Population Genetic Issues for Forensic DNA Profiles
PROGRAM NARRATIVE

## Specific Aims

The particular topics to be covered by the proposed work fall under three main headings:

- Characterizing population structure and relatedness.
- Interpreting lineage marker evidence.
- Adopting next-generation sequencing data.

Work in these three areas is designed to advance the utility of genomic profiling for human identification, with an ultimate goal of quantifying the strength of genomic evidence. There has been tremendous progress over the 32 years since the initial publication by Jeffreys in 1985 on “DNA Fingerprinting” but new technologies and data have raised new issues in these three areas.

## Statement of the Problem

This application is for support to continue work on the population genetic issues affecting the interpretation of forensic DNA profiles. For the period 1/1/12-12/31/14 the work has been supported by NIJ award [redacted], and for the period 1/1/15 - present it has been supported by NIJ award [redacted]. A progress report for the first two years of [redacted] is given in this application package. [redacted] is the PI of the current award and for the requested award, and he is joined by [redacted] of the [redacted] and by [redacted] of the [redacted]. These three senior investigators have worked together on the current award and for many years before that. [redacted] will also continue to work on the project,

The investigators have worked to base their investigations on sound statistical and population genetic theory. They have long advocated the use of likelihood ratios for forensic calculations, and this use makes it important to distinguish between a profile probability, the chance a randomly-chosen person will have a particular DNA profile, and match probability, the chance a randomly chosen person will have a profile given that the profile has already been seen in the present case. [redacted] and [redacted] worked with [redacted] in 2013 to include this distinction in the 2014 Y-STR interpretation guidelines. [redacted] participated in the Y-STR invited panel discussion at the AAFS meeting in 2017 and noticed a need to remind the audience that profile probabilities (e.g. Recommendation 4.1 of the National Research Council report of 1996) were different from, and should not be combined with, match probabilities (e.g. NRC Recommendation 4.2). The present investigators regard their efforts to inform US forensic scientists at least as important as their development of new theoretical methodologies.

## Population structure and relatedness

We take as a starting point that, for example, the match probability for a homozygous profile $AA$ at a single autosomal locus is $[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]/[(1+\theta)(1+2\theta)]$ (Balding and Nichols, 1994) and for Y-STR allele $A$ is $[\theta + (1-\theta)p_A]$ (e.g. Buckleton et al., 2011). For autosomal profiles, it is customary to multiply the resulting match probabilities over loci and for Y-STR profiles to use a multi-locus value of $\theta$. There is some empirical support for these approaches (e.g. autosomal: Weir, 2004; Y-STR: Hall, 2016). Given the increasing number of CODIS loci, the use of highly-mutable Y-STR loci and the introduction of NGS variants, we believe it prudent to re-examine this issue, not to undermine current procedures but to strengthen future practice.

As background, we display in Figure 1 estimated $\theta$ values, using different numbers of loci, and shown on a $-\log_{10}$ scale, for published PPY23 data (Purps et al., 2014). For each of the $n!/[(n-x)!x!]$ subsets of $x$ of the $n$ loci available, we estimated $\theta$ using the methods of Buckleton et al. (2016). There is clear evidence that these estimates are neither constant over the number of loci, nor are they products of single-locus estimates. The dependence of $\theta$ on the number of loci is complex, and so is the match probability for a Y-STR profile.

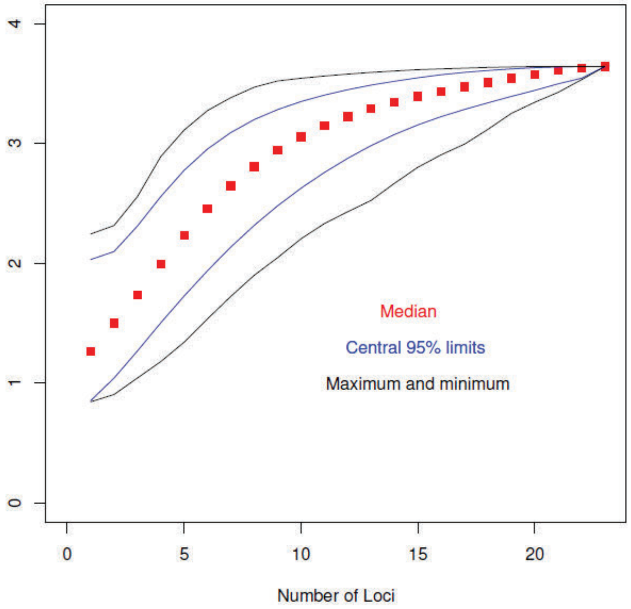

**Figure 1:** Plot of estimated multi-locus $\theta$, on $-\log_{10}$ scale, against the number of loci, for PPY23 data.

The complexities for Y-STR profiles, depending in part on the lack of recombination among Y-STR loci, has been widely recognized and has led to work of the sort described below. The dependence among single-locus match probabilities for autosmal loci has also been recognized for a long time. In the forensic context, Donnelly (1995) said: “after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related”. He provided some theoretical expressions for the degree of dependence. Laurie and Weir (2003) provided both theoretical and empirical demonstrations of dependence, and referred to earlier theory by Cockerham and Weir (1973): “For finite populations, between-locus dependencies can exist even for unlinked loci”. A completely empirical demonstration was given by Weir (2004): "As the number of loci increases, the proportion of cases [sets of loci chosen from a set of nine loci] in which the product rule estimate of the multi-locus number of matches is less than the observed number are: 1/36=0.33 for two loci, 42/84=0.50 for three loci, 88/126=0.70 for four loci and 92/126=0.73 for five loci." At that time we discounted the problem by stating that "under-estimating matching probabilities is prevented by using the products over loci of the match probabilities with $\theta$ greater than zero." Donnelly was concerned with family resemblance, whereas Laurie and Weir were concerned with dependence resulting from finite population size and evolutionary history, as were Balding and Nichols (1994). Evolutionary-dependencies will be much smaller than family-dependencies and we seek to place bounds on them for current data.

