TrueAllele® and STRmixTM: A Comparison of Two Probabilistic Genotyping Software Programs in Forensic DNA Profile Analysis

Orozco, Diana

TrueAllele® and STRmix^{TM}:
A Comparison of Two Probabilistic Genotyping Software Programs in Forensic DNA Profile Analysis

By

DIANA OROZCO
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Forensic Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Ruth E. Dickover, Chair

Ruth E. Ballard

Kent E. Pinkerton

Committee in Charge

2023

Acknowledgements

First and foremost, I want to thank the chair of my thesis committee, Dr. Ruth E. Dickover, for all the help and support in completing this project. I want to thank Dr. Ruth E. Ballard for allowing me the opportunity to take on this project. I also want to thank Dr. Kent E. Pinkerton for accepting the final role in this thesis committee. This study could not have been possible without all three of you.

Next, I want to thank all my friends and family for their endless support, love, and encouragement throughout the time spent on this project. More specifically to my mother and father, who have shown me to strive for what I believe in no matter the obstacles I face within the journey. I dedicate this thesis to all of you.

I want to thank all the professors who have taught me all I need to know to make it this far, and to my supervisors that continue to teach me as the forensics world is ever changing.

Finally, I want to thank Kern Regional Crime Laboratory, specifically Garett Sugimoto, for providing the data and sample files used in the study and answering all my TrueAllele® data questions. I also want to thank California State University, Sacramento, and Santa Clara County District Attorney's Crime Laboratory for providing the software programs used in the study.

TABLE OF CONTENTS

Abstract

Since its initial development in 1985, forensic DNA analysis has become increasingly important in casework evidence analysis. Forensic DNA typing is widely carried out today using sophisticated instrumentation and highly sensitive reagents. These developments have proven very beneficial to crime laboratories in solving criminal cases. However, an increase in forensic typing sensitivity can also lead to problems in interpreting the resulting profiles. DNA casework analysts are often confronted with complicated results including mixtures, degraded DNA, and/or low copy number DNA all of which are particularly difficult to interpret, deconvolute, and evaluate statistically. To make matters worse, there is no "one best way" to interpret challenging profiles that is agreed upon by the forensics community at large, and many labs rely on manual deconvolution techniques prone to human error.

A significant advancement in the analysis of forensic DNA profiles has been the development of complex computer algorithms for mixture deconvolution and/or statistical analysis in a largely automated fashion. Probabilistic genotyping (PG) software has come to the forefront of forensic DNA interpretation and is being used by more and more crime laboratories throughout the world today. There is now a wide variety of forensic DNA analysis programs available. However, no standardization exists, and crime labs may be unsure as to which program they should validate for use in casework analysis. Two popular DNA interpretation tools include the PG software programs, TrueAllele® and STRmix™. These programs use the Markov Chain Monte Carlo method (MCMC) to examine virtually every possible genotype contained in a DNA profile and to provide a statistical value as to the likelihood of each possible profile. Both programs process complicated mixtures more efficiently than manual binary methods, increasing the chances that the findings are robust, reproducible, and admissible in

court.

TrueAllele® and STRmix™ have both been validated by multiple crime laboratories and used in casework. However, in a court case, NY v Hillary (2016), STRmix™ excluded Hillary as a possible contributor, while TrueAllele® generated an inconclusive result after analyzing the same DNA profile. While TrueAllele® and STRmix™ both use the MCMC method, the difference in final outcomes suggests the programs may differ in sensitivity etc. Such a discrepancy, therefore, calls for a comparison study to better understand the programs, and to assist laboratories in making the best choice for their casework.

This study consists of a comparative MCMC analysis of thirty-six mixture sample profiles (Globalfiler®) that included from two to five contributors using TrueAllele® and STRmix™. These mixtures were originally processed using TrueAllele® as part of the Kern Regional Crime Laboratory's software validation study. The sample files were processed here using two versions of STRmix™, version 2.5 and 2.6. Although some individual interpretation requests produced different MCMC statistics across the three programs, no statistically significant differences were identified in this study. Rather, both the coefficient of determination and Analysis of Variance (ANOVA) statistic showed that the overall data were comparable between software programs when reporting mixture weight contributions and likelihood ratios (LRs) for mixture types of two to five contributors. Any significant differences were investigated and are discussed in detail.

List of Abbreviations

DNA	- Deoxyribonucleic Acid
STR	- Short Tandem Repeats
PCR	- Polymerase Chain Reaction
RFLP	- Restriction Fragment Length Polymorphism
VNTR	- Variable Number of Tandem Repeats
RFU	- Relative Fluorescence Unit
SWGDAM	- Scientific Working Group on DNA Analysis Methods
PG	- Probabilistic genotyping
RMP	- Random Match Probability
CPI	- Combined Probability of Inclusion
MCMC	- Markov Chain Monte Carlo
NOC	- Number of Contributors
LR	- Likelihood Ratio

List of Tables

Table:

List of Figures

Figure:

Introduction

Modern forensic DNA analysis largely involves purifying DNA from an evidence sample associated with a crime scene and developing a genetic profile consisting of highly variable short tandem repeats (STRs). STR markers are ideal in forensic and paternity DNA identification because they are highly polymorphic, have fairly even distributions of allelic forms, and are easily amplified by polymerase chain reaction (PCR) (Moretti, et al., 2001).

The technique of STR typing was an improvement over the early forensic DNA analysis technique of restriction fragment length polymorphism (RFLP) (Hammond et. al., 1994). RFLP uses restriction enzymes to cut DNA into smaller fragments of interest. The regions analyzed, referred to as variable number of tandem repeats (VNTRs), are areas in DNA that are highly repetitive and variable among individuals (Nakamura et. al., 1987), which makes them useful in human identification. Unfortunately, however, RFLP analysis can take weeks to accomplish and requires a large sample size, which is often not available in forensic casework (Butler, 2010). Over time, new lab assays and developments in instrumentation and software for analysis of the smaller STR regions helped increase the sensitivity and discriminatory power of DNA analysis techniques and decreased the assay time significantly (Hammond et. al., 1994). For this reason, STR typing dominates the field of forensic DNA analysis throughout the world today.

Currently, 15 to 23 polymorphic STR loci are routinely used for human identification. In a typical DNA testing kit protocol, all the loci are amplified by PCR using primers that flank the STR repeat sites (Wages Jr, 2005). The PCR primers add fluorescent tags to the allelic STR products during the amplification process for use in detecting the PCR products later. STR allele PCR products vary in base pair (bp) size according to the size of the locus and according to the number of repeats they contain. The STR alleles are detected through capillary electrophoresis in which the PCR products are separated by bp size while fluorescent signals are detected and quantified. The variety of colors of the fluorescent tags in combination with the size in bps allow for identification of specific STR alleles across loci. The resulting forensic DNA profile is depicted as a plot, called an electropherogram. The height of an allelic STR peak on an electropherogram is given in relative fluorescence units (RFUs) generated by the associated fluorescent PCR primer tags. The RFUs detected at each allelic peak on the electropherogram enable the analyst to identify the type and amount of each STR allele present in the DNA sample (Butler, 2010).

Once DNA sample evidence is analyzed in the lab, the STR results must be interpreted in order for the analyst to draw conclusions regarding who may or may not have contributed their DNA to the sample. According to the Scientific Working Group on DNA Analysis Methods (SWGDAM), an individual cannot be considered a potential contributor to the DNA profile unless a valid statistic providing a numerical assessment of the match is provided (SWGDAM, 2017). SWGDAM allows for the use of several different statistics to indicate the strength of a potential DNA profile match including those calculated using traditional, manual mathematical formulas. Due to the complexity of forensic DNA statistical calculations, manual statistical analysis of forensic DNA profiles is best suited to single-source (i.e., single-contributor) profiles, or single-source profiles derived from simple DNA mixtures through restriction calculations (Lynch and Cotton, 2018). When a sample contains DNA from a single source, the profile shows one to two allele peaks at each STR locus making it very easy to determine the STR profile of the contributor (Kelly, 2014). Basic single-source samples show one of two patterns of alleles at each locus: either a pair of peaks (heterozygous) of roughly the same height or a single peak (homozygous) approximately double the height of one heterozygous peak. These peak heights remain fairly consistent throughout the electropherogram, allowing straightforward interpretation of results (Butler, 2004). Single-source DNA profiles typically provide reliable results that can be accurately analyzed using manual statistical methodology.

Unfortunately, forensic DNA casework often involves samples from excessively handled evidence items, and/or items exposed to harsh environments. Samples obtained from such items often include mixtures of DNA derived from multiple people, degraded DNA, and/or low copy number DNA (Lee et. al., 1998). Before probabilistic genotyping (PG) software programs became available, DNA analysts had to deconvolute DNA mixtures manually based primarily on STR peak height data from evidence electropherograms. SWGDAM guidelines allow for manual statistical calculations for some DNA mixture profiles, however, the mixtures must be relatively simple with few contributors and sufficient allele RFUs for interpretation. If one or more contributors to a DNA mixture is present at a much higher signal level than others, the profile may undergo deconvolution using restriction calculations to separate major contributor alleles from minor contributor alleles for manual statical analysis (SWGDAM, 2017). These major and minor components may then be used to provide simpler profiles for interpretation and statistical analysis (SWGDAM, 2017).

Unfortunately, restriction calculations are not always able to separate mixed profiles into clearly separate components, or the resulting component profiles can still be too complicated for manual statistics. A sample containing a roughly equal mixture of two or more people can make deriving the individual profiles for each contributor particularly difficult because the peaks at a locus may overlap. A single locus from a mixture can show more than two STR peaks, but the possible genotypes of the various contributors may not be clearly distinguishable (Torres and Sanz, 2003). If the major and minor components are not distinguishable, the analyst cannot determine genotypes eligible for a single-source statistic such as the Random Match Probability (RMP) statistic. Instead, the analyst may calculate less powerful mixture statistics such as the Combined Probability of Inclusion (CPI) using the frequency of each allele as measured in reference populations (Buckleton and Curran, 2008). As the number of contributors to a DNA profile increases, the specificity of the DNA statistic that can be calculated, and hence the resulting probability, decreases. In other words, manual statistics typically provide decreased strength of inclusion probability as the number of contributors to the mixture increases (Perlin et. al., 2014). Another difficulty in interpreting STR profiles from more than one contributor, is that individuals providing a minor portion of the total sample DNA may not have detectable alleles at all STR loci. For these reasons, interpreting forensic DNA mixtures can be very difficult and the results may not be eligible for any manual statistics, threatening the value of these profiles in court. Therefore, manual calculations of match probabilities on a mixture with three or more contributors remain difficult and often generate inconclusive statistical results (Perlin et. al., 2014).

