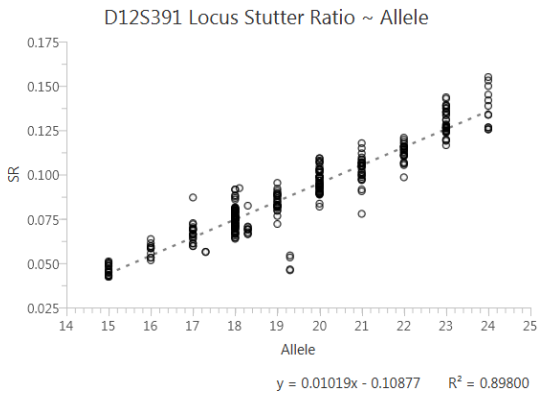

D5S818 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.01057x - 0.06403
R² = 0.54600

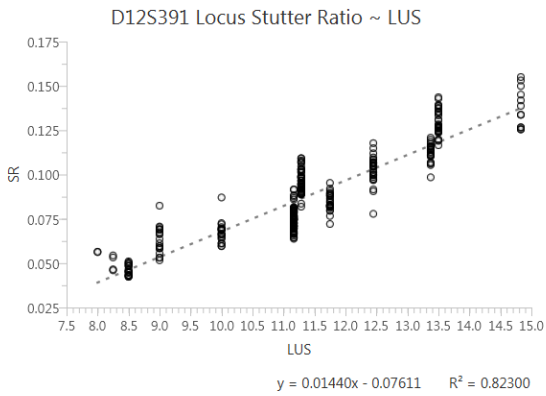

D7S820 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.01111x - 0.05860
R² = 0.64900

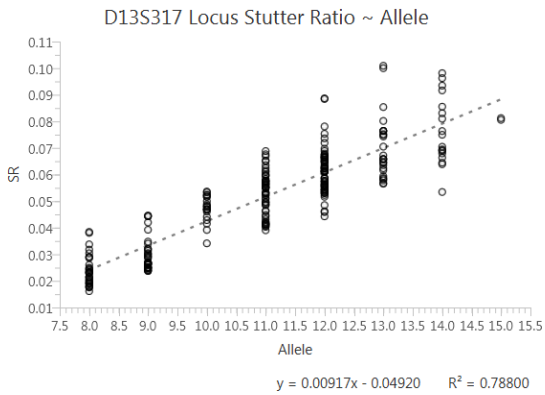

D8S1179 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00416x + 0.01684
R² = 0.30400

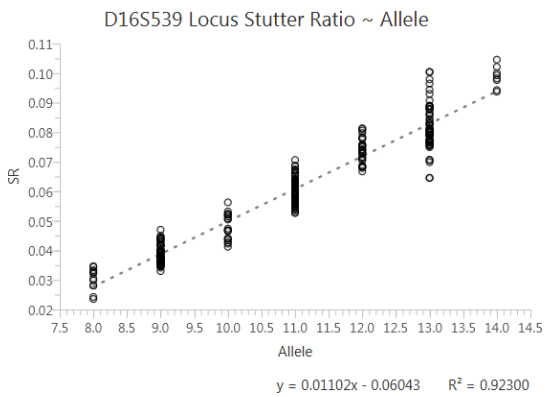

D8S1179 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00620x + 0.00129
R² = 0.25900

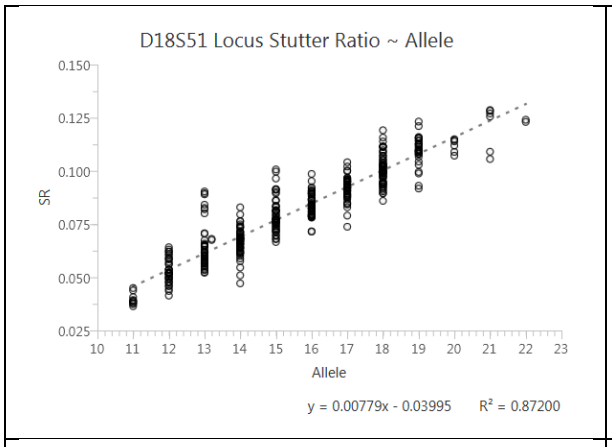

D10S1248 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00940x - 0.05564
R² = 0.78900

