Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Robust detection of individual forensic profiles in DNA mixtures

J. Isaacson [a,1], E. Schwoebel [b], A. Shcherbina [b], D. Ricke [b], J. Harper [b], M. Petrovick [b], J. Bobrow [b], T. Boettcher [b], B. Helfer [b], C. Zook [b], E. Wack [b,*]

[a] URX, 168 South Park St., San Francisco, CA 94107, United States
[b] MIT Lincoln Laboratory, Bioengineering Systems and Technologies Group, 244 Wood St., Lexington, MA 02420, United States

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form 6 August 2014
Accepted 4 September 2014

Keywords:
Single nucleotide polymorphism (SNP)
DNA mixture
Random man not excluded (RMNE)

ABSTRACT

For a forensic identification method to be admissible in international courts, the probability of false match must be quantified. For comparison of individuals against complex mixtures using a panel of single nucleotide polymorphisms (SNPs), the probability of a random man not excluded, *P*(*RMNE*) is one admissible standard. While the *P*(*RMNE*) of SNP alleles has been previously studied, it remains to be rigorously defined and calculated for experimentally genotyped mixtures. In this report, exact *P*(*RMNE*) values were calculated for a range of complex mixtures, verified with Monte Carlo simulations, and compared alongside experimentally determined detection probabilities.

1. Introduction

Present-day forensic analysis encompasses two methods for comparing uncharacterized, human DNA profiles against reference DNA profiles for identification. In the first, short tandem repeats (STRs) are amplified using polymerase chain reaction (PCR), allele sizes determined, generally using capillary electrophoresis [1], and the results used to query the FBI Combined DNA Index System (CODIS) [2]. A suspect's DNA profile is compared against either a suspect's reference sample or the National DNA Index (NDIS) database to look for STR length (allele) similarity. The second method, mitochondrial DNA (mtDNA) sequencing, gained widespread acceptance in the early 1980s. Each mitochondrion in a human cell contains one strand of near identical, maternally-inherited mtDNA [3,4]. Mitochondrial DNA sequence can be used to designate a subject as excluded, inconclusive, or cannot exclude [5].

It is difficult to compare a mixture of DNA from two or more contributors to a set of reference DNA samples. Electrophoretic measurements are not sensitive enough to detect minor contributor DNA signatures when the ratio of minor to major contributor's DNA is less than 1:10. In more complex mixtures of three or more contributors, modern forensic techniques are often unable to accurately differentiate between suspects, victims and innocent individuals [6], although progress is being made in this area [7].

Previously, Homer et al. [8] has shown that a simulated individual DNA profile can be identified in large, complex, simulated DNA mixtures using SNPs extracted during genome wide association studies (GWAS). Shortly after publication, the generalizability of Homer's methodology was questioned due to several underlying statistical assumptions as noted by Braun, et al. [9]. Specifically, Braun concluded that Homer's model assumes individuals must hail from homogeneous populations. Nevertheless, SNP sequencing remains an active area of interest within the forensic community [9,10].

One step towards wider adoption of SNP sequencing by international court systems is to rigorously define a false match rate. Nuclear STR profiling and mtDNA sequencing have been shown to have false match rates ranging from rarer than one in one billion to one in one thousand. These probabilities are affected by sample degradation and the commonality of specific repeat lengths within biogeographic sub-populations [11]. It is notable that the chance of laboratory error (a false positive, as opposed to a random match) is likely to dominate in practice [12]. Herein, the discussion is limited to consideration of statistical calculations regarding the data itself, and does not include error rates in collection, handling, analysis, or interpretation. To estimate this likelihood of falsely matching an innocent suspect, international courts have adopted two key metrics: the probability of a random man not excluded, *P*(*RMNE*), and the likelihood ratio, LR [13,14].

A model to estimate the *P*(*RMNE*) and LR associated with a forensic identification of DNA mixtures using genetic markers has previously been developed by Buckleton et al. [13] and the specific simplified case for SNP genotyping has been adapted by Voskoboinik

[*] Corresponding author. Tel.: +1 781 981 7481; fax: +1 781 981 7848.
E-mail address: wack@ll.mit.edu (E. Wack).
[1] Tel.: +1 518 698 9208.

http://dx.doi.org/10.1016/j.fsigen.2014.09.003
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

and Darvasi [14]. *P*(*RMNE*) was estimated for mixtures ranging from two to ten contributing individuals, sampling from a large collection of simulated individual DNA profiles. For each study, *P*(*RMNE*) was calculated as a function of population-averaged minor-allele frequencies (mAFs). The study considers the effects of the following on *P*(*RMNE*) calculation:

- SNP panel size
- Multiple relatives' DNA present in a mixture
- Population-specific SNP bias
- Fixed genotyping error, binomially distributed among alleles.