In addition to forensic DNA mixtures, analysts can also receive evidence samples that experienced poor environmental conditions, sometimes for long periods of time. Harsh environmental conditions promoting DNA degradation include exposure to ultraviolet light and/or high temperatures (Butler, 2010). Degraded DNA samples typically show a characteristic decrease in allelic peak height RFU's particularly for longer STR loci on the electropherogram. As the degree of DNA sample degradation increases, the associated allelic peak RFU's decrease, which then typically increases the number of peaks below the stochastic threshold. This can be an issue in profile interpretation because if the peak heights are too low, there is potential for allelic loss or “dropout” on the electropherogram resulting in the inability to develop a full DNA profile (Graw, 2000).

To add to the challenges confronting forensic DNA analysts, there is no standardized manual interpretation procedure for difficult STR profiles that the forensic community at large has agreed upon. Instead, most labs rely on their own manual deconvolution procedures, which are lengthy and can be prone to human error (Dror and Hampikian, 2011). Fortunately, powerful computing tools have been developed to help forensic DNA analysts interpret and calculate statistics on complex STR profiles. TrueAllele® and STRmix^{TM} are two of many software programs available to assist laboratories in determining individual contributor genotypes and computing DNA match statistics for complex mixtures (Kadash et al., 2004; Moretti et al., 2017). Cybergenetics, a bioinformation company based in Pittsburgh, Pennsylvania, is the developer of TrueAllele®. The Institute of Environmental Science and Research Ltd. (ESR; New Zealand) and Forensic Science South Australia (Australia) developed STRmix^{TM}. Both software programs were developed using probabilistic statistical analysis methods to separate potential individual genotypes from complex forensic mixtures including degraded and low copy number DNA samples. TrueAllele® and STRmix^{TM} use the powerful Markov Chain Monte Carlo (MCMC) method to determine mixture weight proportions in evidence samples, and most importantly to examine virtually every possible genotype combination across all loci and provide a statistical value to the likelihood of each possible profile (Perlin, et al., 2013). For interpretation, both programs first deconvolute the DNA evidence profile mixtures and then compare any given reference profiles to the mixture derived genotypes (Perlin, et. al., 2011). Additionally, both programs model out stutter and require the analyst to assume the number of contributors (NOC) before interpreting the mixtures.

TrueAllele$^{®}$ and STRmix^{TM}

While TrueAllele$^{®}$ and STRmix^{TM} are the main PG tools used by crime laboratories in the United States today, there are general differences between the two software programs. Beginning with TrueAllele$^{®}$, this software program analyzes the raw data obtained from the capillary electrophoresis stage in the form of .fsa STR genotype profiles and does not use an analytical threshold. When using TrueAllele$^{®}$ an analyst may remove artifacts prior to interpretation, but this is not obligatory. TrueAllele$^{®}$ requires the user to manually assume a total NOC prior to running the MCMC analysis. When comparing references to the derived genotypes profiles, TrueAllele$^{®}$ will calculate a likelihood (LR) for each contributor in a mixture. For example, if a DNA evidence profile is a mixture of at least three contributors, then a single reference will result in three LR values. At the time of this study, results from TrueAllele$^{®}$ were reported in the base 10 logarithm of the LR. Updated versions now allow for view of the results in either base 10 logarithm notation or in scientific notation. In base 10 logarithm notation, anything larger than 0 provides support of inclusion, and anything less than 0 provides support of exclusion. Finally, the creator of TrueAllele$^{®}$, Mark Perlin, states that an attribute of scientific uncertainty is that a match probability cannot be stated as a definite zero when based on real data (Perlin and Sinelnikov, 2009). Therefore, an exclusion will always be some number less than 0.

STRmix^{TM} requires the raw data first be manually processed and edited considering analytical thresholds. It then requires generating a data table for upload to the program. STRmix^{TM}, therefore, requires more initial analyst editing before a profile can be interpreted, resulting in a faster run-time. Similar to TrueAllele$^{®}$, STRmix^{TM} also requires the analysts to enter the assumed NOC prior to running the MCMC analysis. However, STRmix^{TM} will produce an error message if it suspects a NOC was set too low (JCRCL, 2019). When comparing references to the derived genotype profiles, STRmix^{TM} will only report one LR for each comparison request. STRmix^{TM} reports LRs in scientific notation, where an inclusion is anything larger than 1 and an exclusion is anything less than 1. Most exclusions on STRmix^{TM} are reported as an absolute 0 if a reference profile does not compare to a derived genotype simply based on allele calls. If the alleles in a proposed genotype cannot fully exclude a reference, then it will consider peak heights and provide a LR value.

STRmix^{TM} and TrueAllele$^{®}$ have been validated by crime laboratories in the US and are currently being used in casework. As MCMC processing is inherently variable, laboratories typically run each profile several times and report only those results that are concordant across runs. No data exists yet, however, on how similar the results would be if obtained using these two different MCMC programs for the same DNA profile. Therefore, the purpose of this study is to investigate differences in the final results in analyses of the same samples using these two programs. The hypotheses evaluated in the study will consists of:

Hp: The performance of TrueAllele$^{®}$ and STRmix^{TM} are not significantly different.

Hd: The performance of TrueAllele$^{®}$ and STRmix^{TM} are significantly different.

To answer this, DNA profile data will be analyzed to determine any correlation between the statistical values derived by each program for a given sample, as well as the differences in number of inclusions, exclusions, and undetermined results across programs. False positives, false negatives, and the effect of assuming an incorrect NOC will also be compared across both programs. These results will be analyzed statistically to evaluate the sensitivity, specificity, reproducibility, and repeatability across the two software programs.

This study will provide important information to crime laboratories on the effectiveness of the two individual programs as well as the repeatability of MCMC probabilistic forensic genotyping analysis in general to help them decide which, if any, PG system they want to validate. Vital information on the limitations of PG will also be evaluated to inform DNA analysts of the need to qualify MCMC statistics appropriately during court testimony.

Materials and Methods

Probabilistic Genotyping Software

Kern Regional Crime Laboratory (KRCL) evaluated TrueAllele® Server version 3.25.5840.1 and TrueAllele® Visual User Interface for easy review (VUIer) version 3.3.6228.1 to perform a MCMC method for forensic STR analysis on casework-type mixture profiles as part of their validation (Kern Regional Crime Laboratory, 2017) (Curran, J.M., 2008). The same DNA profiles were analyzed using STRmix™ version 2.5.11, provided by California State University at Sacramento for this study. Amid data analysis, STRmix™ released a new version. Therefore, the same procedure was performed using STRmix™ version 2.6.0 provided by Santa Clara County District Attorney's Crime Laboratory.

TrueAllele® Validation Data

In 2017, the KRCL completed their validation study of the TrueAllele® Casework PG system for the interpretation of forensic DNA profiles generated with the Globalfiler® PCR Amplification Kit (Gouveia, N. et.al., 2015), a 6-dye STR amplification kit (Kern Regional Crime Laboratory, 2017). Thirty-six mixture samples were prepared using multiple contributors. The mixture profiles were derived from samples made with varying levels of known template DNA. Additionally, the composition of some mixtures was randomized to mimic typical evidence profiles encountered in forensic DNA casework. The mixture samples were amplified for 28 cycles, and detection was performed using an Applied Biosystems® 3130 Genetic Analyzer (Connon, C. C., et.al., 2016) with a capillary electrophoresis injection time of 5 seconds. The results were interpreted using TrueAllele® (Kern Regional Crime Laboratory, 2017). KRCL provided the resulting DNA profiles and the corresponding raw data used for their validation to perform this comparative research study.

Forensic DNA evidence profiles analyzed included four different mixture types which ranged from two-person to five-person. Each mixture type included nine different evidence samples. The evidence samples were numerically labeled beginning with 32 and ending with 67. The reference samples were also numerically labeled 1 through 10, thus, keeping the DNA sources unknown until the analysis stage. Table 1 summarizes the DNA samples and the corresponding numerical labels designated for each mixture profile type.

Evidence Profile Labels	Mixture Type	Count
32-36, 52-55	2-person	9
37-41, 56-59	3-person	9
42-46, 60-63	4-person	9
47-51, 64-67	5-person	9

Table 1 *A summary of the numerical labels and count for each mixture type analyzed.*

Evidence samples included a range of mixture weight proportions for all types of mixtures. The first set of samples (samples 32 through 51) were made from five known reference samples selected at random. Furthermore, the mixture weight set-up for each contributor in a sample was determined using a uniform distribution computed by Dirichlet sampling (Perlin et. al., 2015). Samples 43, 44, 46, and 47 each included one unknown contributor.

The second set of samples (52 through 67) were made from seven known reference samples, which included known relatives of DNA contributors to the study. The known relatives consisted of full brothers (references 4 and 5), and a father and son (references 4 and 6). Two mixture samples for each mixture type included known relatives.

STRmix^{TM}

As required by STRmix^{TM}, the electropherograms generated by the 3130 capillary electrophoresis for each mixture profile were checked for artifacts by a fully trained analyst from KRCL. All the peaks appearing artifactual were removed from the data prior to analysis. The final .fsa data files were then exported in table format using GeneMapper® IDX software (Applied Biosystems™) with the stutter filter off. The exported mixture sample data tables included rows for each locus and columns for up to twelve alleles followed by columns for the allele base-pair lengths, or "sizes", and columns for allele heights in RFU. The exported known reference sample data tables consisted of rows for each locus and columns for two alleles.

All mixture sample interpretations were completed on both versions of STRmix^{TM} using the default MCMC setting (8 chains of 100,000 burn-in accepts with 50,000 post burn-in accepts per chain). Mimicking the KRCL validation of TrueAllele® as closely as possible, each mixture profile was run against one reference profile at a time on the STRmix^{TM} software platform with the NOC set to the known value, n, for each profile. LRs were assigned using three FBI extended population databases (Caucasian, African American, and Southwest Hispanic) with the most conservative LR reported. Each request was made in duplicate or until two resulting LRs were obtained within 2 log-fold of each other as required by KRCL for concordance. The process was then repeated specifying incorrect NOC, n-1 and n+1.

Data Sets

Between all the resulting concordant LRs for each combination, the most conservative LR result was tabulated into a final data set of 360 data points. These same criteria were used to select the final concordant LRs from the TrueAllele® validation study data for comparison with STRmix^{TM} results.