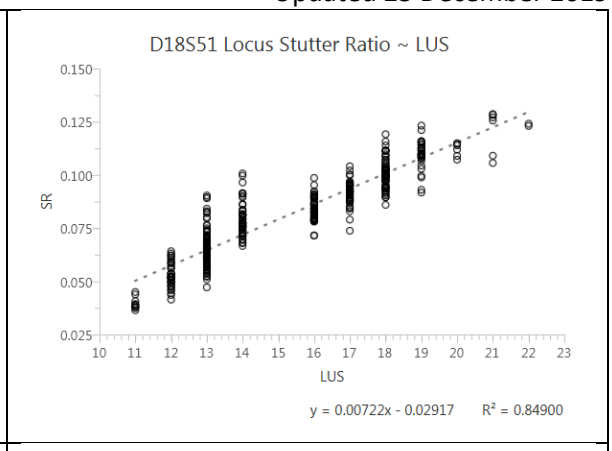

D12S391 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.01019x - 0.10877
R² = 0.89800

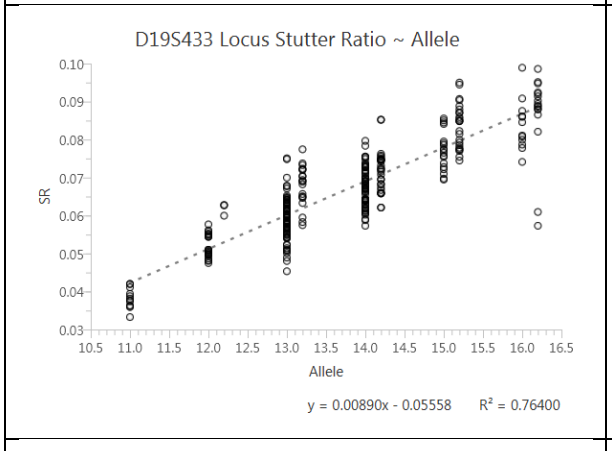

D12S391 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.01440x - 0.07611
R² = 0.82300

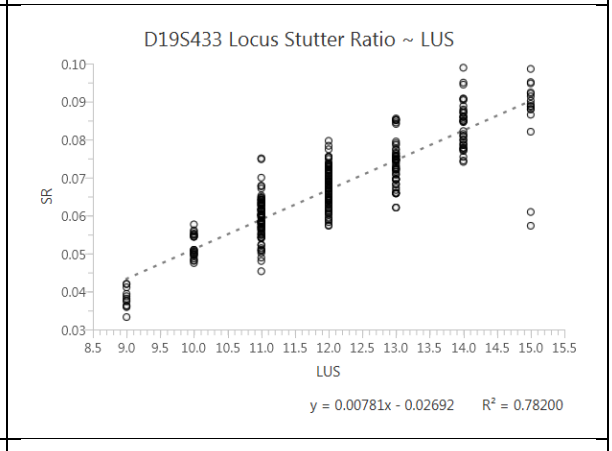

D13S317 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00917x - 0.04920
R² = 0.78800

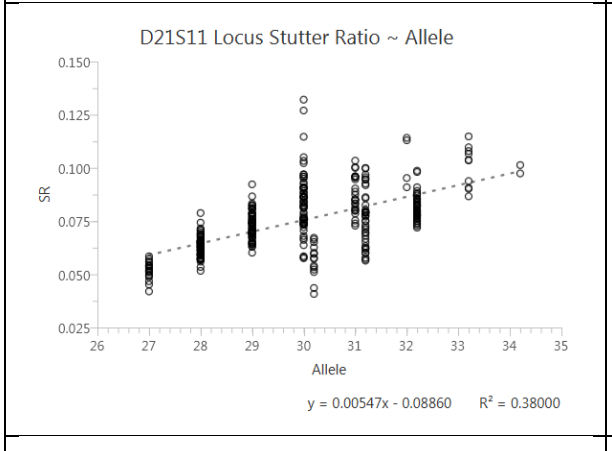

D16S539 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.01102x - 0.06043
R² = 0.92300

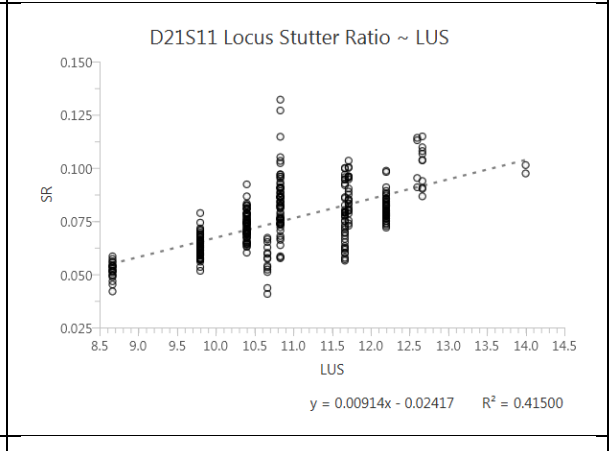

D18S51 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00779x - 0.03995
R² = 0.87200