In Voskoboinik and Darvasi's study [14], it is argued that allele-specific mAF values have little effect on *P*(*RMNE*). Fig. 2 from Voskoboinik and Darvasi's study [14] demonstrates two simulations to calculate *P*(*RMNE*). The first uses a panel of 1000 theoretical SNPs each with a constant mAF defined at 0.075. The second uses a panel of 1000 SNPs with mAFs uniformly sampled from 0.05 to 0.1. Following the central limit theorem, the mean *P*(*RMNE*) of the second simulation equals the *P*(*RMNE*) of the first simulation. It is further demonstrated that the 99% confidence interval of *P*(*RMNE*) for the second simulation ranges approximately three orders of magnitude. Voskoboinik and Darvasi [14] use these results as a justification to assume a constant mAF of 0.075 for all *P*(*RMNE*) calculations.

While this approximation is valid, forensic quality calculation of *P*(*RMNE*) using experimentally derived SNP calls requires precise treatment of the affects of allele-specific mAFs. The difference between an estimated false match rate of 10^{-14} and 10^{-11} may not affect a jury's interpretation of a crime scene, but in more complex mixtures, the difference between 10^{-9} and 10^{-6} may affect the admissibility of DNA evidence.

To this end, we have generalized the calculation of *P*(*RMNE*) of SNP panel sequencing to more accurately identify individuals in complex, multi-contributor mixtures and to incorporate the impact of errors that result in mismatches between reference and sample data. A rigorous calculation of *P*(*RMNE*) has been constructed from sequenced mixtures and compared with both Voskoboinik and Darvasi's [14] approximation and Monte Carlo simulations of random individuals. While the likelihood ratio is a commonly used metric for estimating false matches in courts, it has not been considered in these calculations because the likelihood ratio unrealistically requires knowing *a priori* the number of contributors to a mixture. By comparing reference genotypes with the mixtures, the probability of detecting an individual has been quantified and compared against the *P*(*RMNE*) [15].

2. Methods

2.1. SNP sequencing

We have created a panel of 480 SNPs to analyze complex mixtures. The loci were selected in order to optimize the mixture analysis process. The design criteria included the following elements (using data from the ALFRED [16] and dbSNP [17] databases):

- $F_{st} < 0.06$ to minimize impact of population specific differences
- Minor allele frequency between 0.03 and 0.07
- No health-related SNPs
- Minimum distance between SNPs of 500,000 bases to minimize linkage between loci
- Avoid homopolymer stretches due to higher sequencer error.

We have sequenced a panel of 480 SNPs for twenty-three unrelated individuals and nine mixtures. Each mixture is generated from equimolar combinations of subsets of these twenty-three individuals.

Buccal cells were collected from individuals using either Bode or Epicenter swabs. Genomic DNA was isolated using the Qiagen Investigator Kit, eluted in water, then quantitated by Quant-iT dsDNA High-Sensitivity Assay kit from Invitrogen. For samples that needed concentration, Amicon Ultra-0.5 Centrifugal Filter Device 30 K was used. Devices were first pre-rinsed with dH_2O, according to the protocol. Samples were then added, centrifuged, recovered and re-quantitated using Quant-iT kit. Fluidigm designed panels of primers based on a submitted list of SNPs with a minor Allele Frequency (mAF) ranging from 0.03 to 0.07. The final design resulted in a panel of 480 SNPs. Primers were combined and diluted according to Fluidigm's manual "Access Array System for Ion Torrent PGM Sequencing System". The PCR products were then generated on the 48.48 Access Array System, using Ion Torrent bi-directional sequencing adapters and barcodes. PCR products were pooled and purified using Agencourt Ampure XP beads according to Fluidigm's manual, section: "The Purification of Harvested PCR Products" with the exception that the samples were eluted off beads of low TE (EDTA, Tris-HCl). Samples were then quantitated by Quant-iT, and subsequently diluted to the concentration suggested by Ion Torrent, using an average length of 200 base pairs for the calculation. Molecules/μL = (sample concentration[ng/μL] $\times$ 6.022 $\times$ 10^{23})/(656.6 $\times$ 10^9 $\times$ amplicon length [base pairs]). Amplicon libraries were then prepared for sequencing using the Ion OneTouch Template reaction kit, followed by the Ion Torrent PGM Sequencing kit, and sequenced on the Ion Torrent PGM.

The ION Torrent PGM chemistries have known challenges with homopolymer sequences, particularly A-rich stretches. This has been encountered when we have sequenced STR loci including D18S51, FGE and Penta E and D. For SNP panels, this can be largely avoided when designing the panel, as we have done in this instance.

SNPs were identified using a software suite developed by the authors. The software reads a set of barcoded DNA sequences from the Ion Torrent server. In parallel, reads are aligned to a barcode database and a human reference DNA database using BLAST [18]. If both the barcode and 5′ and 3′ flanking DNA sequences partially match their respective databases, the SNP allele called between the 5′ and 3′ flanks is extracted. Here, a partial match has been heuristically rule driven:

- The five bases nearest to the SNP base must perfectly align to the DNA database.
- The next five bases out may have, in aggregate, at most one error.