All resulting LRs were reported in the base 10 logarithm notation. A LR equal to or greater than 4.00 was considered a positive result, or indicative of a reference's inclusion as a possible source of DNA in the mixture profile. A LR equal to or less than -4.00 was considered a negative result, or indicative of a reference's exclusion as a possible source of DNA in the mixture profile. LRs between -4.00 and 4.00 were considered inconclusive as to whether or not a reference could be a possible source of DNA in the mixture profile.

The website for STRmix^{TM} includes minimum recommendations for computers based on expected sample sizes. For example, for up to three-person mixtures (sometimes four-person mixtures) the website recommends a minimum of a two-core computer processor with at least a 4 gigabyte (GB) random access memory (RAM). For up to four-person mixtures (sometimes five-person mixtures) it recommends a minimum of an eight-core computer processor with at least a 128 GB RAM.

STRmix^{TM} Version 2.5 was installed on a computer with a 4-core processor and was only able to run two- to four-person mixtures, decreasing its expected 360-point data set to only 270 data points for the correct NOC, n, and decreasing even further to 180 data points for the over-assumed NOC, n+1. When comparisons involved version 2.5 of STRmix^{TM}, the appropriate four-person and five-person mixtures from the other software programs were not included in the comparisons.

Additionally, STRmix^{TM} Version 2.6 was installed on a 6-core computer processor and was only able to run two- to four-person mixtures for the over-assumed NOC, n+1, decreasing its data set to only 270 data points.

As mentioned, STRmix^{TM} will produce an error message if it detects the NOC was set too low and will not move forward with the comparison until the NOC is adjusted. Therefore, this feature of STRmix^{TM} lowered the number of data points when the NOC was under-assumed, n-1.

Empirical Examinations

Each PG program was challenged using multiple mixture types, varying mixture proportions, and inputting the correct and incorrect NOC to compare the programs in the aspects of precision, accuracy, sensitivity, specificity.

Since MCMC is known to have some variability when reporting LRs (Bright, et. al., 2015), precision, or the ability to reproduce the same or similar results, of each program was determined by requesting duplicate runs under each program until the concordant results were obtained. The average number of requests made to reach two concordant LR ratios for every combination was calculated for each program. Furthermore, averages of LR ranges for each comparison request were calculated for each program. Precision was examined only for interpretation requests set at the true NOC since the TrueAllele$^{®}$ validation did not focus on concordance when under-assuming or over-assuming NOC.

The final most conservative concordant LRs tabulated for each program were statistically compared to one another. First, each program was statistically tested for normalcy using the Jarque-Bera test. If the LR results were found to be normally distributed for all three programs, the LRs per program were then compared to one another, two programs at a time, using a one-way Analysis of Variance (ANOVA) test. ANOVA tests were used to determine if there are any statistically significant differences between the results provided by the programs. This process was then repeated on separate lists of LRs from known contributors and LRs from known non-contributors.

All 36 evidence-type mixture samples were prepared at KRCL targeting specific DNA mixture weight contributions. The observed mixture weights from each program were compared to one another by calculating the range difference in reported mixture weight proportions between all three programs and then averaging the difference. The same process was repeated on the observed mixture weight proportions from all three programs *and* the expected mixture weight proportions based on the original mixture sample set-up.

Each program's accuracy, or ability to correctly include and exclude a reference profile, was assessed by using the lists of known contributors and known non-contributors for each mixture type and comparing the count of true and false conclusions between each program. A subset of the data was compiled for LR's that resulted in different conclusions between the three programs for the same request. For example, if the request for a mixture sample compared to reference 1 under all three programs resulted in both STRmix^{TM} programs excluding the reference as a contributor to the mixture, but TrueAllele$^{®}$ resulted in an inconclusive LR, then the results for this request under all three programs were added to the data subset. Once all the appropriate results from requests with different conclusions between the programs were added to the subset, a range in difference between LRs was determined and an average difference was calculated. Any difference between LRs larger than 10 was investigated further by comparing LR results, MCMC settings, and/or LR ranges.

Accuracy was further assessed by examining the effect of allele sharing on LR results

(Cheng, 2021). For known contributors, allele sharing ratios were calculated by comparing one known contributor's alleles to the alleles of the other known contributor(s) in the same mixture sample. The total number of alleles possible for a reference was divided by the highest number of alleles the reference shared with any other reference within the same mixture sample. For known non-contributors, allele sharing ratios were calculated comparing alleles from a known non-contributor to the alleles detected in a mixture sample. The total number of alleles detected in a mixture sample were divided by the number of alleles a known non-contributor shared with the mixture sample.

The sensitivity of each program, or the ability to correctly conclude the presence of a contributor when presented in various DNA amounts, was assessed by comparing the mixture weight proportions of all known contributors for each mixture profile to its corresponding resulting LR under each program. The mixture weight proportions for each contributor were evaluated to determine the lowest amount of DNA where each program can reliably conclude the presence of a known contributor. Sensitivity was further assessed by examining the first five sample mixtures with the lowest average RFUs to see the effect on the resulting LRs, if any.

The specificity, or ability to correctly exclude a known non-contributor, was assessed for each program by comparing the resulting LRs to the list of known non-contributors for each sample.

As mentioned, each program was additionally challenged by entering an under-assumed NOC, n-1, and an over-assumed NOC, n+1, for all comparison requests. The effect of the incorrect NOC was assessed by creating a dataset of results where the resulting conclusion changed from the correct NOC, n, dataset. In other words, if a comparison request concluded the reference was "included" under the correct NOC, but "inconclusive" or "excluded" under the incorrect number of contributors, then the results for this comparison request were added to the compiled dataset. Once complete, counts were tabulated and averages of change in LR values were calculated.

Results

Data Sets

As mentioned previously, STRmix^{TM} will produce an error message if it detects the NOC was under-assumed. This was a problem for mixture sample 32, a two-person mixture.

Mixture sample 32 was a robust sample and had a high number of elevated stutter peaks (average peak height per loci was 8596 RFUs, the highest of all samples). Therefore, both STRmix^{TM} programs interpreted this sample as a mixture of 3 contributors, and produced an error message when the NOC was set to the true value of 2 contributors. In order to move forward with this sample under the true NOC, the sample was reanalyzed in GeneMapper ID-X with the stutter filter on. This removed most of the stutter peaks and allowed STRmix^{TM} to run the sample with NOC set to 2. The LRs from sample 32 were the only LRs derived from data with the stutter filter on and used to move forward with comparisons between programs.

Concordance

Figure 1 shows the count of interpretation requests made per sample per program to achieve concordance. For two- to four-person mixtures, an average of 2.5 interpretations were requested on STRmix^{TM} version 2.5 to reach concordance per DNA mixture sample. An average of 2.6 interpretations were requested on STRmix^{TM} version 2.6 to reach concordance per DNA mixture sample. An average of 4.1 interpretations were requested on TrueAllele$^{®}$ to reach concordance per DNA mixture sample. For 5-person mixtures, an average of 3.1 interpretations were requested to reach concordance for STRmix$^{\text{TM}}$ version 2.6 and an average of 2.9 interpretations were requested for TrueAllele®.