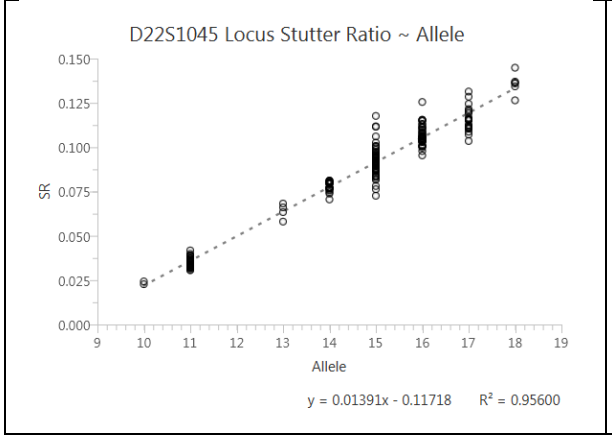

D18S51 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00722x - 0.02917
R² = 0.84900

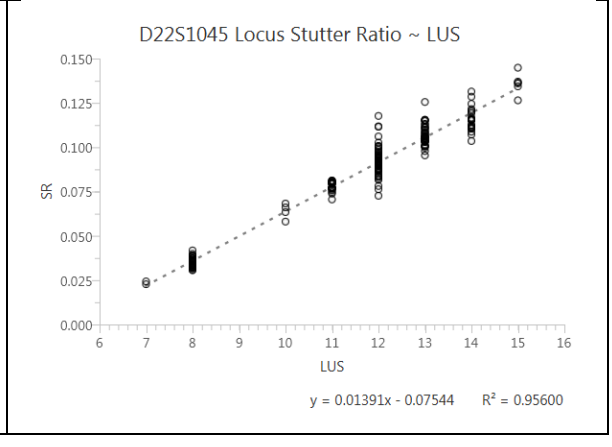

D19S433 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00890x - 0.05558
R² = 0.76400

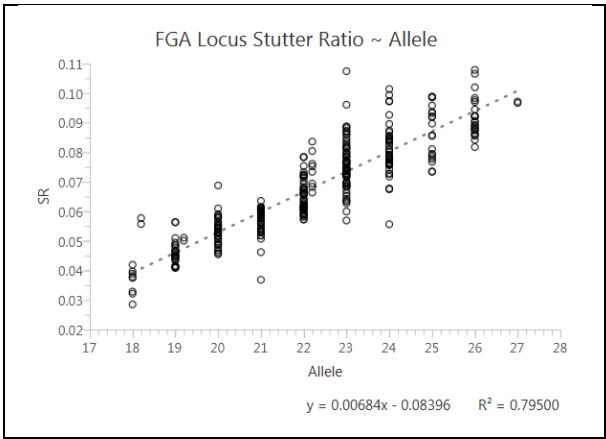

D19S433 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00781x - 0.02692
R² = 0.78200

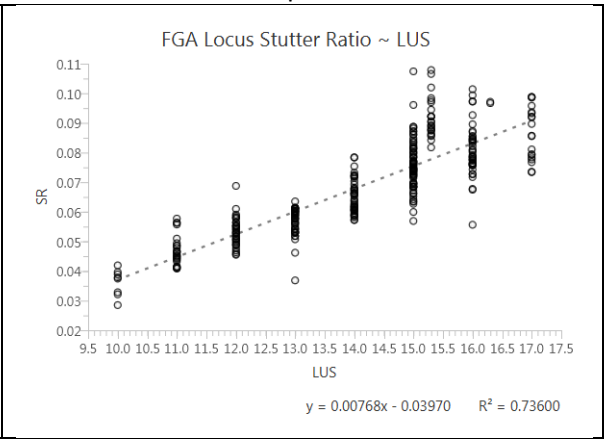

D21S11 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00547x - 0.08860
R² = 0.38000

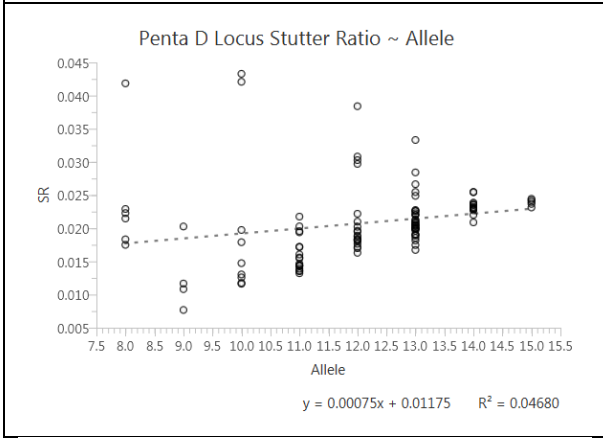

D21S11 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00914x - 0.02417
R² = 0.41500