Calls from alleles with the same identified barcode and locus are aggregated and output to a comma separated value worksheet. Only reads with a mean Phred quality score of 20 or higher are used to make SNP calls.

2.2. Analysis overview

For each sequenced individual and mixture, a minor allele ratio (mAR) is calculated for each of 480 loci by dividing the number of minor allele calls by the total number of calls at each SNP locus. Fig. 1 visualizes the minor allele ratio at 448 of the 480 successfully sequenced loci for one individual. The remaining 32 loci sequenced with less than 50 calls (per locus) on one or both strands and were not used for analysis.

In Fig. 1, an individual's mAR is calculated for each locus, rank ordered and plotted. Allele calls are aggregated into four main clusters:

- Homozygous major: mAR < 0.2
- Heterozygous: mAR ≥ 0.2 and <0.8
- Homozygous minor: mAR ≥ 0.8.

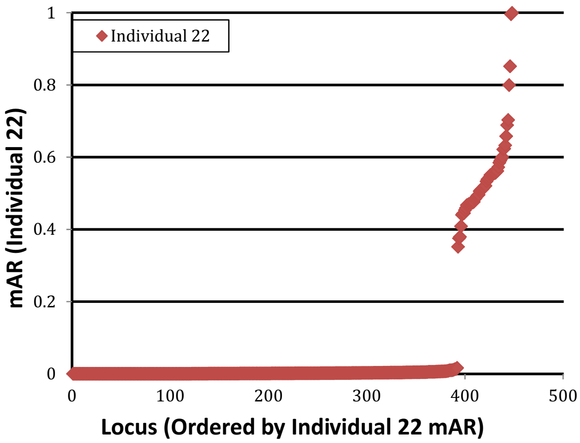

Fig. 1. mAR plotted for one individual (Individual 22). The mAR clusters around 0 (homozygous major), 0.5 (heterozygous) and 1.0 (homozygous minor). Intermediate values may be random or could represent errors in sequencing/amplification or analysis.

This specific individual presents with 392 homozygous major alleles, 52 heterozygous loci and 4 homozygous minor alleles. This distribution of minor alleles is representative of all of the individuals analyzed for this study.

Mixtures can be characterized in a similar manner using a plot of rank-ordered mAR values. Fig. 2 shows mAR data from an equimolar mixture of DNA from 8 individuals plotted on a logarithmic scale.

In a real-world scenario, the threshold for identifying the presence of a minor allele must be set without *a priori* knowledge of the components (number of contributors and molar ratios) of a mixture. We set a mAR threshold such that loci with a mAR of >0.01 are designated as mA-positive. This generally places the threshold near to cross-over between mA-negative and mA-positive loci in the truth data.

In this experiment, the 0.01 threshold matched the distribution of the true presence of minor alleles well. In Fig. 2 there are nine false positive loci, wherein the mixture data for these loci produces a mAR greater than 0.01 but the truth data indicates that none of the reference samples had a minor allele. There are three false negative loci where minor alleles are detected in individual samples but not present in the mixture. Across the ten equimolar mixtures sequenced with this 480 SNP panel, the median number of false positive loci calls was 5 and the median number of false negatives was 2.5.

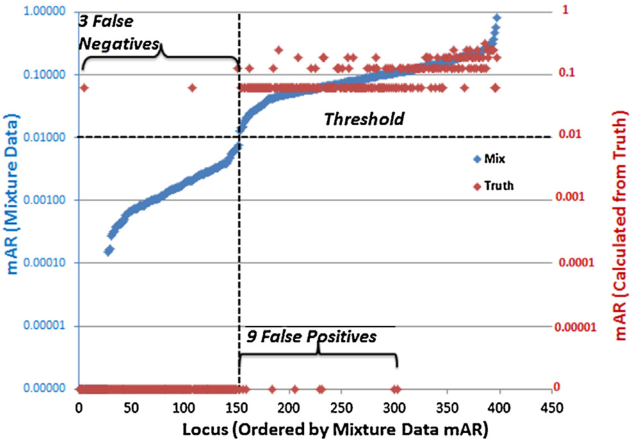

Fig. 2. Plot of minor allele ratio (mAR) values for 398 loci from an equimolar, 8-person mixture. Loci (*Y*-Axis) are rank-ordered by mAR value (*Y*-axis) from the mixture data. Blue Diamonds: mAR determined experimentally from sequence data on an 8-person DNA mixture. Red Diamonds: Proportion of alleles in the sample that are expected to contain mA (as determined from each individual's profile) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

We have also performed replicate sequencing of the reference profiles across the 23 subjects and found the median per locus error rate to be less than 0.06% when filtering the data by Q20 base quality and a minimum of 100 sequence reads per locus (50 reads minimum on each strand). The sources of the false positives and false negatives will be sample, amplification and sequencer technology specific, but the rates observed here are likely representative of the state of technology.