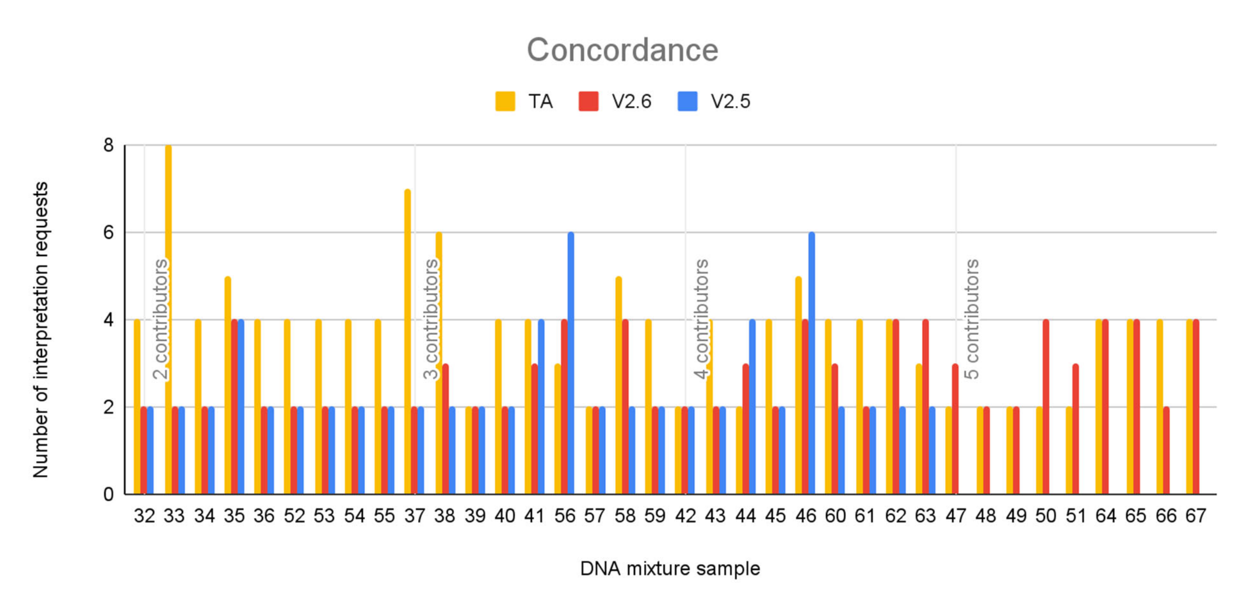


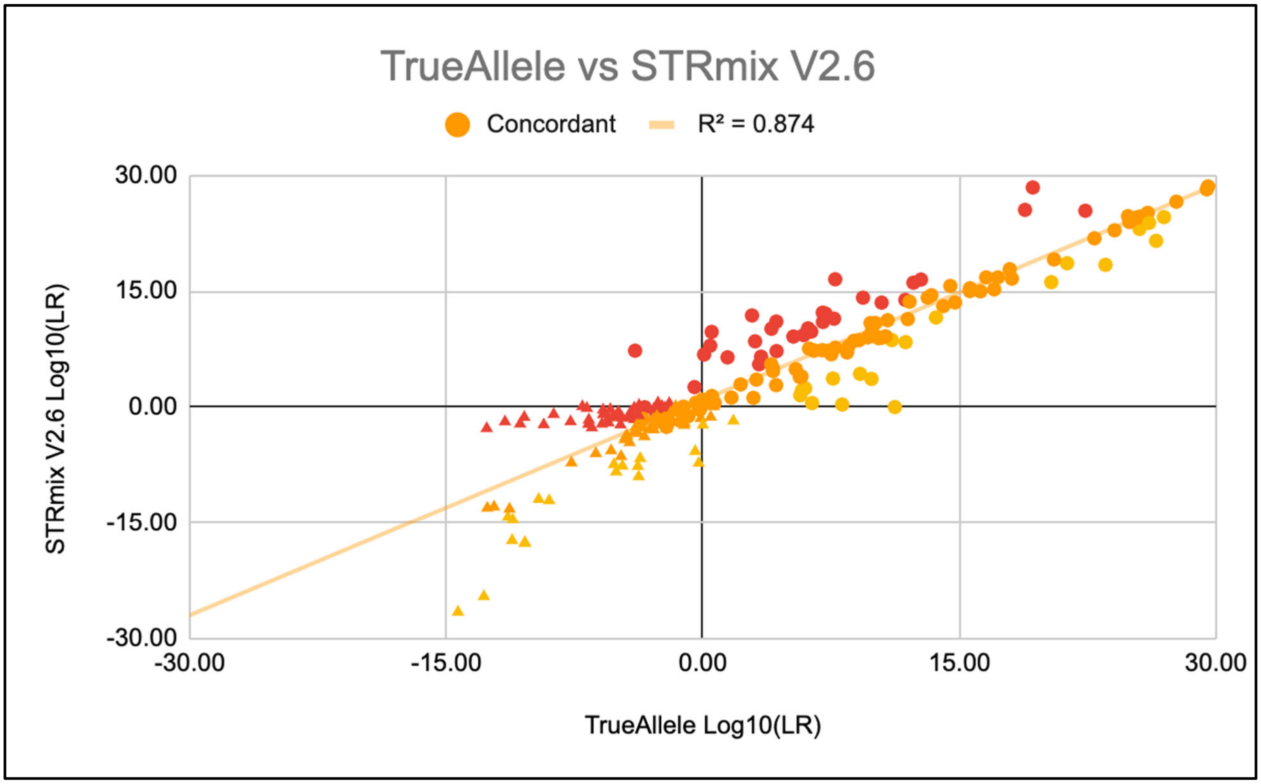
Figure 1. *A column chart with counts of interpretations requested for each mixture under each software program.*

Mixture Types	**TrueAllele®**	**STRmix$^{\text{TM}}$ V2.6**	**STRmix$^{\text{TM}}$ V2.5**
All	3.7778	2.7222	2.5185
2 to 4	4.0741	2.5926	2.5185
5	2.8889	3.1111	N/A

Table 2. *A summary of the average number of interpretations requested to reach concordance for each software program.*

Jarque-Bera and ANOVA Tests

Figure 2 provides a visual representation of how the LRs between two programs compare to one another. The Jarque-Bera test found the LRs from two-person mixtures for both STRmix$^{\text{TM}}$ versions, three-person mixtures for STRmix$^{\text{TM}}$ version 2.6, and five-person mixtures for TrueAllele® were normally distributed. The remaining LRs resulted in p-values below the significance level, 0.05. Therefore, a one-way ANOVA test was not performed on all mixture-type data at once. Instead, the mixture-type data was separated into subsets of known contributors and known non-contributors. The data for LRs from known contributors for each mixture type per program were found to be normally distributed. The one-way ANOVA tests revealed there was no significant difference in the reported values between the programs as all p-values were greater than the alpha level of 0.05. Lastly, the LRs for known-non contributors were normally distributed for three-person mixtures under TrueAllele®, and for five-person mixtures under STRmix™ version 2.6. ANOVA tests were not performed for LRs from known non-contributors.



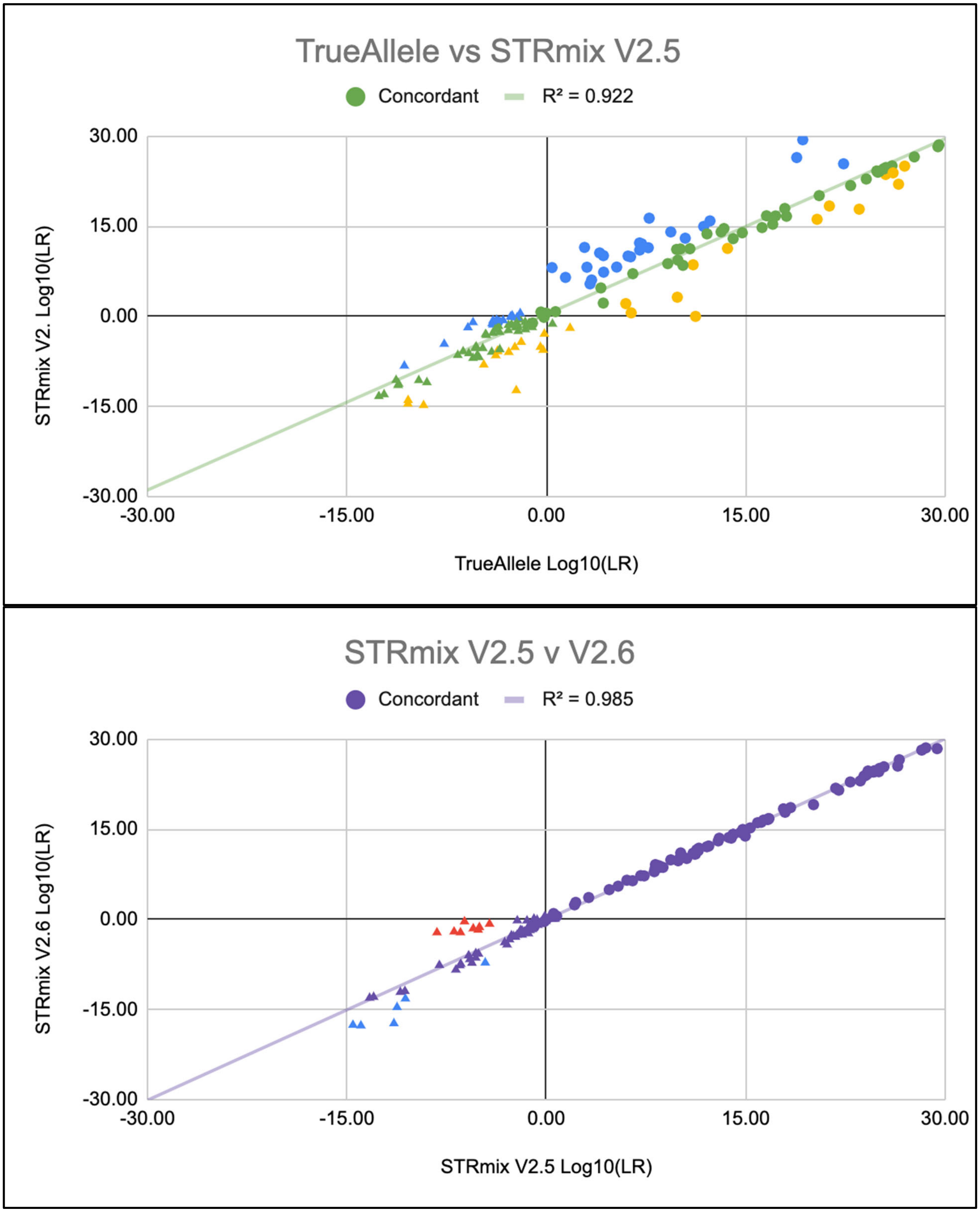


Figure 2. *Scatter plots comparing LRs between two PG programs at once. LRs are in circles for known contributors and triangles for known non-contributors. LRs larger under TrueAllele® are yellow. LRs larger under STRmix™ v2.6 are red. LRs larger under STRmix™ v2.5 are blue. All other colors represent LRs concordant between the programs (within a 2-log ban).*

Mixture Weights

Figure 3 shows a scatter plot of the observed mixture weights as determined by the programs as they relate to the expected mixture weights. When the observed mixture weights were compared between programs, all three programs produced a coefficient of determination trendline, R^2, near 0.7. The range in difference was from 0.00 (0%) to 0.21 (21%) with an average difference of 0.04 (4%). The observed mixture weights from each program were then compared to the expected mixture weights and the range in difference was from 0.00 (0%) to 0.73 (73%) with an average difference of 0.11 (11%). Since the observed mixture weights are more consistent between programs, all further empirical examinations involving mixture weights were conducted using the observed mixture weights of the respective program.

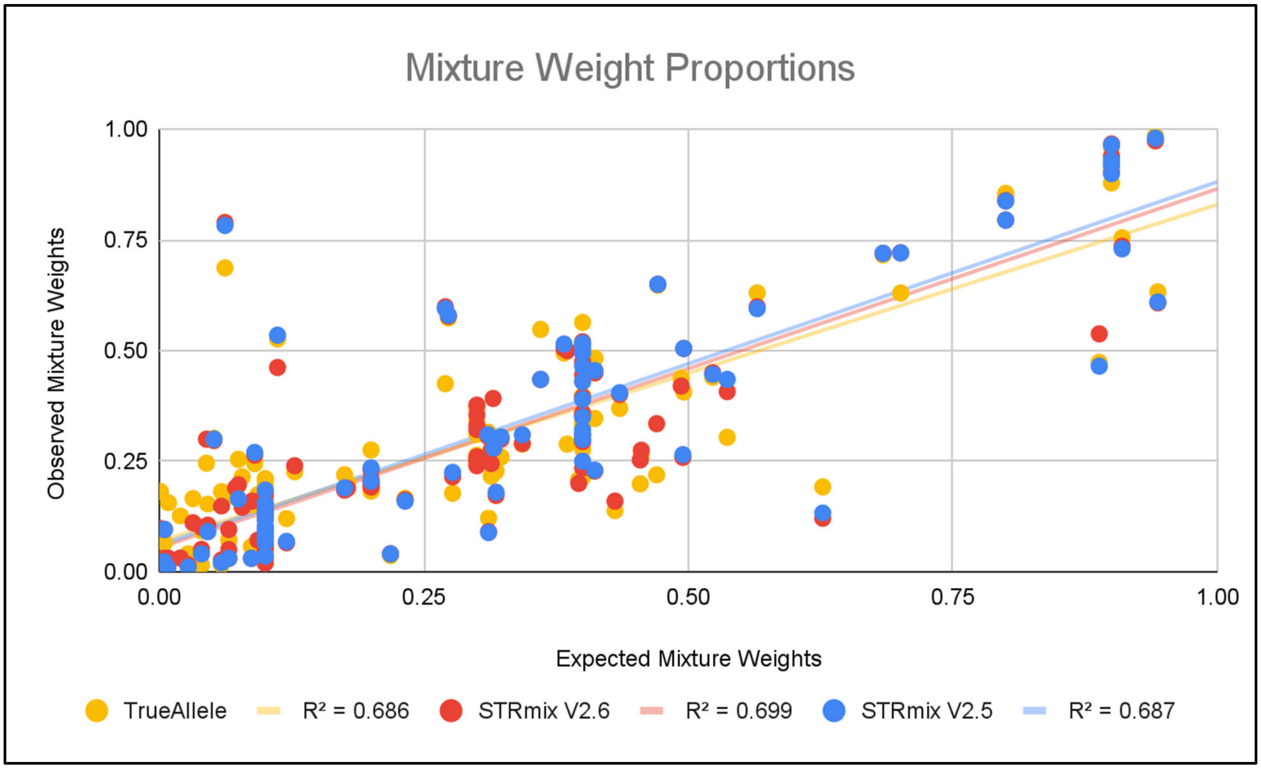


Figure 3. *A scatter plot of expected versus observed mixture weight contributions with a coefficient of determination trendline (R^2) for each software program.*