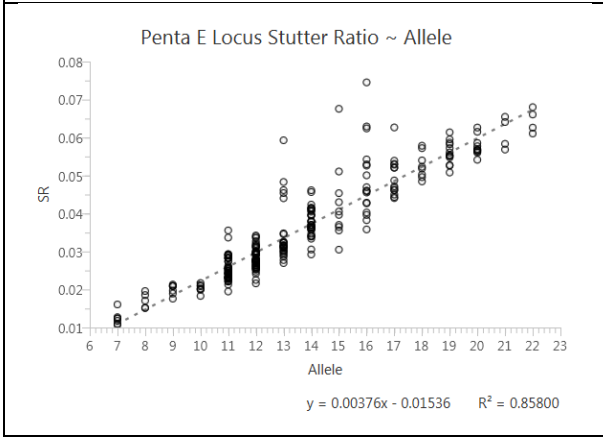

D22S1045 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.01391x - 0.11718
R² = 0.95600

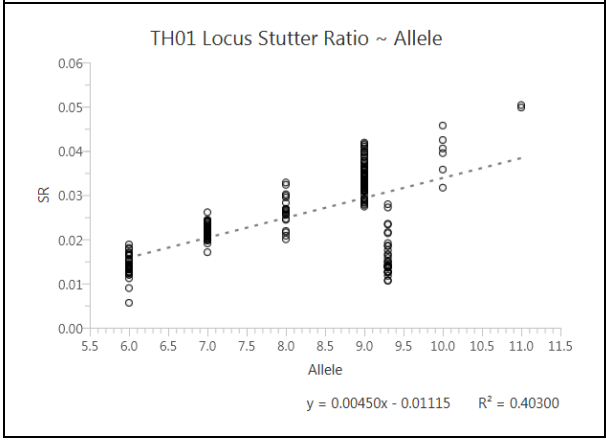

D22S1045 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.01391x - 0.07544
R² = 0.95600

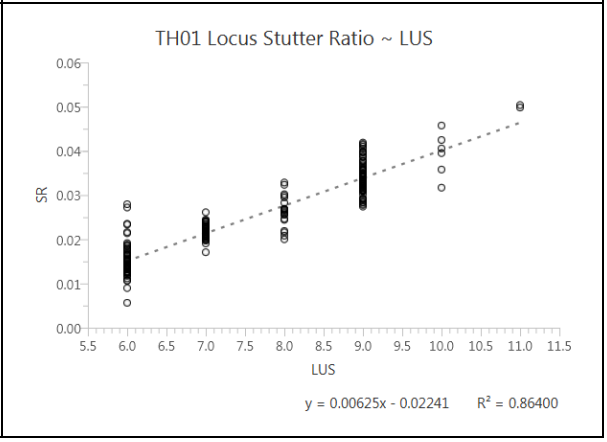

FGA Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00684x - 0.08396
R² = 0.79500

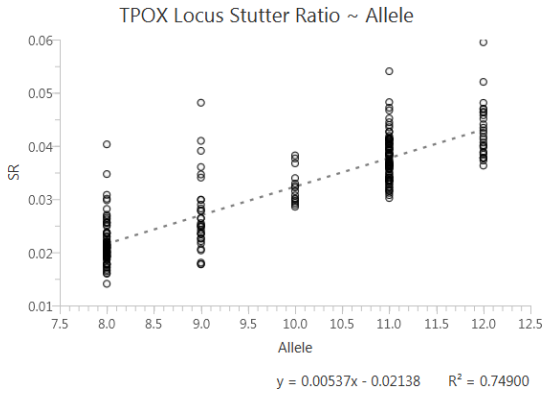

FGA Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00768x - 0.03970
R² = 0.73600

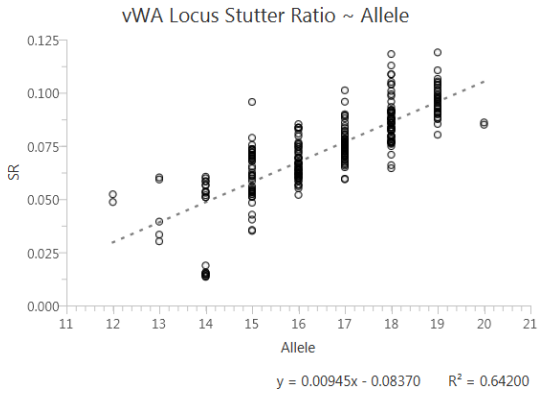

Penta D Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00075x + 0.01175
R² = 0.04680