2.3. P(RMNE) calculation

Building on previous work from Buckleton et al. [13] and Voskoboinik and Darvasi [14], we developed a model to declare if an individual is present within a mixture while accounting for mismatched alleles. Minor allele calls are compared between an interrogatory sample and a mixture, declaring each locus a match or mismatch. Because of all of the sources of non-trivial error discussed above, it is necessary for any detection module to allow for one or more mismatches. For this reason, an individual as a whole is declared "present in mixture" if the number of mismatches is small, or declared "absent from mixture" if the number of allele mismatches is large. Selection of an optimal minor allele mismatch level requires balancing true detection sensitivity and false-positive inclusions. The effects of varying this mismatch threshold are presented below.

At any particular locus, an individual's mAR matches the mixture if both present with homozygous major allele ratio. However, if an individual presents with one or more minor alleles at a particular locus but the mixture appears homozygous major, a mismatch is called. The rare case where a mixture is called homozygous minor at a locus is also considered; a match is called for a homozygous minor or heterozygous individual, and a mismatch otherwise.

A model to evaluate *P*(*RMNE*) for any number of allowed allele mismatches, considering variance in mAF between loci, has been developed. To ensure a precise calculation of the *P*(*RMNE*) value, four methods are calculated and compared:

- Direct Evaluation of the probability distribution function (pdf)
- Approximation of the pdf via a binomial distribution similar to [14]
- Exact calculation of the pdf utilizing the Discrete Fourier Transform–Characteristic Function (DFT-CF) method
- A Monte Carlo simulation.

Previously, the probability of a random man not excluded has been formulated as [14]:

$$P(RMNE) = \prod_{i=0}^{S}(1-p_i)^2 \approx \bar{P}^{2S} \qquad (1)$$

A mixed DNA sample is genotyped at N SNP loci where S is the number of homozygous major loci in a mixture, p_i is the population estimated mAF of allele *i* and $\bar{P}$ is the average mAF taken across all *S* loci. This model of false inclusion assumes all minor allele-containing loci must match between a reference and a mixture. If all *N-S* alleles match, then a reference DNA sample will be called homozygous major with probability $(1-p_i)^2$ at every locus called homozygous major in the mixture. A subject can be excluded from a mixture at a particular locus if they have a minor allele when that locus is homozygous major in the mixture. Exclusion due to homozygous minor loci in the mixture is about three orders of magnitude less likely than exclusion due to homozygous major loci [14]. Allele mismatches are not considered in equation (1). As mentioned in the article, [14], and described previously, averaging minor allele frequencies using the approximation in equation (1) boosts computational speed of evaluating $P(RMNE)$, at the cost of a reduction in the model's accuracy. This approximation of $P(RMNE)$ in equation (1) is valid for small mixtures with no sequencing error. However, the model is not accurate enough for forensic admissibility. A more precise model considering allele mismatches and allele specific mAFs is required.

Rather than fix an estimated allele call error rate as in [14], the approximation in equation (1) can be improved to include allele mismatches by rewriting as a binomial distribution:

$$P_M(RMNE) \approx \sum_{k=0}^{M} \bar{P}^{2k}(1-\bar{P})^{2(S-k)}\binom{S}{k} \tag{2}$$

where M is the maximum number of allowed mismatching loci. M greater than zero results from any number of deficiencies in the genotyping process such as drop out, low copy number, amplification, sequencing errors and others. For equation (2), $P(RMNE)$ is approximated as a Binomial distribution with a "probability of success" equal to the average minor allele frequency across all forensically valid loci, $\bar{P}$. Using this improvement over equation (1), $P_M(RMNE)$ can be quickly evaluated for all M less than S. Given a desired probability of false inclusion tolerance t, a threshold can be placed on the maximum number of allowed allele mismatches, M, such that $P_M(RMNE) < t$.

The probability of a random man not excluded can be further improved to include both locus-dependent mAFs, the inclusion of possible mismatches where the mixture is homozygous minor and the suspect is not, and an arbitrary number of allele mismatches:

$$P_M(RMNE) = \sum_{k=0}^{M}\left[\prod_{A_k}\prod_{i\in A_k^c}(1-(1-P_i^2)) + \prod_{B_k}\prod_{i\in B_k} p_i^2 \prod_{i\in B_k}(1-p_i^2)\right] \tag{3}$$

where A_k represents all sets of (S-k) loci wherein both the individual and the mixture are called homozygous major and A_k^c represents all sets of k loci wherein the individual is called with one or more minor alleles and the mixture is called homozygous major. Similarly, let Q equal the number of alleles wherein the mixture is called homozygous minor (mAR > .99). B_k represents all sets of (Q-k) loci wherein both the individual and the mixture are called homozygous minor and B_k^c represents all sets of k loci wherein the individual is called with less than two minor alleles and the mixture is called homozygous minor. In practice, Q is observed as zero and equation 3 reduces to:

$$P_M(RMNE) = \sum_{k=0}^{M}\left[\prod_{A_k}\prod_{i\in A_k^c}(1-p_i)^2 \prod_{i\in A_k^c}(2*p_i - p_i^2)\right] \tag{4}$$