Accuracy

Table 3 tabulates the number of true conclusions, inconclusive results, false conclusions, and occurrences of no results. Out of 270 total conclusions for two- to four-person mixtures, STRmix^{TM} version 2.5 resulted in 217 true conclusions, 52 inconclusive results, and 1 false conclusion. STRmix^{TM} version 2.6 resulted in 211 true conclusions, 58 inconclusive results, and 1 false conclusion. TrueAllele® resulted in 202 true conclusions, 68 inconclusive results, and no false conclusions. Out of 90 total conclusions for five-person mixtures, STRmix^{TM} version 2.6 resulted in 27 true conclusions and 63 inconclusive results. TrueAllele® resulted in 50 true inclusions and 40 inconclusive results. These results are visually represented in figure 4a, known contributors, and 4b, known non-contributors.

2- to 4-person mixtures (5-person mixtures)						
Conclusion	**STRmix^{TM} V2.5**		**STRmix^{TM} V2.6**		**TrueAllele®**	
True (+)	67 (0)	25% (0%)	67 (26)	25% (29%)	66 (29)	24% (32%)
True (-)	150 (0)	56% (0%)	144 (1)	53% (1%)	136 (21)	50% (23%)
INC (+) known contributor	7 (0)	3% (0%)	7 (14)	3% (16%)	7 (8)	3% (9%)
INC (-) known contributor	3 (0)	1% (0%)	3 (4)	1% (4%)	5 (7)	2% (8%)
INC (+) known non-contributor	4 (0)	1% (0%)	6 (5)	2% (6%)	4 (1)	1% (1%)
INC (-) known non-contributor	38 (0)	14% (0%)	42 (40)	16% (44%)	52 (24)	19% (27%)
False (+)	0 (0)	0% (0%)	0 (0)	0% (0%)	0 (0)	0% (0%)
False (-)	1 (0)	0% (0%)	1 (0)	0% (0%)	0 (0)	0% (0%)
No Data	0 (90)	0% (100%)	0 (0)	0% (0%)	0 (0)	0% (0%)

Table 3. *Number of conclusions for two- to four-person mixtures (five-person mixtures) per software program.*

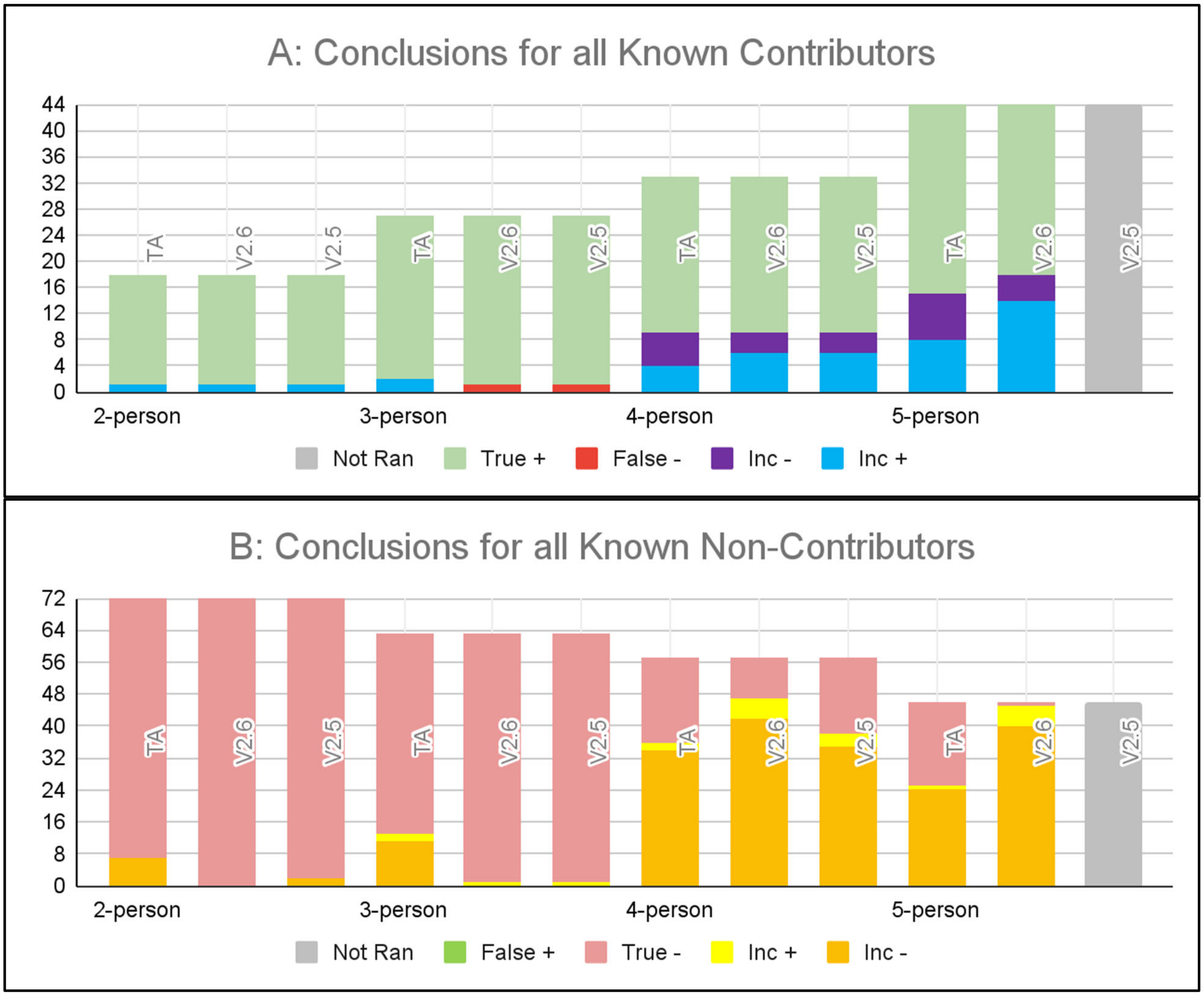


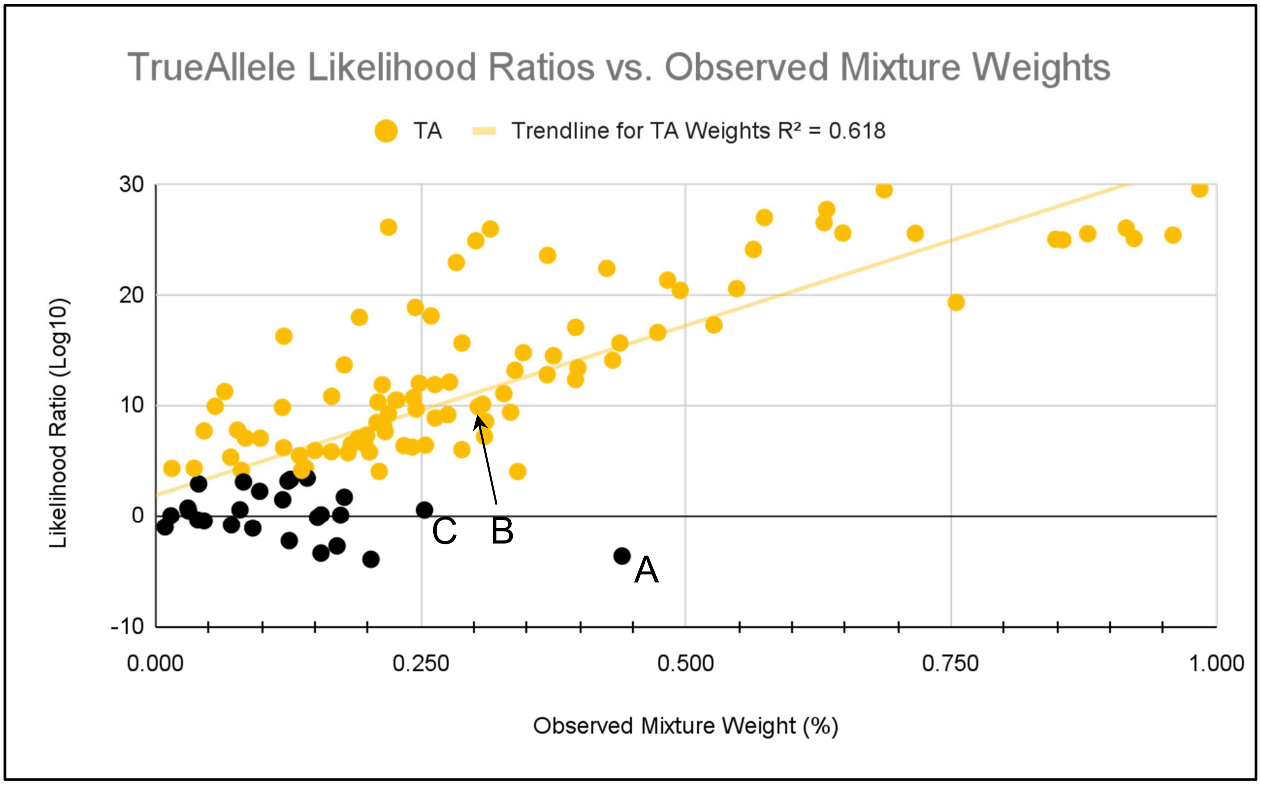
Figure 4 (a and b). *Stacked column chart for all a) known contributors and b) known non-contributors.*

The dataset of LR's that resulted in different conclusions between programs for the same request was assessed. For known contributors, 20 of the same requests had different conclusions between the three programs. The average range difference in LR (log10) was 5.63, with the minimum difference as 1.79 and the maximum difference as 11.26. For known non-contributors 56 requests had different conclusions across the three programs. The average range difference in LR (log10) was 4.43, with a minimum difference of 0.12 and a maximum difference of 12.31.