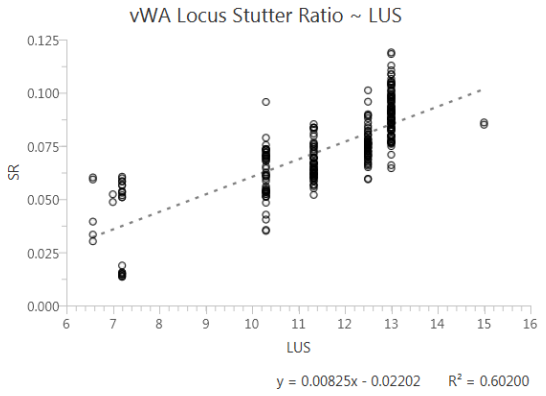

Penta E Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00376x - 0.01536
R² = 0.85800


TH01 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00450x - 0.01115
R² = 0.40300

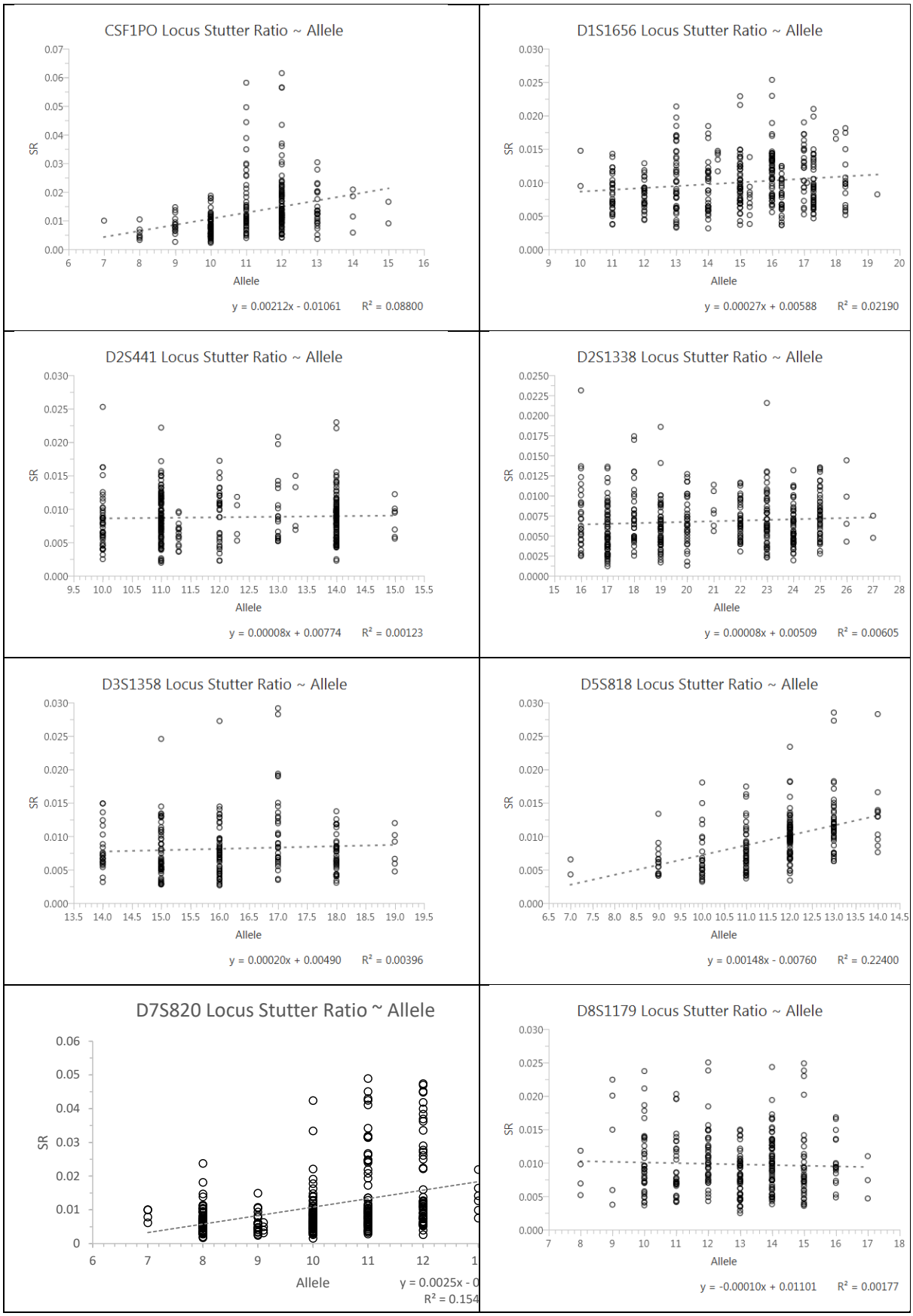

TH01 Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00625x - 0.02241
R² = 0.86400