Two factors suggest we revisit the issue: firstly there are now more loci being used in forensic science (Hares, 2015). Although the CODIS additional loci were chosen from among those with (relatively) low mutation rate, the CODIS loci generally do have mutation rates of the order of $10^{-3}$. As Laurie and Weir (2003) showed: "the dependency effects increase as the mutation rate increases . . . the between-locus dependency effects are magnified when considering more loci."

## Lineage markers

The lack of recombination among Y-STR loci argues for them not being independent. Judicious choice of loci can reduce dependencies, reflecting the independence of mutation events at different loci. Hall (2016) showed that, among all pairs of Y-STR loci in the three databases she examined, most pairs did not show significant linkage disequilibrium. In Table 1 we show the most diverse of PPY23 loci, ordered by their single-locus entropies, the conditional entropy of each locus when added in the order shown, and the resulting final entropy. For a haplotype $A$ with alternative forms $A_u$ having sample frequencies $\tilde{p}_u$, the entropy is $-\sum_u \tilde{p}_u \ln(\tilde{p}_u)$. This is a useful measure of diversity for the set of loci represented in the haplotype, although it differs from the sample average match probability $\sum_u \tilde{p}_u^2$ and does not solve the forensic issues (Caliebe et al., 2015). Nevertheless, Table 1 shows the diminishing benefits of adding additional Y-STR loci: there is little change in combined entropy beyond 10 loci (read across columns for each row.)

**Table 1:** Entropy measures for Y-STR markers.

| Added | Entropy | | | Added | Entropy | | |
|---|---|---|---|---|---|---|---|
| Marker | Single | Combined | Conditional | Marker | Single | Combined | Conditional |
| YS385ab | 4.750 | 4.750 | 4.750 | DYS481 | 2.962 | 6.972 | 2.222 |
| DYS570 | 2.554 | 8.447 | 1.474 | DYS576 | 2.493 | 9.318 | 0.871 |
| DYS458 | 2.220 | 9.741 | 0.423 | DYS389II | 2.329 | 9.906 | 0.165 |
| DYS549 | 1.719 | 9.999 | 0.093 | DYS635 | 2.136 | 10.05 | 0.053 |
| DYS19 | 2.112 | 10.08 | 0.028 | DYS439 | 1.637 | 10.10 | 0.024 |
| DYS533 | 1.433 | 10.11 | 0.010 | DYS456 | 1.691 | 10.12 | 0.006 |
| GATAH4 | 1.512 | 10.12 | 0.005 | DYS393 | 1.654 | 10.13 | 0.003 |
| DYS448 | 1.858 | 10.13 | 0.002 | DYS643 | 2.456 | 10.13 | 0.002 |
| DYS390 | 1.844 | 10.13 | 0.002 | DYS391 | 1.058 | 10.13 | 0.002 |

## NGS Data

"Massively parallel sequencing (MPS) is adding a new dimension to the field of forensic genetics, providing distinct advantages over CE [capillary electrophoresis] systems in terms of captured information, multiplex sizes, and analyzing highly degraded samples. In recent years, MPS has been applied to the generation of STR sequence data with the general outcome that STRs can be successfully typed producing genotypes compatible with those of CE analyses, even from compromised forensic samples. Furthermore, MPS derived STR genotypes provide additional information to that generated by CE separation by capturing the full nucleotide sequence underlying the repeat units and nearby flanking regions. It was demonstrated by earlier studies using mass spectrometric (MS) systems that the discrimination power of STR typing could be increased by differentiating the nucleotide sequences of alleles with identical size. With MPS, forensic tests will further discern STR variants that cannot be distinguished by MS, e.g. repeat motifs that are shifted relative to each other in the repeat region. Early assessments of MPS STR typing show it will be highly beneficial to routine casework by increasing the discrimination power, improving resolution of mixtures, and enhancing the identification of stutter peaks and artifacts." (Parson et al., 2016)

Parson's listing of the many advantages of NGS data did not include the possibility of including single nucleotide polymorphism (SNP) markers in typing kits to provide additional information on phenotype and/or ancestry (e.g. Illumina ForenSeq, Promega PowerSeq kits). We will not consider here the considerable potential of large-scale SNP array data for phenotypes such as face morphology (http://parabon-nanolab.com).

The key finding has been the ability to infer the underlying STR genotypes when STR loci are sequenced. The STRait Razor software (Warshauer et al., 2015) has proven useful in this regard. We have begun an investigation of how this approach may be enhanced in terms of quality and speed. We have used $C^{++}$ with an $R$-wrapper rather than *Perl* used by STRait Razor. We also modified the method of scoring the alignments between STR-region sequence data and reference sequences for STR alleles.

# Project Design and Execution

## Population Structure and Relatedness

We have developed a new approach to estimating $\theta$ from published STR allele frequencies and we applied that method to a worldwide survey we extracted from 250 publications (Buckleton et al., 2016). Although our findings were broadly consistent with those of other authors, we did suggest that somewhat-larger values be used than has been the practice. In Table 2 we summarize some of our results. There is considerable variation of estimates over loci. Averaging over loci shows the smallest $\theta$ (0.0038) is for the set of African populations - the most diverse of human populations. The largest value (0.1050) is for the small number of quite homogeneous Inuit populations. We stressed that estimates depend on the reference set of populations: in Table 2 the reference is the entire set of 446 populations in the survey.