All the data was assessed to determine the effect of allele sharing on the resulting LRs for each program. Table 4 summarizes the coefficient of determination, R^2, results from scatter plots for known contributors and known non-contributors per software program. The R^2 results were as low as 0.001 and up to 0.012.

	TrueAllele®	**STRmix^{TM} V2.6**	**STRmix^{TM} V2.5**
Known Contributors	0.012	0.001	0.006
Known Non-Contributors	0.001	0.006	0.007

Table 4. *Coefficient of determination, R^2, values for known contributors and known non-contributors for each program when graphing the relationship between allele sharing ratios and resulting LR values.*



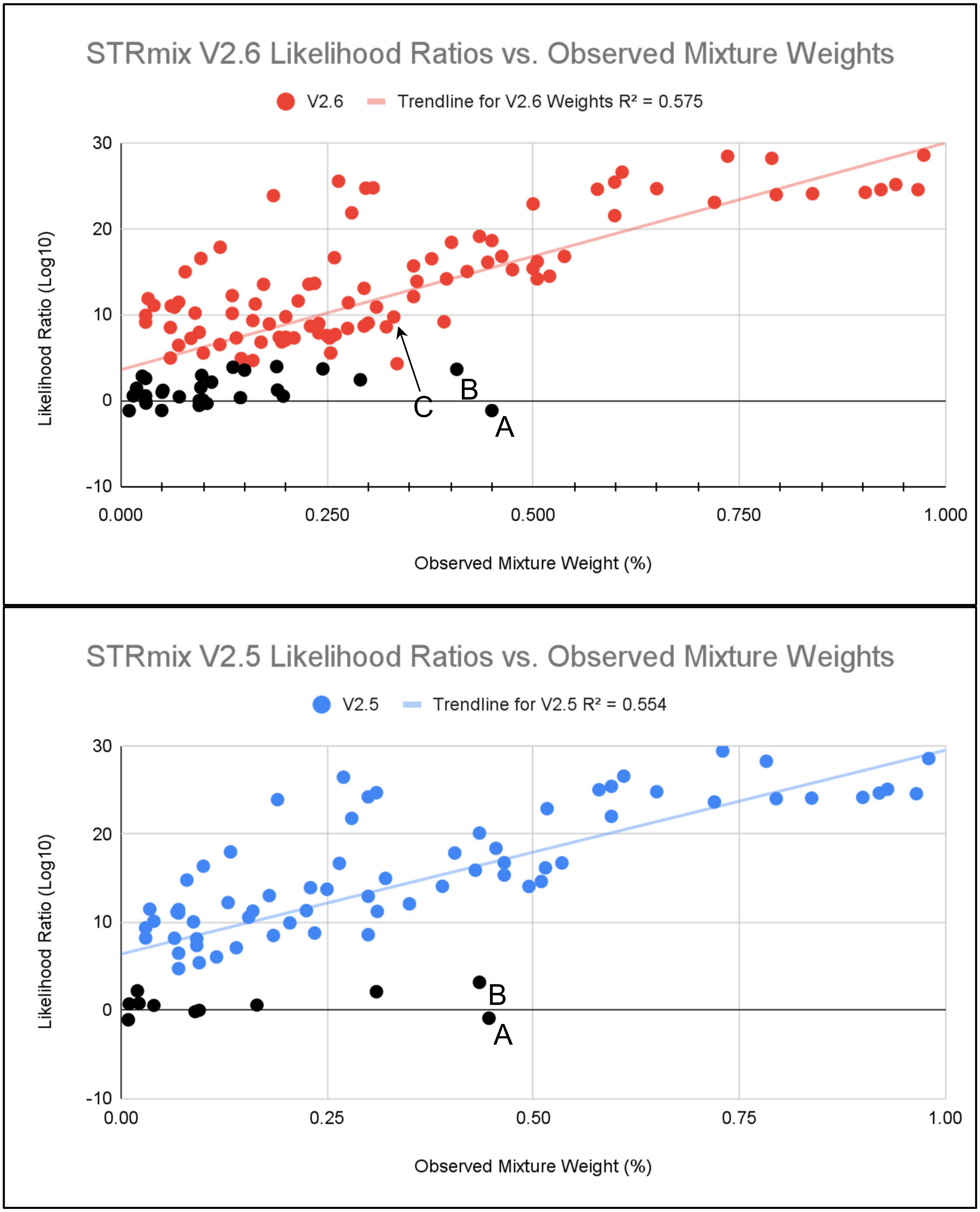


Figure 5 (a, b, and c). *Scatter plots of observed mixture weight proportions versus likelihood ratios (log10) for known contributors under a) TrueAllele®, b) STRmix^{TM} v2.6, and c) STRmix^{TM} v2.5. All non-inclusionary results are represented by black dots.*

Sample Mixture Type	Software Program	Number of non-inclusions	Lowest Mixture Weight Value	Highest Mixture Weight Value	Average Mixture Weight Value
All mixtures	TrueAllele®	27	0.9%	44.0%	11.9%
	STRmix^{TM} V2.6	29	1.0%	45.0%	12.1%
	STRmix^{TM} V2.5	11	0.9%	44.7%	15.0%
	Total	**67**	**0.9%**	**45.0%**	**13.0%**
Two- to Four-Person Mixtures	TrueAllele®	12	0.9%	44.0%	10.3%
	STRmix^{TM} V2.6	11	1.0%	45.0%	15.2%
	STRmix^{TM} V2.5	11	0.9%	44.7%	15.0%
	Total	**34**	**0.9%**	**45.0%**	**13.5%**
Five-Person Mixtures	TrueAllele®	15	4.6%	25.4%	13.3%
	STRmix^{TM} V2.6	18	1.9%	24.5%	10.2%
	STRmix^{TM} V2.5	N/A	N/A	N/A	N/A
	Total	**33**	**1.9%**	**24.5%**	**11.8%**

Table 5. *Ranges and averages of the observed mixture weights for all non-inclusionary results of known contributors.*

Sensitivity

Each program was challenged with mixture weights as low as nearly 0% to as high as 99%. Out of 78 expected known inclusions for two- to four-person mixtures, both versions of STRmix^{TM} included 85.9% (67) of known contributors, and TrueAllele® included 84.6% (66) of known contributors. Out of 44 expected inclusions for five-person mixtures, STRmix^{TM} version 2.6 included 59.1% (26) known contributors, and TrueAllele® included 65.9% (29) known contributors. Figure 5a, 5b, and 5c show the relation between mixture weight proportions and the resulting LRs. All inconclusive and false exclusionary results are represented by black dots on the graph. Table 5 shows the range and average MW proportions for all inconclusive and false exclusionary LRs, in other words, all non-inclusions. All three software programs resulted in a coefficient of determination, R^2, of approximately 0.6. All the non-inclusions had mixture weight proportions lower than 50.0% with an average mixture weight of 12.5%. The first two outliers for each program were investigated further in the discussion section (marked A, B and C on the graphs).

The first five mixtures with the lowest average RFUs per locus were sample mixtures 42 (196 RFUs), 49 (327 RFUs), 51 (937 RFUs), 48 (1083 RFUs), and 47 (1204 RFUs). Table 6 summarizes the count of true inclusionary LRs and non-inclusionary LRs per program for the known contributors in each of these mixture samples.

Sample and NOC	Average RFUs per Locus	PG Software	Number of true inclusions	Number of inconclusives for known contributors
42 4 contributors	196	TrueAllele®	3	1
		STRmix™ V2.6	0	4
		STRmix™ V2.5	0	4
49 5 contributors	327	TrueAllele®	3	2
		STRmix™ V2.6	1	4
51 5 contributors	937	TrueAllele®	4	1
		STRmix™ V2.6	2	3
48 5 contributors	1083	TrueAllele®	5	0
		STRmix™ V2.6	4	1
47 5 contributors (1 unknown)	1204	TrueAllele®	2	2
		STRmix™ V2.6	2	2

Table 6. *Conclusion counts per PG software for the first five samples with the lowest average peak heights per locus in RFUs.*

Specificity

References known to not be in a mixture were compared to each sample mixture profile for each software program. Referring to Table 3, neither of the three programs resulted in a false inclusion of a known-non contributor for any sample mixture. Out of 192 expected exclusions for two- to four-person mixtures, TrueAllele® excluded 70.8% (136) of all known non-contributors, STRmix^{TM} version 2.6 excluded 75% (144) of all known non-contributors, and STRmix^{TM} version 2.5 excluded 78.1% (150) of all known non-contributors. Out of 46 expected exclusions for five-person contributors, TrueAllele® excluded 45.7% (21) of all known non-contributors, and STRmix^{TM} version 2.6 excluded 2.2% (1) of all known non-contributors.

Under-assumed NOC

All comparison requests were re-run with the NOC under-assumed by 1 (n-1). For two- to four-person mixtures, both versions of STRmix^{TM} did not run 16 mixture samples as they resulted in error messages from a failed pre burn-in due to at least one locus showing evidence of a higher NOC. Of these 16 mixture samples, TrueAllele® reported different conclusions for 27 comparison requests in which 1 request increased in LR, and 26 requests decreased in LR for an average LR change of -22.95. For the remaining two- to four-person mixtures that were successfully interpreted on STRmix^{TM}, TrueAllele® reported different conclusions for 45 comparison requests, in which 8 requests increased in LR and 37 requests decreased in LR for an average change of -10.59. STRmix^{TM} version 2.6 reported different conclusions for 56 comparison requests, in which 39 requests increased in LR and 17 requests decreased in LR for an average change of -0.56. STRmix^{TM} version 2.5 reported different conclusions for 49 comparison requests, in which 35 requests increased in LR and 14 requests decreased in LR for an average change of -0.57.