The cumulative distribution function (cdf), evaluated over a range of M mismatched loci, presented in equation (4) is equivalent to the Poisson-Binomial distribution where the probability of a homozygous major allele call, $(1-p_i)^2$ defines the probability of success for trial i. Equations (3) and (4) directly calculate the probability that a random individual, with T minor alleles, has M mismatching and T-M matching minor allele loci for all combinations of possible mismatching alleles. This is a powerful tool allowing a user to immediately quantify, for a given mixture, how the choice of the number of allowed mismatched loci due to all sources of analytical error affects the likelihood that a random individual would be falsely called "present in mixture", or the $P(RMNE)$. This will also lend confidence to the judicial system that $P(RMNE)$ is being accurately calculated for imperfectly matching data.

While this generalization adds precision to previous models, equation (4), which requires computing all combinations of M mismatching loci over S-M homozygous major loci, must be considered. This creates a large computational running time prohibiting calculation for large k. To be precise, the cost of evaluating this distribution function for M mismatching loci is $O(S^M)$—a polynomial algorithm. Calculating this probability distribution function for $M = 4$ and $N = 299$ on a MacBook Pro with 4 GB of RAM and an Intel 2.7 GHz i7 takes upwards of an hour.

To mitigate the long running time of direct calculation but retain the fidelity of the Poission-Binomial model, the DFT-CF method of evaluating the Poisson-Binomial distribution is utilized [19]. It is an exact method, simultaneously evaluating the above probability distribution function for all $M \in [0, S]$ in $O(S\log S)$ running time. The same MacBook Pro is able to evaluate the RMNE for any number of mismatches in seconds.

To verify the accuracy of equation (4) and study its improvement over equation (2), Monte Carlo simulations are performed. Approximately one billion individuals (i.e., two billion alleles) are generated with minor allele frequencies pulled from ALFRED, an allele frequency database for anthropology [16]. Each individual is compared against an input mixture and the number of locus mismatches is tabulated. The probability of a RMNE for k mismatches is calculated as the number of simulated individuals with k mismatches divided by the total number of simulated individuals. This Monte Carlo simulation runs in approximately one hour across 64 processors on our internal parallel grid computing environment.

The four methods are compared in Fig. 3.

Plotted here is the P($RMNE$) cumulative distribution function versus number of allowed mismatches. The plot visualizes the likelihood of a random individual from the population matching this 15-person mixture given k or fewer allowed locus mismatches. This simulation used a 975-locus SNP panel with similar design parameters as the 480-locus SNP panel used in the experimental data shown later. Here direct calculation (green diamonds) refers to directly evaluating the P($RMNE$) probability distribution function. Note that it is computationally prohibitive to directly calculate more than the first 4 terms in this cumulative distribution function. Poisson Binomial (blue stars) refers to the DFT-CF method of evaluating the $P(RMNE)$. It is observed in Table 1 that the direct calculation and DFT-CF methods match with an

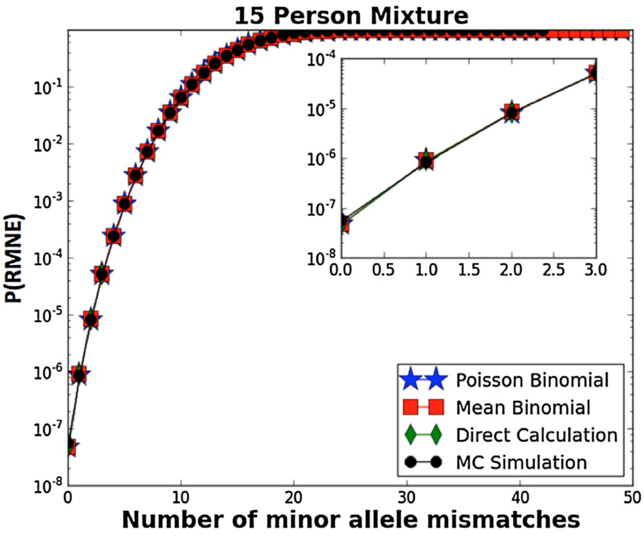

Fig. 3. Comparison of $P(RMNE)$ calculation methods.

Table 1
Root mean square error (RMSE) between direct evaluation of the *P*(*RMNE*) cumulative distribution function and the DFT-CF method (exact) and the Binomial method (approximation) The DFT-CF method will be the default method used in subsequent mixture identifications. It is an accurate, exact method, robust to finite size effects and fast to calculate.