The estimates we show were values of $F_{ST}$ calculated as $(\tilde{M}_W - \tilde{M}_B)(1 - \tilde{M}_B)$ and is the appropriate quantity to use for $\theta$ when allele frequencies are taken from a database representing all populations in the reference set. The estimate uses two sample proportions of matching pairs of alleles: $\tilde{M}_W$ within populations and $\tilde{M}_B$ between pairs of populations.

For the set of African populations, the average within-population matching proportion was $\tilde{M}_W = 0.1884$ and the average between-population-pair averages were $\tilde{M}_B = 0.1691$ within the African region and $\tilde{M}_B = 0.1726$ for all pairs of populations. There is a larger $F_{ST}$ for the set of African populations ($\hat{\beta}_W = 0.0082$) with Africa as a reference set than there is ($\hat{\beta}_W = 0.0038$) with the world as a reference set. The opposite was found for a collection of Inuit populations: the average within-population matching proportion was $\tilde{M}_W = 0.4379$ whereas the average between-population-pair matching proportions were $\tilde{M}_B = 0.1726$ for pairs within the Inuit group and $\tilde{M}_B = 0.0090$ for all pairs in the study: so $F_{ST}$ is less with Inuit as a reference set ($\hat{\beta}_W = 0.0205$) than with the world as a reference set ($\hat{\beta}_W = 0.1050$).

**Table 2** $\theta$ estimates from world-wide survey.

| | Africa | AusAb | Asian | Caucn | Hisp | IndPk | NatAm | Inuit | Polyn | World |
|---|---|---|---|---|---|---|---|---|---|---|
| CSF1PO | -0.0668 | 0.0130 | 0.0154 | 0.0127 | 0.0165 | 0.0197 | 0.0616 | 0.0406 | 0.0291 | 0.0117 |
| D1S1656 | 0.0339 | —— | 0.0658 | -0.0018 | 0.0189 | 0.0316 | —— | 0.0812 | —— | 0.0157 |
| D2S441 | 0.0153 | —— | 0.0265 | 0.0316 | 0.1005 | -0.0285 | —— | 0.1625 | —— | 0.0332 |
| D2S1338 | 0.0029 | 0.0313 | 0.0319 | 0.0129 | 0.0234 | 0.0134 | 0.1210 | 0.1255 | 0.0035 | 0.0292 |
| D3S1358 | 0.0145 | 0.0279 | 0.0578 | -0.0345 | 0.0239 | 0.0227 | 0.2200 | 0.2196 | 0.0426 | 0.0254 |
| D5S818 | 0.0102 | -0.0229 | -0.0132 | 0.0465 | 0.0474 | 0.0197 | 0.1192 | 0.0461 | -0.0243 | 0.0337 |
| D6S1043 | -0.0006 | —— | 0.0126 | 0.0669 | 0.0030 | —— | —— | —— | —— | 0.0233 |
| D7S820 | 0.0244 | 0.0557 | 0.0345 | 0.0001 | 0.0165 | 0.0039 | 0.0842 | 0.0443 | -0.0078 | 0.0222 |
| D8S1179 | 0.0405 | -0.0153 | -0.0187 | 0.0169 | 0.0273 | -0.0207 | 0.0885 | 0.1264 | 0.0227 | 0.0179 |
| D10S1248 | -0.0397 | —— | 0.0383 | 0.0047 | 0.0473 | -0.0195 | —— | 0.1345 | —— | 0.0102 |
| D12S391 | 0.0317 | —— | 0.0448 | -0.0097 | 0.0745 | 0.0258 | —— | 0.0522 | —— | 0.0120 |
| D13S317 | 0.1221 | 0.0806 | 0.0235 | 0.0445 | 0.0051 | 0.0093 | 0.0252 | 0.0990 | 0.0375 | 0.0384 |
| D16S539 | -0.0018 | 0.0597 | 0.0237 | 0.0288 | 0.0093 | -0.0025 | 0.0720 | 0.1635 | 0.0227 | 0.0250 |
| D18S51 | -0.0012 | 0.0064 | 0.0345 | 0.0064 | 0.0026 | 0.0323 | 0.0503 | 0.0733 | 0.0538 | 0.0181 |
| D19S433 | -0.0095 | 0.1661 | 0.0226 | 0.0410 | 0.0053 | 0.0166 | 0.0132 | -0.0013 | 0.0015 | 0.0254 |
| D21S11 | -0.0076 | -0.0225 | 0.0422 | 0.0084 | 0.0126 | 0.0013 | 0.0702 | 0.0492 | 0.0393 | 0.0200 |
| D22S1045 | -0.0626 | —— | -0.0078 | 0.0300 | 0.0872 | -0.0211 | —— | 0.0836 | —— | 0.0204 |
| FGA | 0.0027 | 0.0038 | 0.0183 | 0.0164 | 0.0011 | 0.0072 | 0.0226 | 0.0296 | 0.0655 | 0.0142 |
| PENTAD | -0.0402 | —— | 0.0567 | 0.0180 | 0.0015 | 0.0160 | 0.0380 | —— | —— | 0.0227 |
| PENTAE | 0.0185 | —— | 0.0163 | 0.0235 | 0.0136 | 0.0137 | 0.0409 | —— | —— | 0.0202 |
| SE33 | 0.0234 | —— | 0.0138 | 0.0205 | 0.0152 | 0.0041 | —— | 0.1081 | —— | 0.0219 |
| TH01 | 0.0731 | 0.0679 | 0.1465 | 0.0189 | 0.0369 | 0.0199 | 0.2084 | 0.5200 | 0.0464 | 0.0755 |
| TPOX | -0.1336 | -0.0031 | 0.0911 | 0.0578 | 0.0064 | -0.0369 | 0.0736 | 0.0395 | 0.0412 | 0.0339 |
| VWA | -0.0021 | 0.0246 | 0.0195 | 0.0087 | 0.0373 | 0.0055 | 0.0808 | 0.0231 | 0.0213 | 0.0198 |
| All loci | 0.0038 | 0.0328 | 0.0328 | 0.0193 | 0.0258 | 0.0065 | 0.0804 | 0.1050 | 0.0265 | 0.0244 |