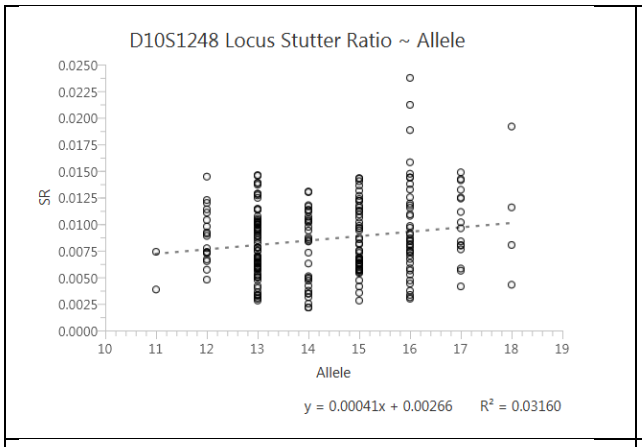

TPOX Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00537x - 0.02138
R² = 0.74900

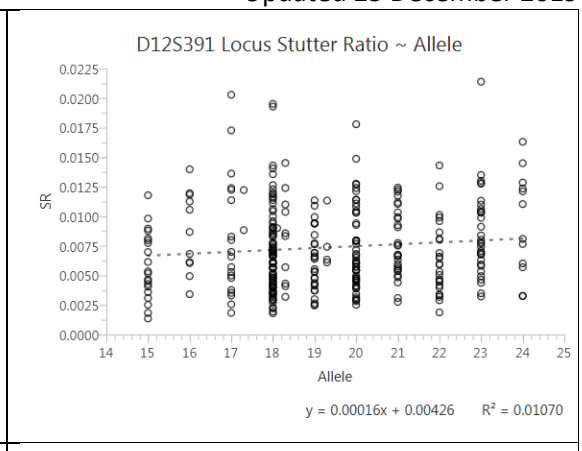

vWA Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00945x - 0.08370
R² = 0.64200

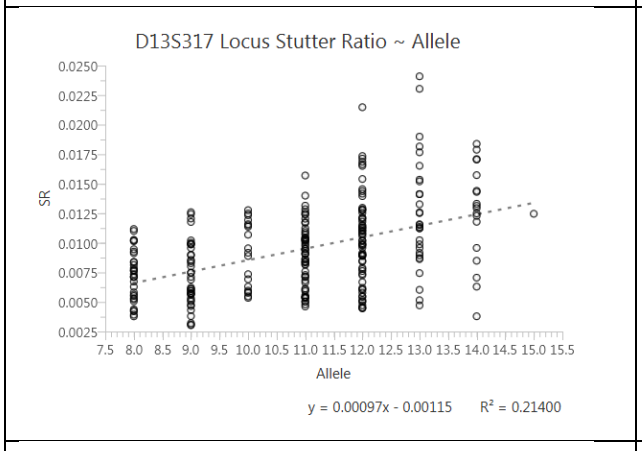

vWA Locus Stutter Ratio ~ LUS
SR
LUS
y = 0.00825x - 0.02202
R² = 0.60200

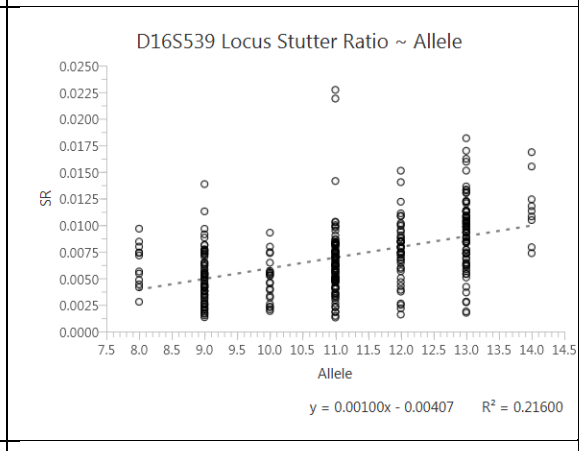
Appendix 2 Forward stutter versus Allele

