Forensic Science International: Genetics 19 (2015) 107–122

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Probabilistic characterisation of baseline noise in STR profiles

Ullrich J. Mönich [a,*], Ken Duffy [d], Muriel Médard [a], Viveck Cadambe [c], Lauren E. Alfonse [b], Catherine Grgicak [b]

[a] Research Laboratory of Electronics, Massachusetts Institute of Technology, United States
[b] Biomedical Forensic Sciences, Boston University School of Medicine, United States
[c] Department of Electrical Engineering, Pennsylvania State University, United States
[d] Hamilton Institute, Maynooth University – National University of Ireland Maynooth, Ireland

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 22 May 2015
Accepted 2 July 2015
Available online 8 July 2015

Keywords:
Short tandem repeat
Noise
Peak height
Distribution
G-test
Stutter

ABSTRACT

There are three dominant contributing factors that distort short tandem repeat profile measurements, two of which, stutter and variations in the allelic peak heights, have been described extensively. Here we characterise the remaining component, baseline noise. A probabilistic characterisation of the non-allelic noise peaks is not only inherently useful for statistical inference but is also significant for establishing a detection threshold. We do this by analysing the data from 643 single person profiles for the Identifiler Plus kit and 303 for the PowerPlex 16 HS kit. This investigation reveals that although the dye colour is a significant factor, it is not sufficient to have a per-dye colour description of the noise. Furthermore, we show that at a per-locus basis, out of the Gaussian, log-normal, and gamma distribution classes, baseline noise is best described by log-normal distributions and provide a methodology for setting an analytical threshold based on that deduction. In the PowerPlex 16 HS kit, we observe evidence of significant stutter at two repeat units shorter than the allelic peak, which has implications for the definition of baseline noise and signal interpretation. In general, the DNA input mass has an influence on the noise distribution. Thus, it is advisable to study noise and, consequently, to infer quantities like the analytical threshold from data with a DNA input mass comparable to the DNA input mass of the samples to be analysed.

1. Introduction

Short tandem repeat (STR) allele signal interpretation is a central tool in forensic analysis, as the number of repeated copies of a basic motif at given loci can be used to uniquely identify an individual. Currently, forensically relevant STR loci are processed via capillary electrophoresis. The relative intensity of the fluorescent signal is an indication of the number of amplicons in the sample. The transit time of the amplicons through the capillary is utilised to determine the size of the fragment, which is translated into the number of repeat units, i.e. allele. A DNA STR profile will likely contain the following elements: the desired DNA signal, disturbance due to the instrument, non-specific amplification, or other reasons, often termed noise, and artefacts, such as stutter, -A, and pull-up [1]. The interpretation of DNA profiles, especially from low DNA input mass samples, is complicated, because of the distortion of the DNA signal by the latter elements.

The effects of noise in STR profiles are usually suppressed by applying an analytical threshold to the data [2–6]. All peaks smaller than the analytical threshold, which is also referred to as detection threshold, or minimum distinguishable signal threshold, are considered indistinguishable from noise and thus are ignored. Further, a second threshold, the stochastic threshold, has been used in casework. The stochastic threshold is defined as the value above which it is reasonable to assume that allelic dropout has not occurred within a single-source sample [7].

Recently, it was shown that the inclusion of peak height information results in some improvement over interpretation methods that use a drop-out probability for low-template samples [8]. Interpretation methods that utilise peak height or area information, rather than the binary information of whether a peak height or area exceeds a threshold or not, have garnered much attention, and a variety of fully-continuous interpretation platforms have been developed [9–16].

The peak area information is used in [9–12] to establish a quantitative method for the interpretation of STR mixtures. In

* Corresponding author. Tel.: +1 6172536171.
E-mail addresses: moenich@mit.edu (U.J. Mönich), Ken.Duffy@nuim.ie (K. Duffy), medard@mit.edu (M. Médard), viveck@engr.psu.edu (V. Cadambe), alfonse@bu.edu (L.E. Alfonse), cgrgicak@bu.edu (C. Grgicak).

http://dx.doi.org/10.1016/j.fsigen.2015.07.001
1872-4973/

these methods the peak areas are usually modelled as Gaussian random variables and baseline noise is not handled explicitly. Sometimes, noise is modelled implicitly as small peaks with mean zero [10,12]. In [13], a continuous modelling of the peak height is used if the peak exceeds a user defined analytical threshold. A quantitative computer evaluation of DNA mixtures is proposed in [15,16]. In the latter work, peak heights are modelled as truncated Gaussian random variables and baseline noise is implicitly treated as small peaks with mean zero. Although these continuous methods use peak height information, an analytical threshold is typically still applied.

Taking the idea of the continuous methods further, it is consequent to remove the analytical threshold as well, and to consider the information in the peaks that were previously below the analytical threshold. The potential of this approach is illustrated in [17], where the authors present an algorithm to determine the number of contributors to a DNA mixture. They compare the maximum allele count method to their method, which uses no threshold, and show that the latter results in increased accuracy for samples with low DNA input mass. However, little is known about the properties of these peaks. In order to derive a meaningful signal model, it appears necessary to study the distribution of the baseline noise.

Even in settings where an analytical threshold is applied, the information of the noise peak height distribution is of relevance because it can be used to specify the analytical threshold. The fact that the conservative nature of analytical thresholds has been criticised led [18] to analyse the baseline noise. Information loss may be kept low by applying an optimal analytical threshold that suppresses most of the noise while enhancing the detection of alleles [6]. However, a common validation platform or best practise have yet to be established for this purpose [19]. In order to establish a best practise, a detailed understanding of the behaviour of noise appears necessary.