We have recently shown that our population-structure estimates can be used to estimate kinship or coancestry coefficients between pairs of individuals. In essence, each individual is regarded as its own population and our population-level $\theta$ estimates give kinship estimates by reducing sample sizes to one individual/two alleles per population. It is more common to write estimates in terms of allele dosages: if individual $j$ has $X_{ju}$ copies of allele $u$ at a locus the estimated kinship for individuals $j, j'$ is written as $\hat{\beta}_{jj'}$:

$$\hat{\beta}^i_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_B}{1 - \tilde{M}_B}$$

where $\tilde{M}_{jj'} = \sum_u X_{ju} X_{j'u}/4$ and $\tilde{M}_B = \sum_{j=1}^{n} \sum_{j'=1, j' \neq j}^{n} \tilde{M}_{jj'}$ for individuals in a sample of $n$ individuals. For multiple loci, the numerator and denominator are summed separately over loci. These estimates behave much better than standard estimates (Ritland, 1996; Yang et al., 2011):

$$\hat{\theta}_{jj'} = \frac{(X_{ju} - 2\tilde{p}_u)(X_{j'u} - 2\tilde{p}_u)}{4\tilde{p}_u(1 - \tilde{p}_u)} \tag{1}$$

where $\tilde{p}_u$ is the sample allele frequency for the set of $n$ individuals in a study.

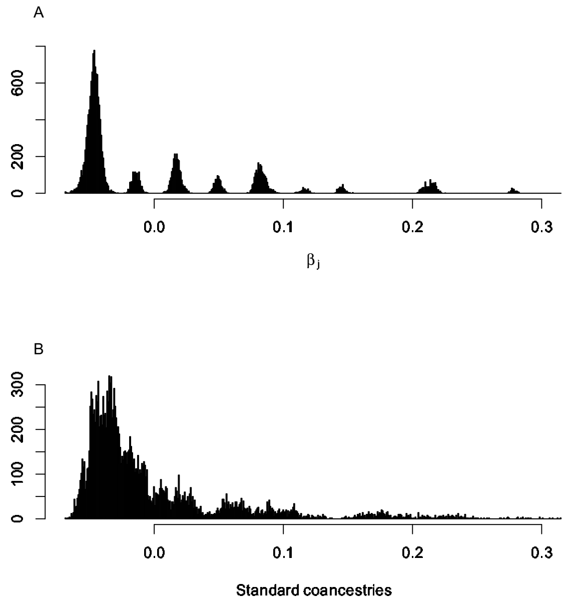

**Figure 2.** Comparison of new and standard kinship estimates.

In Figure 2 we compare our new estimates with the standard ones for a set of SNP data. The true values were $i/32, i = 0, 1, \ldots 10$ and our estimates were clustered around these values (they are all relative to the average over all kinship values for pairs of individuals in the study). Kinship estimation has not been possible with only 20 or so STR loci, but with the coming addition of SNP data we can expect forensic scientists to have the ability to provide meaningful estimates and we will advocate the use of our new procedure, which holds for both STRs and SNPs.

To estimate multi-locus $\theta$ values we need allele-pair-matching proportions for all loci. This is not feasible for 20 or more loci in databases of only a few thousand profiles, and even the plot we showed in Figure 1 suffers from having only one set of 23 loci – no indication is given in that plot of the sampling variation for the right-most value. We will explore the application of the theoretical work of Donnelly (1995) but we are more optimistic about the use of simulation. It is straightforward to set up forward simulations for realistic population sizes under the assumptions of random mating and discrete generations. We can model either autosomal or Y-chromosome STR loci with stepwise mutation, SNP markers with infinite-alleles mutation, or a combination of both. We can keep track of actual identity-by-descent to produce actual multi-locus $\theta$ values. As an illustration, in Table 3 we show joint autosomal and Y-STR $\theta$ values, for a single STR locus and 20 Y-STR loci. The mutation rate was the same at all loci. The joint coancestry $\theta_{AY}$ is the probability that a pair of autosomal alleles, one from each of two men, is identical by descent at the same time as their Y profiles are identical by descent. There are two conditional coancestries: autosomal given Y, $\theta_{A|Y} = \theta_{AY}/\theta_Y$ and Y given autosomal, $\theta_{Y|A} = \theta_{AY}/\theta_A$. This work was motivated by that of Walsh et al. (2006) and it continues our own recent work (Buckleton and Myers, 2014).

**Table 3:** Predicted autosomal and Y-STR $\theta$ values.

| $N$ | $\mu$ | $\theta_Y$ | $\theta_{Y\|A}$ | $\theta_A$ | $\theta_{A\|Y}$ |
|---|---|---|---|---|---|
| $10^4$ | $10^{-3}$ | 0.00244 | 0.00370 | 0.01233 | 0.01868 |
| | $10^{-4}$ | 0.02434 | 0.02447 | 0.11110 | 0.11168 |
| $10^5$ | $10^{-3}$ | 0.00024 | 0.00151 | 0.00125 | 0.00768 |
| | $10^{-4}$ | 0.00249 | 0.00262 | 0.01234 | 0.01300 |
| $10^6$ | $10^{-3}$ | 0.00002 | 0.00129 | 0.00012 | 0.00656 |
| | $10^{-4}$ | 0.00025 | 0.00038 | 0.00125 | 0.00191 |