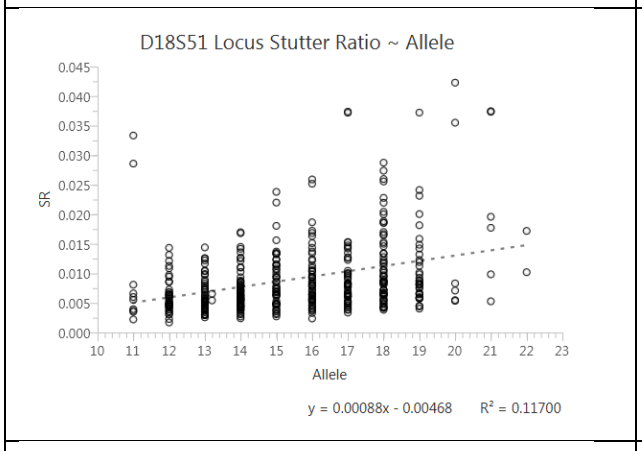


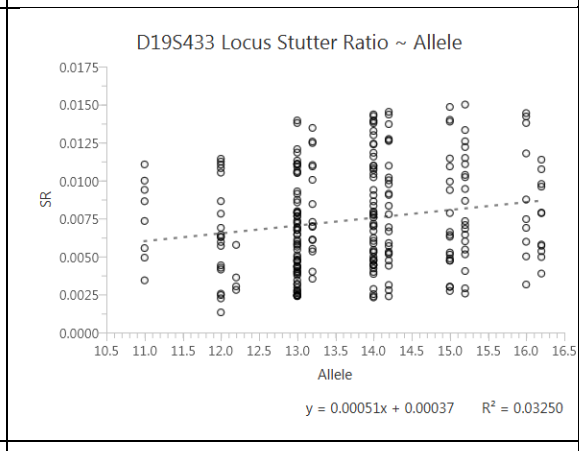

D10S1248 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00041x + 0.00266
R² = 0.03160

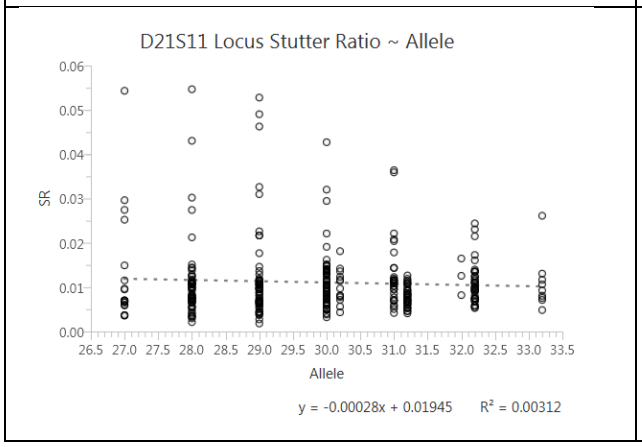

D12S391 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00016x + 0.00426
R² = 0.01070

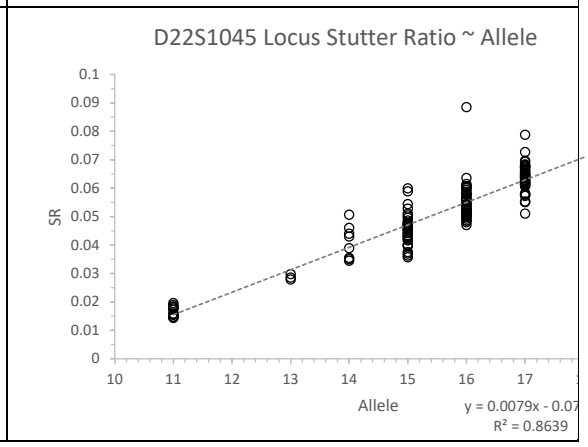

D13S317 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00097x - 0.00115
R² = 0.21400