Noise in STR profiles is commonly modelled via Gaussian distributions [18,17], though log-normal distributions have also been suggested [6]. In [17], a Gaussian noise model is used for determining the likelihood that a given number of individuals contributed to a mixed sample. Although utilisation of the Gaussian model to describe noise provides improved identification over previous techniques, the authors do not provide analysis independent of determining the most likely number of contributors to confirm its appropriateness. In this paper, we revisit this premise.

Using data sets created at a range of DNA input masses using both the Identifiler Plus kit and the PowerPlex 16 HS kit, we first ask in Section 3.2 if there is a mass dependent per-dye colour representation of the noise. This hypothesis is rejected, particularly at high input masses. That is, there is no single statistical distribution that can jointly describe the noise observed at distinct loci measured in the same dye colour. As a result, it is necessary to characterise noise on a mass dependent per-locus basis. In Section 3.4, using statistical analysis, we assess three classes of distributions, Gaussians, log-normals and gamma distributions, for their ability to describe measured noise data.

2. Methods

2.1. Data generation and representation

There are different ways to represent the information contained in an electropherogram. In an idealised signal, i.e., in a signal with no artefacts or noise, each peak corresponds to an allele that is present in the DNA sample. Hence, it is possible to specify each peak location by a tuple consisting of locus and allele name.

We choose a vector representation of the data, similarly to [20]. To obtain the vector representation from the electropherogram peak list, we index all possible allele and stutter positions with an integer from 1 to I, where I denotes the total number of positions. Then we collect the signal peak heights from all the possible positions in a vector $\underline{y} = (y_1, y_2, \ldots, y_I)^{\mathrm{T}}$ and call this vector the measurement vector. If there is no peak detected at a particular position, a 0 is included in the vector $\underline{y}$ at the corresponding index.

The data that are used for the analysis in this paper consist of 643 distinct single person profiles from 60 individuals for the Identifiler Plus kit and 303 distinct single person profiles from 48 individuals for the PowerPlex 16 HS kit. For the IDplus kit, the 643 profiles were generated by targeting an input mass of 0.0078 ng (97 profiles), 0.0156 ng (94 profiles), 0.0313 ng (94 profiles), 0.047 ng (74 profiles), 0.0625 ng (90 profiles), 0.125 ng (97 profiles), and 0.25 ng (97 profiles). For the PowerPlex 16 HS kit, the 303 profiles were generated by targeting an input mass of 0.0078 ng (48 profiles), 0.0156 ng (47 profiles), 0.0313 ng (48 profiles), 0.0625 ng (48 profiles), 0.125 ng (48 profiles), and 0.25 ng (48 profiles).

It is to be noted that additional samples targeting 0.5 ng (45 profiles) and 1 ng (45 profiles) were amplified using both kits, but a substantial number of these resulted in at least one locus with "offscale" peaks, indicating that the detector was saturated. This resulted in only 29 and 14 0.5 ng Identifiler Plus and PowerPlex 16 HS usable profiles, respectively. In the 1 ng case, only 2 Identifiler Plus profiles showed no saturation. Given the number of saturated profiles, the statistical evaluations described below were not conducted on these "higher" template samples.

The data were generated with the GeneAmp PCR System 9700, the 3130 Genetic Analyzer, and the GeneMapper ID-X v1.1.1 software, where a constant RFU threshold of one was used. The injection time was 10 s and the injection voltage 3 kV. In the case of the Identifiler Plus amplification kit, the 29-cycle protocol was used and, in the case of the PowerPlex 16 HS kit, the 32-cycle protocol was chosen. The Identifiler Plus kit includes 15 tetranucleotide STR loci and the PowerPlex 16 HS kit 13 tetranucleotide and 2 pentanucleotide STR loci. All amplifications and sample preparations were performed according to the manufacturers' recommended protocols [21,22].

After generation of the data, all peaks originating from pull-up or -A were manually removed. In order for peaks to be classified as pull-up, the peak in question had to be in the same position (±0.3 bases) as the allelic peak in another colour channel and have a peak height of 5% or less of the allelic peak. Further, if a peak fell between two adjacent allelic peaks in a different dye colour and had a "plateau-like" shape, then the peak in question was classified as complex pull-up. A peak was determined to be -A if it was one base shorter (±0.3 bases) than an allelic peak. There were no height restrictions for the complex pull-up and -A artefacts. For a definition and explanation of the previous terms we refer to [1].

The genotypes of all individuals had been determined beforehand from high DNA input mass samples.

2.2. Data analysis

We enumerate the profiles from 1 to P, where P denotes the total number of profiles that are available for a given kit, and denote the measurement vectors by $\underline{y}^p = (y_1^p, y_2^p, \ldots, y_I^p)^{\mathrm{T}}$, $p = 1, \ldots, P$. The components y_i^p, $i = 1, \ldots, I$, of the measurement vector are called measurement values. Note that the measurement values that are given by the GeneMapper ID-X v1.1.1 software are non-negative integers.

Knowing the genotypes of the individuals, we can group the components $y_1^p, y_2^p, \ldots, y_I^p$ of each measurement vector into three categories: true peak, stutter, and noise. We call a component y_i^p of the measurement vector $\underline{y}^p$ true peak if at index i there is either a homozygote allele or a heterozygote allele. Stutter is an artefact