It is profile match probabilities, rather than $\theta$ values, that are of forensic relevance. For moderate numbers of loci we can compare empirical matching proportions with those predicted by a combination of haplotype frequencies and $\theta$ values. For autosomal profiles we expect to be able to use the Balding-Nichols approach with $A$ representing a gamete (one allele per locus) and summing over pairs of gametes consistent with the profile of interest. For larger numbers of loci, and rare profiles, we will investigate the use of $\theta$ as a default value for any rare profile and (combinations of) $6\theta^2/[(1+\theta)(1+2\theta)]$ for autosomal homozygotes and $2\theta^2/[(1+\theta)(1+2\theta)]$ for autosomal heterozygotes. The combinatorial issues are analogous to those discussed in the next section for lineage profiles. This use of the same value for all profiles with the same number of loci is consistent with the approach of Brenner (2010, 2014) but our value does depend on the number of loci, and accommodates population structure (Buckleton et al., 2016). The advantage is the avoidance of the assumption of independence of loci.

## Lineage Markers

Y chromosome STR typing is often used when autosomal typing has failed. As such it often presents mixed and low template profiles.

Consider a two person mixed profile at 21 loci. Each locus will show 0,1,2 peaks depending on masking and drop-out. There may be additional peaks from drop-in. Without consideration of drop-out and drop-in every locus showing two peaks results in two possible combinations of haplotypes. If there are 21 of these this is $2^{21} = 2,097,152$. Consideration of drop-out and drop-in greatly increased this number.

Because of the high diversity of the Y chromosome, databases of realistic size are poor at informing haplotype probabilities. Many of the haplotype pairs mentioned above will have one or both haplotypes unrepresented in the database and hence return a sample estimate of zero. Haplotypes in modern population databases display substructure. Say more. Bruce some comments about these being genetic entities will do.

The SWGDAM Y chromosome working group has highlighted the need for strong interpretation tools for such mixtures. Exploratory work in this area has highlighted the following:

1. The need for a computationally efficient and foundationally valid method for informing haplotype probabilities for haplotypes observed rarely or not at all in a database, and

2. The probable need for use of MCMC and importance sampling methods to give realistic run times for casework mixture interpretation.

As previously noted, the space of possibilities for haplotype sets –sets of haplotypes corresponding to one or more contributors – can be extra-ordinarily large. This means that mixture interpretation, which involves summation over all possibilities will typically be impossible to carry out in finite time. A traditional statistical approach to such problems is to employ sampling methods. That is, we sample a large, but finite, subset of the haplotype sets with replacement, evaluate the probability of the evidence with respect to each of these sets and average. If we are using autosomal markers, then sampling genotype sets with probability proportional to the frequency of the alleles present at each locus, on a per locus basis works reasonably well. This is because the space of genotype sets at a single locus is relatively small, and therefore simple random sampling will cover most of the possibilities. However, the same cannot be said when we consider haplotype sets. The inability to treat loci as independent means we would have to consider whole haplotypes rather than locus specific haplotypes. The resulting estimates, although unbiased, would have remarkably poor precision. Important sampling and MCMC based methodology offer us two potential solutions.

Importance sampling is a relatively simple idea. Imagine that we want to sample from a probability distribution with probability (density) function $f(x)$, but we are only interested in a subset of the outcomes for which $f(x)$ is defined, and these outcomes have very low probability. Importance sampling proceeds by considering an importance density, say $h(x)$, which assigns much higher probability to the outcomes of interest. If we sample from $h(x)$, then we will see many more instances of the outcomes we are interested in than if we sampled from $f(x)$. To counter this over-sampling we simply re-weight the observations we sample by the ratio of $f(x)$ to $h(x)$. This is the essence of importance sampling. It biases the sampling towards the outcome(s) of interest, and then re-weights the sample. Estimators based on the resulting sample will be unbiased and generally have smaller precision than a simple random sample. This that for an equivalent sized sample, the importance sample estimate of the probability of the evidence given the haplotype sets will generally be considerably more accurate than that obtained through simple random sampling.

Monte Carlo Markov Chain based methods are another approach to estimation in spaces with many many outcomes. In order to make this relevant, we describe here how we envisage this method working for this problem. Again we start with the idea that the space of haplotype sets is huge. We need to estimate:

$$\Pr(E|H) = \sum_{s\in S}\prod_{i=1}^{r}\Pr(R_i|s)\Pr(s)$$

where $R_i$ is the $i^{th}$ repeat measurement (PCR analysis) of the evidential stain, and $s$ is a particular haplotype set in the space of haplotype sets $S$. We are unable to sum over all possible value of $s$, but we might be able to approximate this sum by choosing a random sample of sets that covers those possibilities which explain the evidence well, and not spending much time dealing with possibilities that dont. A potential method would be to propose a particular haplotype set, $s_0$ at random (perhaps with uniform probability). We then evaluate

$$L_0 = \prod_{i=1}^{r}\Pr(R_i|s_0)\Pr(s_0)$$

We then propose $N$, where $N$ is large, alternative sets. At each proposal we randomly select $s_1$, and evaluate

$$L_1 = \prod_{i=1}^{r}\Pr(R_i|s_1)\Pr(s_1)$$

If $L_1$ is larger than $L_0$ or, a random value $u \sim U[0,1]$ is less than

$$\min(1, L_1/L_0)$$

then we store $s_1$ and $L_1$, set $s_0 = s_1$ and $L_0 = L_1$ otherwise we store $s_0$, $L_0$. For a large enough value of $N$, the average over all stored values of $L$ should approximate the probability of interest, i.e.

$$\Pr(E|H) \approx \frac{1}{N}\sum_{I=1}^{N}\prod_{i=1}^{r}\Pr(R_i|s_l)$$

The idea in proposing a new haplotype set is to choose one that is reasonably similar to the one under consideration. For example we might alter the alleles observed at just one locus. This produces correlated samples (and hence we need to take a large sample to deal with the inefficiency of the estimator), but it also means that we more spend time considering haplotype sets that are more probable (i.e. good explanations for the evidence), than we would if we used simple random sampling. We propose to continue our use of these techniques as we consider Y-STR mixtures. Such an approach will not be very precise because of the sheer size of the space. However, both importance sampling and MCMC are strategies which have some promise here, and require some investigation. We have extensive experience in the implementation of such approaches as Curran wrote the first commercial program, LoComatioN, that implemented the semi-continuous model for the United Kingdom's Forensic Science in 2004/2005. We have other experience with MCMC methodology used in the STRMix package.