P(*RMNE*) RMSE	DFT-CF	Binomial
Direct evaluation 10 Person	5.37E−17	1.37E−10
Direct evaluation 15 Person	5.78E−17	5.29E−07
Direct evaluation 20 Person	1.80E−16	2.02E−05

error rate equivalent to machine precision (1E−17). The fixed minor allele frequency, Binomial calculation is a reasonable approximation for small mixtures. Indeed, even for ten individuals the error compared against an allele specific calculation is only 1E−10. However, this method incorrectly calculates the *P*(*RMNE*) by several orders of magnitude for the larger 15 and 20 person mixtures. This again lends credence to the argument that per-locus minor allele frequencies are important in accurately determining if an individual is contained within a mixture. The Monte Carlo simulation (MC, black circles) closely follows the DFT-CF calculation and verifies that the above methods closely predict minor allele mismatches of simulated individual. There is a slight deviation between exact calculations and the Monte Carlo simulation around $k = 0$. This error is caused by finite size effects: the *P*(*RMNE*) is approaching one-in-one hundred million, about one order of magnitude smaller than the minimum resolution of the MC simulation, one-in-one billion (one divided by the number of individuals). As the *P*(*RMNE*) approaches the minimum resolution of a MC simulation, errors increase.

2.4. Probability of detection (Pd) calculation

Since the individuals' genotypes have been determined, a probability of detection, Pd, can thus be quantified for each mixture. The number of allele mismatches between a mixture and each contributing individual is recorded. Here a mismatch is defined as before, where a locus is determined to be homozygous for the major allele in the mixture and an individual in the mixture has one or more minor alleles at that locus. Given a minor allele mismatch tolerance, the number of individuals successfully identified within a mixture is calculated. Results from comparing Pd with *P*(*RMNE*) are presented as receiver-operator curves (ROC) within the results section below.

3. Results

3.1. P(RMNE) *for 480 panel*

The *P*(*RMNE*) has been evaluated for a panel of 480 SNPs. Several equimolar mixtures with varying numbers of contributing individuals have been compared. The minor allele frequency values used in calculations are taken from the Allele Frequency Database (ALFRED) SNP population frequencies.

Fig. 4 plots P(*RMNE*) curves for nine mixtures sequenced as described in the methods section. At each locus, a minor allele ratio is calculated as the number of observed minor allele calls divided by the total number of calls. Of the original 480 SNPs, only those appearing homozygous major, defined as loci with a minor allele ratio less than 0.01 are used for P(*RMNE*) calculation. These major allele subsets of the SNP panel range from approximately 10% (3 person mixture) to 80% (15 person mixture) of the original panel size.

As expected, the larger the number of contributing individuals, the more likely a random man may be falsely called "in the mixture." Given a false match tolerance of one-in-one million the

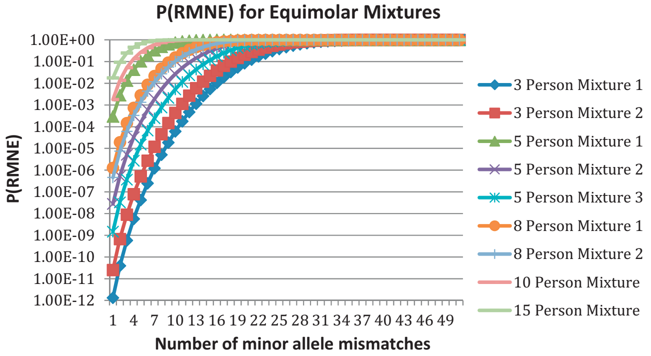

Fig. 4. Analysis of equimolar mixtures using a panel of 480 SNP loci.

3 and 5 person mixtures can be admissible allowing between 0 and 10 allele mismatches. Increasing the false match tolerance to one-in-one billion, only the 3 person mixtures can be admissible at a strict tolerance of 2 or 5 respective mismatches.

While the *P*(*RMNE*) of these more complex mixtures fail to meet the forensic guideline false-positive rate of one in one billion, this particular SNP panel is small [14]. With minimal increase in computational or experimental effort, a larger panel can be run on the same mixtures to increase detection fidelity.

Several mixtures of the same size have been sequenced, providing insight into the sensitivity of a mixture to the specific contributing individuals. The three person mixtures have a variance of several orders of magnitude. This variance is attributed to the inclusion of an individual in *3 Person Mixture 2* with approximately 50% more minor alleles than the other contributors. While the mAFs associated with an individual's minor allele SNP profile are not directly considered in equation 4, minor alleles in an individual also present as minor alleles in DNA mixtures containing that individual. As only homozygous major-allele loci are considered in equation (4), individuals with a large number of minor alleles reduces the number of sequenced alleles available for *P*(*RMNE*) calculation. This increases *P*(*RMNE*).

Conversely, the five and eight person mixtures are relatively robust to specific contributors. It is worth noting the aforementioned person with a 50% higher occurrence of minor alleles is also contained within one of the two 8-person mixtures and two of the three 5-person mixtures. While this individual causes the number of minor alleles in those mixtures to rise, there are enough total contributing individuals that this person's minor allele contributions are small. As the number of individuals within a mixture increases, any one individual's contribution to the mixture's minor allele ratio decreases.