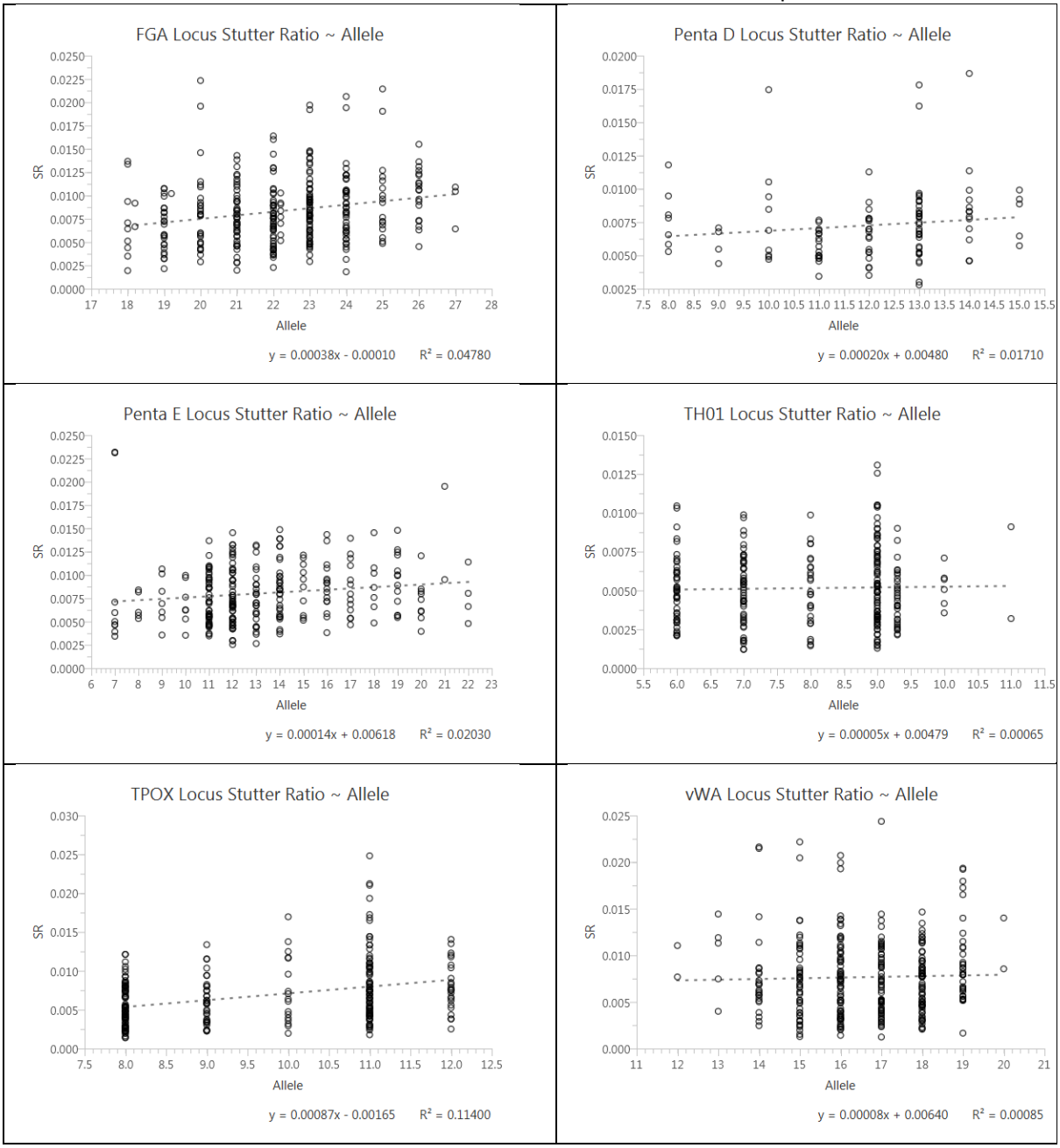

D16S539 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00100x - 0.00407
R² = 0.21600


D18S51 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00088x - 0.00468
R² = 0.11700


D19S433 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00051x + 0.00037
R² = 0.03250


D21S11 Locus Stutter Ratio ~ Allele
SR
Allele
y = -0.00028x + 0.01945
R² = 0.00312


D22S1045 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.0079x - 0.07
R² = 0.8639


FGA Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00038x - 0.00010
R² = 0.04780
Penta D Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00020x + 0.00480
R² = 0.01710
Penta E Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00014x + 0.00618
R² = 0.02030
TH01 Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00005x + 0.00479
R² = 0.00065
TPOX Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00087x - 0.00165
R² = 0.11400
vWA Locus Stutter Ratio ~ Allele
SR
Allele
y = 0.00008x + 0.00640
R² = 0.00085

Updates to summary as of 23 December 2019 (original version 2 August 2016)

The written summary for the Estimation of STRmix™ Parameters for OCME New York Laboratory was reviewed and updated in December 2019. These updates were made in order to correct transcriptional errors that were found. A summary of the updates is listed below.

Page 3: For the locus vWA listed in Table 2, the explanatory variable listed was changed from Allele to Average due to a transcriptional error.

Page 4: The reference to an Excel spreadsheet used to determine drop -in parameters was removed.

Page 6: Figure 3 on Page 6 was replaced with an updated version of the graph to include data points that had been inadvertently left out. Due to the update, the coverage mentioned within the paragraph above the graph changed from 97.2% to 93.5%.

Updates to this summary do not change any standard operating procedures. All NYC OCME standard operating procedures for the settings and use of the STRmix™ software remain the same after review and updates to this document.