We have previously (Buckleton et al., 2011) reviewed the various counting and $\kappa$ (Brenner, 2010, 2014) methods for estimating haplotype profile or match probabilities. We can also mention the Good method and the Discrete-Laplace method. The Generalized Good method (hereafter the Good method) is based on work by Good (1953). This method calculates a likelihood ratio (LR) rather than a haplotype probability or a match probability. A derivation of the LR for the Good method was given by Cereda (2015). Considering two matching haplotypes (one evidential and one reference from the defendant) the propositions of the LR are:

$H_p$: The defendant left the crime stain.
$H_d$: Someone other than the defendant left the crime stain.

If the haplotype is unobserved, then the probability of the evidence under $H_p$ is the probability of observing a singleton (of which there are $n_s$ singletons) in the database. That is $\Pr(E|H_p) = n_s/n$. Under $H_d$, the probability of the evidence is the probability of observing a matching pair $(n_d)$ out of all possible pairwise comparisons is $\Pr(E|H_d) = 2n_d/n(n-1)$. The rare haplotype LR is therefore:

$$\mathrm{LR} \;=\; \frac{(n-1)n_s}{2n_d} \approx \frac{nn_s}{2n_d}$$

Following a similar rationale, for haplotypes that have been observed $x$ times in the database, the LR is:

$$\mathrm{LR} \;=\; \frac{(n-x-1)n_{s+1}}{(x+2)2n_{s+2}} \approx \frac{nn_{x+1}}{(x+2)n_{x+2}}$$

where $n_{x+1}$ is the number of groups of matching haplotypes of size $(x+1)$ within the database. Note that by setting $x = 0$ the LR simplifies to the rare haplotype LR.

The discrete Laplace method (Anderson et al., 2013) also uses the idea of changes in genetic diversity over time through mutation to describe modern populations. The method starts with the idea that there were one (or more) ancestral haplotypes, and through mutation we have arrived at the distribution of haplotypes we observe currently. The method makes three assumptions:

1. A population of haplotypes is composed of clades of haplotypes,

2. Each clade has arisen from one ancestral haplotype by stepwise mutation,

3. Mutations occur independently of each other.

It is possible, given these assumptions, to multiply probabilities of allele differences from a central haplotype across loci within the haplotype clade to assign a probability to a full haplotype in that clade. Hence, given the clade (or within the clade), the loci are assumed to be independent. The probability of the haplotype in the full population is assigned by weighting the assignment from each clade by the estimated contribution of that clade to the total population. The method takes its name from the discrete Laplace distribution which is used to calculate the haplotype probabilities. The Discrete Laplace method has great advantage in that it can work on a per-locus basis. It requires moderate computational effort initially, to compute the distribution parameters, but after this has been done, computation in a specific case is substantially lower.

We are also considering a method we term The Approximate Product Method (APM). This attempts to model the dependency between loci through the continued application of an approximate correction. We start with a problem involving a single haplotype described by alleles at just two loci. If we treat these loci as being completely independent, then the haplotype probability is given by $\Pr(x_1)\Pr(x_2)$ where $x_1$ and $x_2$ are the alleles at the two loci. Similarly if we regard these loci as being completely dependent, then the haplotype probability is given by $\Pr(x_1, x_2) = \Pr(x_1)$. If we write $\alpha$ for the probability that these two loci are dependent, then using the law of total probability we can express the haplotype probability as

$$\Pr(x_1, x_2) = (1-\alpha)\Pr(x_1)\Pr(x_2) + \alpha\Pr(x_1) = \Pr(x_1)[\alpha_{2|1} + (1-\alpha_{2|1})\Pr(x_2)]$$

to define $\alpha_{2|1}$. For three loci, by extension,

$$\Pr(x_3|x_1, x_2) = \alpha_{3|1,2} + (1-\alpha_{3|1,2})\Pr(x_3)$$

If we proceeded in this fashion for all loci, then we would end with many $\alpha$ terms and a significant estimation problem. To manage the number of terms we posit that there is some single value of $\alpha$ that will allow the estimation process across all loci to perform credibly. In other words, for an $L$-locus haplotype

$$\Pr(x_1, x_2, \ldots x_L) = \prod_{l=1}^{L}\{1-(1-\alpha)^{l-1}[1-\Pr(x_l)]\}$$

The motivation for this approximate method is that it allows calculation of haplotype probabilities on a per-locus basis, and it has a very simple computational form. It is admittedly *ad-hoc* but we plan to evaluate it through simulation.

These various methods/models range from using naive to complex population genetics. However, there is no agreement in the forensic community about which best quantifies the evidence of matching Y-STR profiles and some further evaluation seems worthwhile.