For five-person mixtures, TrueAllele® reported different conclusions for 23 comparison requests, in which 11 LRs increased and 12 LRs decreased for an average change of 0.99. STRmix^{TM} version 2.6 reported different conclusions for 31 comparison requests, in which 5 LRs increased and 26 LRs decreased for an average change of -5.05.

Over-assumed NOC

The same procedure as above was repeated with the NOC over-assumed by 1 (n+1). For two- to three-person contributors, TrueAllele® reported different conclusions for 72 comparison requests, in which 64 LRs increased and 8 LRs decreased for an average LR change of 10.90. STRmix^{TM} version 2.6 reported different conclusions for 130 comparison requests, in which 22 LRs increased and 108 LRs decreased for an average change of 0.00. STRmix^{TM} version 2.5 reported different conclusions for 135 comparison requests, in which 27 LRs increased and 108 LRs decreased for an average change of -0.08.

For four-person mixtures, TrueAllele® reported different conclusions for 21 comparison requests, in which 16 LRs increased and 5 LRs decreased for an average change of 2.75. STRmix^{TM} version 2.6 reported different conclusions for 10 comparison requests, in which all 10 LRs increased for an average change of 7.52.

For five-person mixtures, TrueAllele® reported different conclusions for 18 comparison requests, in which 14 LRs increased and 4 LRs decreased for an average change of 2.11.

Discussion

Data Selection

Data collection for STRmix^{TM} was highly limited in this study. Since STRmix^{TM} is installed on the computer of choice upon purchasing, the performance of the program is highly dependent on the strength and size of the computer it is installed onto. STRmix^{TM} Version 2.5 was installed on a smaller computer. Therefore, comparison requests were limited to four-person mixtures for the true number of contributors and three-person mixtures for the over-assumed number of contributors. Similarly, STRmix^{TM} Version 2.6 was limited to up to four-person mixtures when the NOC was over-assumed due to the size of the computer it was installed on. The purchase of the TrueAllele® software program includes an entire computer system including parallel processors. Though this increases the cost of TrueAllele® compared to STRmix^{TM}, this allows the program to perform more consistently from laboratory to laboratory.

Data collection for STRmix^{TM} was also limited by a feature in STRmix^{TM} (JCRCL, 2019). This feature affected the first sample that had a large amount of stutter peaks. STRmix^{TM} assumed it was a three-person mixture instead of a two-person mixture. Therefore, this sample was re-analyzed with the stutter filter on to help STRmix^{TM} remove some of the stutter peaks. Also due to this feature, the under-assumed NOC dataset was significantly decreased where 16 out of 18 total two-person and three-person mixture results from TrueAllele® could not be compared to either version of STRmix^{TM}.

Mixture Weights

The evidence samples used for this study were mixture samples prepared at KRCL targeting specific DNA proportions. Results show the expected mixture weight proportions varied nearly three times more when compared to the observed mixture weights proportions versus when the observed mixture weights were compared across the programs. Variation is likely to occur between expected and observed values when DNA samples are prepared due to several factors such as pipetting technique, volume of aliquots, and quantitation estimation errors (Perlin et. al., 2015). While the observed values varied significantly in comparison to the expected, these varied very little across the three programs. Nonetheless, the results involving mixture weights were reported separately. Unfortunately, this caused slight variation in the data when compared and summarized in charts.

Concordance

Within the samples of two- to four-person mixtures, Figure 1 shows sample 33 was interpreted 8 times in TrueAllele®. Many of these interpretations were concordant to one another, but since the order of requests was not known, all interpretations were counted. On the other hand, mixture samples 57 and 63 never reached concordance between LRs (they were only concordant in mixture weight proportion values). Therefore, these mixture samples would have required at least one more interpretation to obtain concordant results.

The average number of interpretation requests made for five-person mixtures was similar between TrueAllele® and STRmix™ version 2.6. However, at least three mixture samples (samples 47, 50, and 66) under TrueAllele® did not produce concordant LRs to one another. They were only concordant in mixture weight proportion values. Therefore, at least one more request was required to reach concordance between two LRs. This would theoretically increase the average from 2.9 to at least 3.2 for TrueAllele®.

Accuracy

Figures 4a and 4b showed the number of inconclusive (Inc + and Inc -) results generally increased as the NOC increased. Inconclusive results also mostly occurred when requests were made to known non-contributors, or when requests were made to known contributors of the minor components of a mixture profile.

Overall, all three programs performed similarly when comparing known references to the mixture samples for two- to four-person mixtures, apart from the single false exclusion under both versions of STRmix^{TM} discussed later. When comparing known non-contributors to the mixture samples for two- to four-person mixtures, TrueAllele® produced the most inconclusive results out of the three software programs. Possibly indicating TrueAllele® is the more conservative of the three software programs. However, for five-person mixtures, TrueAllele® had slightly higher accuracy than STRmix^{TM} version 2.6. This could be because the MCMC settings on STRmix^{TM} were set to 100k accepts for all runs, while the MCMC settings on TrueAllele® were sometimes set to 200k allowing the MCMC method more opportunity to deconvolute mixtures of higher complexity more accurately.

When the two- to four-person mixture samples of the known non-contributors were compared, STRmix^{TM} version 2.5 had higher accuracy with the highest count of true exclusions, and the lowest count of inconclusive results. For five-person mixtures, TrueAllele® was more accurate with 20% more true exclusions for known non-contributors than STRmix^{TM} version 2.6. Overall, when challenged with a higher number of contributors, TrueAllele® was able to deconvolute most if not all the contributors in these more complex mixtures, as well as successfully exclude the appropriate references.

Only one interpretation was a false exclusion for both versions of STRmix^{TM}. Mixture sample 39 was a three-person mixture with references 7, 9, and 10. TrueAllele® produced true inclusionary LRs for all three references; 24.86, 27.68, and 11.22 respectively. On the other hand, both versions of STRmix™ produced two true inclusionary results and one false exclusionary result. The LRs from STRmix™ version 2.6 were 24.71, 26.58, and 0 respectively. The LRs from STRmix™ version 2.5 were 24.20, 26.55, and 0 respectively. Further investigations showed STRmix™ gave weight to a stutter peak under locus SE33 when deconvoluting the mixture. The final genotype table for the minor contributor was listed as "29, F", meaning the person of interest must have a 29 allele and the second allele could be anything else. Reference 10 is the known minor contributor, with alleles 16 and 17 at locus SE33. Since the proposed alleles from the genotype table at locus SE33 did not match those from reference 10, the LR at this locus was 0, producing an overall final LR of 0. This scenario is an example of why analyst oversight and confirmation of PG deconvolution results is still important and required.

When looking at the comparison requests for known contributors that produced different conclusions between the three programs for the same comparison requests, two requests had a LR range larger than 10. One was sample mixture 39 compared to reference 10, the single false exclusion for both versions of STRmix™ discussed previously. The second comparison request was sample mixture 66 compared to reference 3. However, it was discussed previously that the results for this mixture sample were not concordant under TrueAllele®. The highest LR for the most concordant runs when mixture 66 was compared to reference 3 were -3.97 and 6.62. From there, the more conservative LR was reported for this study's dataset, -3.97. If one or more runs were requested for this comparison, the high LR difference between the programs could have possibly been resolved.

For known non-contributors, only one comparison request had an LR range between the three programs larger than 10, sample 56 compared to reference 7. For this comparison request TrueAllele® resulted in an LR of -2.27 (inconclusive), STRmix™ version 2.6 resulted in an LR of 0 (excluded), and STRmix™ version 2.5 resulted in an LR of -12.31 (excluded). Both versions of STRmix™ resulted in the same genotype profile for the major contributor, and full F, F (anything, anything) genotype profiles for the second and third contributors. However, when the deconvoluted genotypes were compared to reference 7, STRmix™ version 2.6 resulted in a locus LR of 0 for D19S433 which produced a total LR of 0. STRmix™ version 2.5 resulted in all negative locus LRs, therefore, producing an overall negative LR.

Lastly, allele sharing within mixtures samples and between mixture samples compared to the references had little to no effect on the resulting LRs. All coefficients of determination values, R^2, for the scatterplots (not provided) were less than 0.02 with no strong positive or negative trends. The highest percent of allele sharing for known contributors was 66% and 68% in mixtures 54 and 55 (2-person mixtures) when compared to references 4 and 6 (known relatives). All the LRs produced across the three programs produced a strong positive LR above 7.0, except when reference 6 was compared to mixture 54 in TrueAllele®. This, however, can be due to the comparison being done on the minor component of the mixture (~4% mixture weight proportion). The highest percent of allele sharing for known non-contributors was 52% in mixture 42 (a 4-person mixture) when compared to references 4 and 9. All the results were inconclusive across all three programs, but so were all other comparisons within the mixture. Therefore, this is most likely not a result of allele sharing, but just a result of a complex mixture in general.

Sensitivity

As mentioned previously, all three programs performed similarly when comparing results for known contributors for two- to four-person mixtures. Even when challenged with varying molecular weights proportions, they all produced similar positive trends, Figure 5a, 5b, and 5c. Furthermore, the LR from mixture sample 45 compared to reference 9 was the highest outlier for all three programs, outlier labeled A. The average observed mixture weight proportion for this contributor was 44.5%, which was the highest mixture weight proportion that produced a non-inclusionary result. According to the experiment set up by KRCL, this sample contains DNA from references 1, 7, 8 and 9 with respective mixture weights of 1%, 36%, 9%, and 52%. The average observed mixture weights are 0.9%, 50.5%, 3.9%, and 44.5% respectively. All programs produced an inconclusive LR for reference 1 and reference 9. Since the study targeted a 1% contribution for reference 1, it is understandable that this reference could not be included or excluded with certainty. However, as mentioned previously, reference 9 is an outlier point, A, in Figures 5a, 5b, and 5c despite it being the second highest contributor in the sample. Taking a closer look at the electropherograms, mixture sample 45 had fourteen total drop-in alleles within twelve loci that did not coincide with references 1, 7, 8 and 9, and the heights of these alleles ranged from 399 to 2221 RFUs. Of the fourteen total foreign alleles, five of the alleles did not match any of the ten known references used to prepare the mixture samples. On the other hand, sixteen alleles (one of the alleles being a homozygote) from reference 9 were not detected under twelve different loci in sample mixture 45. The inconclusive result for the reference in this sample appears justified, and not a result of any of the software programs' functionality.