The only parameter input into a *P*(*RMNE*) calculation is the minor allele ratio (mAR) threshold. This value defines the maximum allowed mAR for a given allele within a mixture to be called homozygous major. The above calculation uses a mAR threshold of 0.0125, less than the minimum expected mAR for a twenty person equimolar mixture, 1/40. As described previously, it is difficult to determine a mAR threshold for any given mixture. For this reason, the sensitivity of *P*(*RMNE*) calculations to mAR threshold is investigated. Fig. 5 below details the sensitivity of RMNE calculations to choice of mAR Threshold for one 10-person mixture.

As shown in Fig. 5, choice of mAR threshold has a significant impact on the RMNE curve. Larger mAR thresholds allow for a majority of loci to appear homozygous for the major allele, thus reducing the likelihood of a random man not excluded. Conversely, smaller mAR thresholds restrict all loci but those with the lowest mAR from being called homozygous for the major allele, removing a majority of the panel from computation and increasing *P*(*RMNE*).

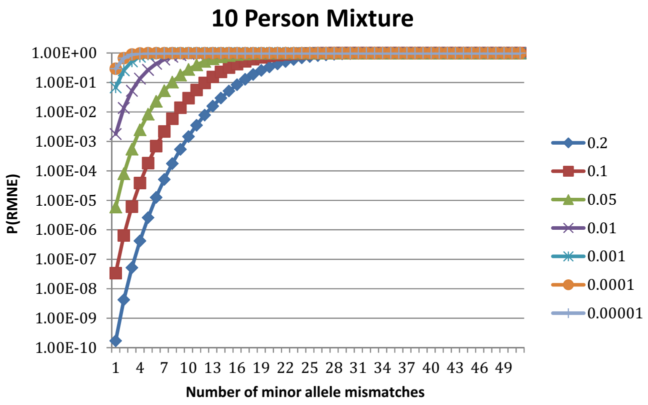

Fig. 5. P(*RMNE*) variability with changes in mAR threshold.

A balance must be struck between ensuring obvious, heterozygous loci are not called homozygous for the major allele, and erroneously removing those loci from the calculation. An upper bound can be drawn for equimolar mixtures at a mAR threshold of $1/(2 \times N)$ where N is the number of contributing individuals. However, this limit is not practical for non-equimolar mixtures with unknown numbers of contributors.

3.2. *Pd vs* P(RMNE)*–fixed mAR Threshold*

Given known reference genotypes for each of the mixtures, a probability of detection is calculated as described above, and displayed in Fig. 6.

Each plot marker in Fig. 6 displays a Pd vs *P*(*RMNE*) value for a fixed number of allowed minor allele mismatches ranging from 0 to 19. For each mixture, as the number of allowed mismatches is increased the Pd vs *P*(*RMNE*) curve shifts from lower left: worse detection power and decreased likelihood of random non-exclusion to upper right: improved detection power at the cost of increased likelihood of random non-exclusion.

As previously described, minor allele mismatches must be allowed in any detection module because of non-trivial levels of sequencing error. Here, the tradeoff between detection power and false random man inclusion can be visualized. For example, the five-person mixtures present with a probability of detection of 0.8 or higher at a *P*(*RMNE*) of 5×10^{-7} for zero mismatches. This means that 80%(4/5) or more of the individuals are correctly identified as present in the mixture with a very small likelihood of false match when no allele mismatches are allowed.

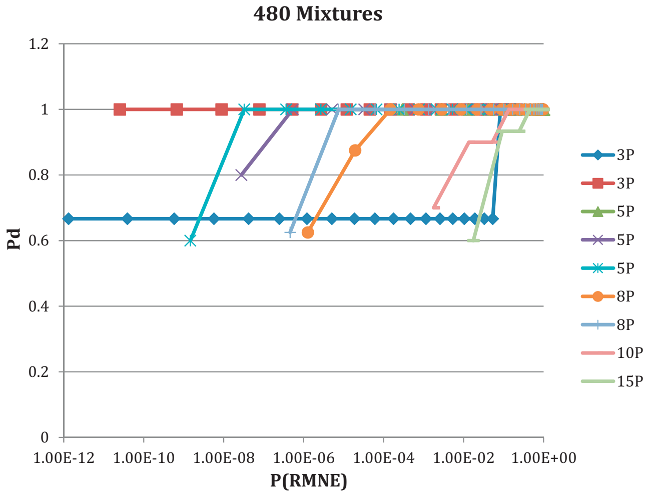

Fig. 6. A ROC-like curve demonstrating the probability of detecting individuals in a mixture, Pd, versus the likelihood of falsely including random individuals, *P*(*RMNE*), as the number of allowed minor allele mismatches increases from 0 to 19.