Further, we plan to investigate the application of Chow-Liu trees (Chow and Liu, 1968) for the estimation of Y-STR haplotype frequencies. Chow-Liu trees attempt to reconstruct or approximate, a discrete multivariate probability function, The Chow-Liu method describes a joint probability distribution as a product of second-order conditional and marginal distributions. For example, the six-dimensional distribution might be approximated as

$$\Pr(X_1, X_2, X_3, X_4, X_5, X_6) = \Pr(X_6|X_5)\Pr(X_5|X_2)\Pr(X_4|X_2)\Pr(X_3|X_2)\Pr(X_2|X_1)\Pr(X_1)$$

where each new term in the product introduces just one new variable, and the product can be represented as a first-order dependency tree, as shown in Figure 3.

This kind of factorization exists in the field of Bayesian Networks. Chow and Liu (1968) provide a simple algorithm for constructing the optimal tree. At each stage of the procedure the algorithm simply adds the maximum mutual information pair to the tree, where mutual information between a pair of random variables is given by

$$I(X,Y) \quad = \quad \sum_{x \,\mathrm{in}\, X} \sum_{y \,\mathrm{in}\, Y} \Pr(x,y) \log\left(\frac{\Pr(x,y)}{\Pr(x)\Pr(y)}\right)$$

The Chow and Liu algorithm gives an approximation which is the minimum Kullback-Leibler distance from the true distribution.

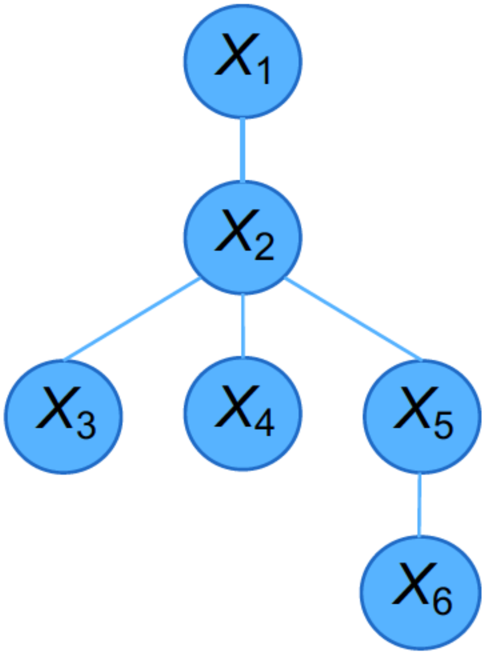

**Figure 3:** Example of Chow-Liu Tree.

**Y-mixture models.** There is a great need in the forensic community for statistical software that helps with the interpretation of mixed Y-haplotypes. Initially we plan to apply semi-continuous interpretation methods to Y-STR mixtures and later move to fully continuous mixtures. Gill et al. (2000), and Balding and Buckleton (2009) proposed a model for autosomal DNA profiles which indirectly accounted for peak height information by the inclusion of terms that modeled the phenomena of DNA dropout, and drop-in (contamination). We call these models semi-continuous because they weight each putative genotype by a probability term between 0 and 1. For example, if we observed alleles 11,12,13 in a crime scene sample and we thought that these originated from two contributors with genotypes 11,12 and 13,14 respectively then we would weight the probability of the respective genotypes by an additional term

$$\Pr(R|G_i) \quad = \quad \Pr(11,12,13|11,12,13,14) = \Pr(\bar{D})^3 \Pr(D) \Pr(\bar{C})$$

This equation models the idea that the alleles 11, 12, 13 have all not dropped out with probability $\Pr(\bar{D})^3$ , the 14 allele from the second contributor has dropped out with probability $\Pr(D)$, and there are no drop-in alleles which occurs with probability $\Pr(\bar{C})$. These methods are useful because they explicitly model stochastic phenomena we know exist in every PCR reaction. A fully continuous model takes its name from the fact that it both assigns a probability to each putative genotype set considered, and that this weight is calculated directly from height information taken from the electropherogram (epg). The heights of peaks at allelic positions in the epg are thought to be proportional to the amount of genetic material contributed to the stain by each donor. Modeling this information directly rather than making arbitrary calls of drop-in, drop-out, or stutter is inherently appealing to all forensic geneticists because it potentially removes one or more source variability, namely that from analysts, from the interpretation process. Neither model has been successfully implemented for Y-STR case work.

The rate-limiting step for using such models (or more complex models), is primarily, that the haplotypes cannot be considered locus by locus, because the loci are not independent. This means that we have to consider whole haplotypes at once. Complete enumeration of all haplotypic combinations may be possible in simple situations but grows combinatorially with the addition of multiple contributors and multiple unknowns. Consider a simple case in which we consider a crime scene stain which appears to be a mixture of two individuals. It is alleged by the prosecution that two individuals identified by a complainant, and who have haplotypes which could explain the crime scene stain, are the only contributors to this stain. The defense argue that it is two random individuals. To consider all possible haplotypes across 20 or more loci would require evaluation over billions of combinations  in the order of $10^{10}$. Such a problem is not currently computationally feasible, and equally, it may not make sense to consider all possible haplotypes because different haplotypes do not arise from random recombinations of alleles. They arise because of mutation events. To make this problem tractable we need to consider efficient schemes for touring the haplotype combination space. Two methods that might offer some way forward are importance sampling and Markov Chain Monte Carlo as described above.

## NGS Data

In the early 90s the US forensic community engaged in what was described as the DNA wars. These debated what population genetics assumptions were appropriate for assigning DNA match probabilities. The large number of genotypes present at each STR locus makes testing for Hardy-Weinberg and linkage equilibrium difficult. An improvement in testing such models came with our realization (Weir, 2004) that partially matching profiles in databases can be used to assess the performance of population genetic models. This was undertaken by counting the number of various classes of partially matching profiles. These observed counts were compared with those expected under various models. Generally conservative behavior of the models was noted. With NGS data two new challenges will emerge.