The second major outlier, B, with the second highest mixture weight contribution for an inconclusive result is only seen under both STRmix^{TM} programs, reference 1 in sample mixture

42. Reference 1 is the highest contributor (38.2% average weight between programs) in sample mixture 42. Taking a closer look at the electropherograms, sample mixture 42 has an average peak height of 196 RFUs per locus, which is the lowest average peak height in RFUs out of all the sample mixtures. STRmix^{TM} had inconclusive results for all four known contributors in this mixture, while TrueAllele$^{®}$ positively included the first three contributors with the highest mixture weight proportions in the sample. With this in mind, the next four samples with the lowest peak heights are examined later for similar trends.

The next major outlier for TrueAllele$^{®}$ was reference 7 in mixture sample 66, a mixture of 5 contributors. The two highest contributors in sample mixture 66, references 4 and 7, had an average MW of 29% between TrueAllele$^{®}$ and STRmix^{TM} version 2.6. TrueAllele$^{®}$ resulted in 4 inconclusive results (including reference 7) and STRmix^{TM} resulted in 1 inconclusive result. As mentioned previously, the resulting LRs for this sample mixture were not concordant with each other under TrueAllele$^{®}$, but only concordant in mixture weight contributions. This outlier seems justified as at least one more run was required to reach concordance.

Table 6 shows that when challenged with low RFUs TrueAllele$^{®}$ tends to overcome low level DNA samples. TrueAllele$^{®}$ successfully included more contributors in 4 out of the 5 mixture samples with the lowest average RFUs per locus compared to STRmix^{TM}. As the average RFUs per locus increased within the five mixture samples in Table 6, the number of true inclusions increased for STRmix^{TM}.

Specificity

As shown in Table 3, STRmix^{TM} version 2.5 successfully excluded more known non-contributors for two- to four-person mixtures than the other two software programs. On the other hand, TrueAllele® successfully excluded more known non-contributors than STRmix™ version 2.6 for five-person mixtures.

Under-assumed NOC

It is expected for minor contributors to be falsely excluded when the NOC is under-assumed (Bright, et.al., 2014). When the NOC was under-assumed (n-1) TrueAllele® had the least amount of comparison requests that resulted in a different conclusion than when run under the correct NOC (n). Majority of comparison requests that reported a different conclusion were LRs that decreased for all three programs.

Over-assumed NOC

When the NOC was over-assumed (n+1) for two- to three-person mixtures, TrueAllele® had the least amount of comparison requests that resulted in a different conclusion than when run under the correct NOC, n. Majority of these comparison requests increased in LR for TrueAllele®, but for both versions of STRmix™ most of the comparison results that resulted in different conclusions had LRs that decreased. Only a single comparison request for TrueAllele® resulted in a false inclusion (a true exclusion for n contributors), sample mixture 39 compared to reference 8. No other false inclusions occurred in the study.

For four-person mixtures, STRmix™ version 2.6 had the least amount of comparison requests that resulted in different conclusions. Most, if not all, of these comparison requests had LRs that increased for both STRmix™ version 2.6 and TrueAllele®. This supports findings from other studies showing previously exclusionary LRs will then increase to an inconclusive LR (Moretti, et. al., 2017).

Conclusion

Previous studies have shown the advantages of utilizing PG software programs as tools in forensic DNA analysis. These programs are powerful enough to perform complex calculations on the most complicated mixture samples. But with a competing market of PG software programs, it can be difficult for a laboratory to choose which is right for them.

As seen in this study, TrueAllele® and STRmix$^{\text{TM}}$ *can* produce different conclusions (NY v. Hillary, 2016). Especially as the number of contributors increases in a mixture sample.   This is why it is highly recommended to only report the most conservative result between strictly concordant runs. It is also important to emphasize that PG software programs are tools, in which analyst input is still required to avoid any false conclusions. Nonetheless, empirical examinations and statistical calculations for significant variance between the programs performed within this study showed the performance of all three programs are more similar to one another than they are different. The most discernable result from this study was that both versions of STRmix$^{\text{TM}}$ produced more true conclusions for mixtures with two and three contributors than TrueAllele®. Possibly indicating TrueAllele® is more conservative. However, then the opposite occurred in mixtures with five contributors in which TrueAllele® produced more true inclusions than STRmix$^{\text{TM}}$. This was possibly due to the computer processor strength, and/or the number of accepts in the MCMC settings.

When choosing between these programs a laboratory will need to consider their computer capabilities and the cost of upgrading them if needed. TrueAllele® is purchased along with parallel processors powerful enough to handle low level DNA samples and complex mixtures (Bauer, 2019). If a laboratory chooses STRmix$^{\text{TM}}$, the laboratory will have to purchase separate processors powerful enough to be comparable with the TrueAllele® parallel processors. If the laboratory chooses STRmix and decides to install the software program on their already owned computers, the laboratory will need to limit mixture interpretations to mixtures with no more than four contributors.

Finally, the LRs between the programs all had a strong positive correlation with similar coefficients of determination values. Of all the LRs produced for comparisons of known contributors, the average LR value within each program was similar to one another (within 1 unit difference). In other words, not one program consistently produced LRs significantly higher or lower than the other programs. However, there were LRs that were non-concordant between programs. Some of these non-concordant LRs were around a 6-log difference. The high variance between LRs was not thoroughly investigated in this study. Therefore, this can be an area of focus for future research.

References

Bauer, D. W., Butt, N., Hornyak, J.M., and Perlin, M. W. (2020) Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Science, 65*(2), 380-398.

Bright, J., Curran, J. M., and Buckleton, J. S. (2014) The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. *Forensic Science International: Genetics, 12*, 208-214.

Bright, J., Stevenson, K. E., Curran, J. M., and Buckleton, J. S. (2015) The variability in likelihood ratios due to different mechanisms. *Forensic Science International: Genetics. 14*, 187-190.

Bright, J., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., Buckleton, J. (2016) Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. *Forensic Science International: Genetics, 23*. 226-239.

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

Butler, J. M. (2007). Short tandem repeat typing technologies used in human identity testing. *BioTechniques, 43*(4), Sii-Sv.

Butler, J. M. (2010). Fundamentals of forensic DNA typing. Academic Press.

Cannon, C. C., LeFebvre, A. K., Benjamin, R. C. (2016) Validation of alternative capillary electrophoresis detection of STRs using POP-6 polymer and a 22 cm array on a 3130xl genetic analyzer. *Forensic Science International: Genetics. 22*. 113-127.

Cheng, K., Bleka, Ø., Gill, P., Curran, J., Bright, J., Taylor, D., Buckleton, J. (2021) A comparison of likelihood ratios obtained from euroformix and strmix, *Journal of Forensic Science, 66*, 2138-2155.

Curran, J. M. (2008) A MCMC method for resolving two person mixtures. *Science and Justice, 48.* 168-177.

Dror IE, Hampikian G (2011) Subjectivity and bias in forensic DNA mixture interpretation. Science & Justice 51: 204–208.

Gill, P. (2002). Role of short tandem repeat DNA in forensic casework in the UK-- Past, present and future perspectives. *BioTechniques, 32*, 366-372.

Gouveia, N., et.al,. (2015) Direct amplification of reference sampled with Globalfiler® PCR Amplificaltion Kit. *Forensic Science International: Genetics Supplement Series, 5*. 135-137.

Graw, M., Weisser, H. J., Lutz, S. (2000) DNA typing of human remains found in damp environments, *Forensic Science International, 113*(1-3), 91-95.

Hammond, H. A., Jin, L., Zhong, Y., Caskey, C. T., & Chakraborty, R. (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American Journal of Human Genetics, 55*(1), 175–189.

Jefferson County Regional Crime Laboratory (2019) Internal validation of STRmix^{TM} V2.6 for the analysis of GlobalFiler^{TM} profiles at the Jefferson County Regional Crime Laboratory.

Kadash, K., Kozlowski, B. E., Biega, L. A., Duceman, B. W. (2004) Validation study of the TrueAllele® automated data review system, *Journal of Forensic Science, 49*(4), 660-667.

Kelly, H., Bright, J., Buckleton, J. S., Curran, J. M. (2014) A comparison of statistical models for the analysis of complex forensic DNA profiles, *Science and Justice, 54*, 66-70.

Kern Regional Crime Laboratory (2017) Validation of the TrueAllele® Casework System interpretation of samples analyzed with the GlobalFiler® PCR Amplification Kit.

Lee, H. C., Ladd, C., Scherczinger, C. A., Bourke, M. T. (1998) Forensic applications of DNA typing: Part 2 Collection and Preservation of DNA Evidence, *The American Journal of Forensic Medicine and Pathology, 19*(1) 10-18.

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.

Moretty, T. R., Baumstark, A. L., Defenbaugh, D. A., Keys, K. M., Smerick, J. B., and Budowle, B. (2001) Validation of short tandem repeats (STRs) for forensic usage: Performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples, *Journal of Forensic Science, 46*(3) 647-660.

Moretti, T. R., Just, R. S., Kehl, S. C., Willis, L. E., Buckleton, J. S., Bright, J., Taylor, D. A., Onorato, A. J. (2017) Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles, *Forensic Science International: Genetics, 29*, 126-144.

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, et al. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping, *Science, 235*, 1616-1622.

Perlin, M. W., Dormer, K., Hornyak, J., Schiermeier-Wood, L., and Greenspoon, S. (2014) TrueAllele® casework on Virginia DNA mixture evidence: Computer and manual interpretation in 72 reported criminal cases, *PLoS ONE, 9*(3) e92836-e92837.

Perlin, M. W., Hornyak, J. M., Sugimoto, G., and Miller, K. W.P. (2015) TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Science, 60*(4) 857-868

Perlin, M. W., Legler, M. M., Spencer, C. E., Smith, J. L., Allan, W. P. Belrose, J. L., Duceman, B. W. (2011) Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Science, 56*(6) 1430-1447.

Perlin, M. W., Sinelnikov, A. (2009) An information gap in DNA evidence interpretation. *PLoS One 4*(12) e8327.

Scientific Working Group on DNA Analysis Methods (SWGDAM) (2017) SWGDAM Interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories; available at https://www.swgdam.org/_files/ugd/4344b0_3f94c9a6286048c3924c58e2c230e74e.pdf.

Wages Jr., J. M. (2005) Encyclopedia of Analytical Science, POLYMERASE CHAIN REACTION, Editor(s): Paul Worsfold, Alan Townshend, Colin Poole, Encyclopedia of Analytical Science (Second Edition), Elsevier, 2005, Pages 243-250, ISBN 9780123693976, https://doi.org/10.1016/B0-12-369397-7/00475-1. (https://www.sciencedirect.com/science/article/pii/B0123693977004751)