For a forensically valid *P*(*RMNE*) of one in one billion, one of the 3 person mixtures can allow two to three allele mismatches, while retaining a Pd of 1.0. The 5 person mixture cannot allow any mismatches to retain a *P*(*RMNE*) less than 10^{-8}, causing Pd to decrease to 0.8. More complex mixtures have too large of a false match rate to be court admissible. Nevertheless, there is less than a one-in-a-million chance of falsely including a random man for 8 person mixtures while retaining a Pd of .6.

4. Conclusion

A model has been developed for evaluating the exact probability of a random man not excluded for SNP mixture analysis. This model extends previous work by incorporating allele specific population frequencies into the *P*(*RMNE*) calculation. Through this inclusion, more complex mixtures can be evaluated with a higher accuracy than previously possible. This *P*(*RMNE*) does not require knowledge of reference genotypes; given an unknown mixture, the likelihood of false inclusion can be calculated. For nine mixtures sequenced in this experiment, a probability of detection has also been calculated. This Pd is compared against *P*(*RMNE*) to ensure that true positive detection rates remain high, even with stringent requirements on false positive rates.

Nevertheless, five of the nine mixtures sequenced were unable to meet the forensic guideline false positive threshold of one-in-one-billion. The simplest means of increasing this model's detection performance and reducing false positive identifications is to increase the number of loci in the panel. An increase in panel size adds more terms into equation (1) and lowers *P*(*RMNE*). It is valuable to observe how allowing minor allele mismatches affects probability of detection and probability of random man not excluded calculations.

Work continues to develop a definitive method to determine the optimal mAR threshold for an unknown mixture. Tracking true positive and false negative calls across mixtures may provide insight into an optimal threshold. A mixture-dependent threshold will also be considered. As insight is gained into prediction of number of contributors (and their input ratios), a robust model to predict contributor-dependent mAR thresholds can be developed.

In parallel, we continue to investigate the affect of panel size on *P*(*RMNE*) calculation. As individuals are sequenced with larger SNP panels, *P*(*RMNE*) naturally decreases. We wish to strike a balance in panel size large enough for *P*(*RMNE*) of forensic mixtures to be below one-in-one billion, but small enough that there are a statistically significant number of reads per locus per run. We are expanding the sequencing to the Ion Torrent Proton platforms on touch samples to ensure a sufficient number of reads are generated per locus, particularly for larger panel sizes and greater number of contributors to complex DNA mixtures.

Acknowledgements

This work is sponsored by the Joint Improvised Explosive Device Defeat Organization under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

The authors wish to thank Jason Kulakowski, Deputy Division Chief, Weapons Technical Intelligence for the Joint Improvised Explosive Device Defeat Organization (JIEDDO) for his continued support.

References

[1] J.M. Butler, Short tandem repeat typing technologies used in human identity testing, Biotechniques 43 (2007).
[2] http://www.cstl.nist.gov/strbase/fbicore.htm.
[3] T. Melton, C. Holland, M. Holland, Forensic mitochondrial DNA analysis: current practice and future potential, Forensic Sci. Rev. 24 (July (2)) (2012).
[4] H. Chial, J. Craig, mtDNA and mitochondrial diseases, Nat. Educ. 1 (1) (2008).
[5] http://swgdam.org/SWGDAM%20mtDNA_Interpretation_Guidelines_APPROVED_073013.pdf.
[6] Peter Gill, James Curran, Cedric Neumann, Amanda Kirkham, Tim Clayton, Jonathan Whitaker, Jim Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Sci. Int.: Genet. 2 (March (2)) (2008) 91–103, ISSN 1872-4973.
[7] D. Taylor, K. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, FSI Genet. 7 (2013) 516–528.
[8] N. Homer, et al., Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays, PLOS Genet. 4 (August (8)) (2008).
[9] R. Braun, et al., Needles in the haystack: identifying individuals present in pooled genomic data, PLOS Genet. 5 (October (10)) (2009).
[10] S.H. Lee, et al., Predicting unobserved phenotypes for complex traits from whole-genome SNP data, PLOS Genet. 4 (10) (2008).
[11] J.M. Butler, Fundamentals of Forensic DNA Typing, Elsevier Academic Press, San Diego, 2010.
[12] W.C. Thompson, et al., How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (January (1)) (2003).
[13] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2005.
[14] Lev Voskoboinik, Ariel Darvasi, Forensic identification of an individual in complex DNA mixtures, Forensic Sci. Int.: Genet. 5 (November (5)) (2011) 428–435, ISSN 1872-4973.
[15] John Buckleton, James Curran, A discussion of the merits of random man not excluded and likelihood ratios, Forensic Sci. Int.: Genet. 2 (September (4)) (2008) 343–348, ISSN 1872-4973.
[16] ALFRED: An Allele Frequency Database for Anthropology. http://alfred.med.yale.edu.
[17] dbSNP: Short Genetic Variations. http://www.ncbi.nlm.nih.gov/projects/SNP/.
[18] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[19] Yili Hong, On computing the distribution function for the Poisson binomial distribution, Computat. Statist. Data Analysis 59 (March) (2013) 41–51, ISSN 0167-9473.