First, NGS provides many more genotypes, meaning that genotype arrays even larger and more sparsely populated. Second, up to half of the new diversity is in the flanking regions. These variants have a lower mutation rate and, consequently, different $\theta$ values. Countering these issues, however, is the promise of improved mixture interpretation.

Modern interpretation strategies require an understanding of the behavior of allele and stutter peak heights. Peak height translates in the NGS setting into read counts. Appropriate models require the expectation and variance about that expectation. For CE data the expectation has been modeled using degradation, template, locus effects and stutter. The variation is modeled as inversely proportional to the template. Current NGS protocols loosen the relationship between template and signal (read counts) and this seriously challenges current models.

The maintenance of a relationship between template and signal is important for the interpretation of DNA mixtures. Qualitative work suggests that relative template at each locus within a sample retains a relationship to relative signal at each locus. However, between-locus relationships and relationship to absolute values for template are adversely affected. The first step that we discuss in this regard is the library preparation stage. At this stage a library is built from various amplifications of DNA extracted from samples. These various amplifications are quantified and a target DNA amount is carried forward into the library. This has the effect that high template or low template (as judged at extraction) and now normalized to similar values. Hence the models based on low relative variance for high template and low relative variance for high template (Bright et al., 2013) are likely to be challenged. The other factor suspected of influencing the relationship between template and signal is purification. Most protocols currently use beads for the purification stage. These beads have an optional molecular weight (say 200bp) that adheres to them. Amplicons near to 200bp are preferentially enriched. This is likely to superimpose a distribution of unknown form but potentially normal centered on 200 on top of the normal exponential degradation curve. The outcome is not yet investigated quantitatively.

If we lose or loosen the relationship between signal and template, existing models will be inappropriate. A parallel consequence is that the value of any shift to sequencing for mixtures now represents a gain from the extra discrimination and a loss due to less information from signal strength. This is an important matter to quantify. [redacted] is exploring changes to the protocols to restore this relationship and we are in discussions with him.

The current model for STR stutter ratio suggests that stutter is proportional to the sum of the maximum of each sequence length in the STR less a lag and zero. This model has an intuitive molecular biological interpretation. The polymerase enzyme does not start to stutter until after a number of uninterrupted repeats. For the NGS situation we will test the suitability of this model and to test whether the changes to the protocols have restored the relationship between template and signal.

In addition we can apply a new test to the model. Consider for example one of the variants of the 20.2 allele at SE33: $[AAAG]_2$ AG $[AAAG]_3$ AG $[AAAG]_{11}$ AA AAAG $[AAAG]_8$ G AAGG $[AAAG]_2$ AG. The lag for SE33 is thought to be about 5. Hence the sequences of two and three repeats contribute nothing and we should see no reads with stutters in these places. However the sequences of 8 and 11 should both stutter and the stutters should be in the ratio of (11-5) to (8-5). This can be directly tested with new NGS data, and we have a collaboration with [redacted] at [redacted] to examine such data. Our work with [redacted] will inform the work we propose for this project. We are proceeding to model stuttering and the read-count process. The need for a new stutter model is illustrated by these two possible stutter products of allele 9.3 from TH01:

**A** $[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_2$ and **B** $[\mathrm{AATG}]_5\mathrm{ATG}[\mathrm{AATG}]_3$.

Suppose that stuttering causes the loss of one repeat sequence (one step loss). In CE, stutter product of allele 9.3 would have been designated as 8.3, while in NGS, two stutter products are possible: stutter A and B. In the current continuous model, the stutter ratio $\pi_l$ for locus $l$ is a function of the longest uninterrupted sequence (LUS). The following linear model describes the relationship between $\pi$ and the explanatory variables LUS and locus, $l$:

$$\pi_l \quad = \quad \beta_{0,l} + \beta_{1,l} \times \mathrm{LUS}$$

TH01 allele 9.3, has the sequence $[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_3$ and hence a LUS of 6. The stutter ratio fits the regression on LUS much better than if stutter is regressed is on the repeat number. This shows the importance of knowing the sequence to determine the appropriate alleles LUS values, and the advantage of NGS over CE.

The stutter modeling needs to be modified for NGS data and this affects the expected stutter peak height $E^l_{(\alpha-1)n}$, now considered the total expected stutter peak height $E^l_{\alpha n}$. For the THO1 example:

$$E_{9.3n} \quad = \quad \phi^{9.3}_A E_{An} + \phi^{9.3}_B E_{Bn}$$

where $\phi_A$ is the proportion of type A stutter product.

# Potential Impact

The criminal justice significance of the proposed work lies in the enhanced interpretation of DNA profiles for human identification, with an emphasis on data being generated by new technologies.

The proposal challenges current practice by questioning the assumption that autosomal profiles are independent across loci, and that STR electropherograms can give unambiguous present/not present determinations for all (potential) alleles in a profile.

The methodology described in this application is novel, and has been developed by the investigators. The proposal is to continue and evaluate these theoretical advances.

The impact of the proposed research will be achieved primarily through the resulting publications, with one measure of impact being citations. [redacted] has over 25,000 citations to his work ($h$ index of 56), [redacted] has 2,500 citations ($h$ index of 26) and [redacted] has over 2,000 citations ($h$ index of 23).

Although this application does not seek funding for training activities, the research will inform the training activities undertaken by the investigators. [redacted] has initiated a new undergraduate course in forensic genetics at the [redacted], and he teaches an online graduate course every second year through the [redacted] on statistical and genetic aspects of forensic DNA typing. Together with [redacted] he teaches a module in forensic genetics at the annual [redacted] at the [redacted] [redacted] has been involved in short course delivery in [redacted] and [redacted], [redacted] and the [redacted]. [redacted] teaches a forensic genetics course at the [redacted]. He and [redacted] supervise Masters and PhD research in forensic science at that university.