Forensic Science International: Genetics 37 (2018) 6–12

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Modelling allelic drop-outs in STR sequencing data generated by MPS

Søren B. Vilsen[a,*], Torben Tvedebrink[a,1], Poul S. Eriksen[a,1], Christian Hussing[b,2], Claus Børsting[b,2], Niels Morling[b,2]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Probability of drop-out
Modelling allele coverage
Poisson-gamma distribution
Short tandem repeat
Massively parallel sequencing
Forensic genetics

ABSTRACT

We used a Poisson-gamma model to analyse the allele coverage of autosomal short tandem repeat (STR) systems obtained by massively parallel sequencing (MPS). The Poisson-gamma coverage model was created using the peak height models from capillary electrophoresis (CE) based detection of PCR products as a starting point. The CE models were modified to account for the differences between CE and MPS signals by accounting for the large marker imbalances seen for MPS data and by using the Poisson-gamma distribution instead of the normal, log-normal, or gamma distributions that were applied for CE data. We took two approaches to estimate the marker imbalance parameters by (1) using a work-flow data base, and (2) using the results of replicate investigations of the samples.

The Poisson-gamma model was used to estimate the rate of drop-outs of (1) single contributor dilution series experiments and (2) the minor contributor in two-person mixture samples. We examined the predictive capabilities of the model by comparing the observed and expected Brier scores of each sample. We derived the expected Brier scores and their variances to create asymptotic confidence intervals of the Brier scores. We found that the Poisson-gamma model performed well when using the work-flow data base, but that the replicate approach is not necessarily a viable option.

1. Introduction

DNA recovered at a crime scene is often found in low quantity and may be highly degraded. This affects the efficiency of the PCR and may lead to partial profiles with frequent locus and allele drop-outs, and in some situations also allele drop-ins, due to stochastic effects [1–13]. We will only consider drop-outs. Typically, forensic genetics laboratories define a minimum threshold for acceptance of an allele. If this threshold is not reached, the allele is not detected and the DNA profile will suffer from allele drop-out. Accurate estimation of the probability of allelic drop-out is important since allele drop-outs will lead to mismatches between the DNA profile of the trace sample and the reference sample from a victim, a potential suspect, or an unrelated third party. The methods developed for assessing allelic drop-out of amplified PCR products detected by capillary electrophoresis (CE) cannot be used directly for massively parallel sequencing (MPS) assays. Thus, the methods need to be modified for use in the MPS setting.

Estimation of the probability of allelic drop-out in CE has generally taken one of two forms: (1) direct marker dependent estimation by logistic regression [14,15], or (2) modelling the peak height signal and calculating the probability of being smaller than the threshold under the model [16–20].

The logistic regression model relied on H, the average peak height [14,15]. In a previous paper [21], we concluded that the average allele coverage (equivalent to the average peak height in CE) was not an apt covariate of the probability of drop-out because the libraries were normalised prior to sequencing. While this was the case for the Ion PGM 10-plex data, it was not necessarily the case for the data generated by the Illumina® ForenSeq™ Signature Prep Kit. In our previous paper [21], we suggested the standard deviation of the heterozygote balance as a possible replacement. This hypothesis was tested during the preliminary data analysis. However, the performance was deemed too poor and, thus, not included.

Therefore, in order to estimate the probability of allelic drop-out, we will model the allele coverage. Many of the lessons learned by modelling peak heights in CE should translate when modelling the

* Corresponding author. Permanent address: Skjernvej 4A, 9220 Aalborg East, Denmark.
E-mail addresses: svilsen@math.aau.dk (S.B. Vilsen), tvede@math.aau.dk (T. Tvedebrink), svante@math.aau.dk (P.S. Eriksen), christian.hussing@sund.ku.dk (C. Hussing), Claus.Boersting@sund.ku.dk (C. Børsting), niels.morling@sund.ku.dk (N. Morling).
[1] Permanent address: Skjernvej 4A, 9220 Aalborg East, Denmark.
[2] Permanent address: Frederik V's Vej 11, 2100 Copenhagen, Denmark.

https://doi.org/10.1016/j.fsigen.2018.07.017
Received 21 March 2018; Received in revised form 14 July 2018; Accepted 23 July 2018
Available online 25 July 2018
1872-4973/

coverage, because the data are based on analysis of PCR products. As peak heights are inherently continuous (though often round to nearest integer), they have been modelled using normal, log-normal, and gamma distributions [16–20]. Even though the choice of distribution differs, these models exhibit similar mean and variance structures. We have chosen to focus on the 'gamma model' [22,18,23], which assumes that for sample s and the marker m, the peak height of allele a, H_{sma}, follows a gamma distribution with mean and variance given as:

$$\mathbb{E}[H_{\text{sma}}] = \eta_s \rho_s \sum_{c=1}^{C} \{(1-\xi_s) g_{\text{smac}} + \xi_s g_{\text{sm}(a+1)c}\} \varphi_c, \quad \text{and} \tag{1.1}$$

$$\mathbb{V}\text{ar}(H_{\text{sma}}) = \eta_s \mathbb{E}[H_{\text{sma}}], \tag{1.2}$$

respectively, where c is a contributor to the mixture, C is the total number of contributors to the mixture, ρ_s is proportional to the amount of input DNA, φ_c is the mixture proportion of individual c, ξ_s is the average stutter proportion of the sample, and g_{smac} is the number of alleles a that contributor c has. Thus, g_{smac} is 0, 1, or 2 if the contributor c does not have any allele a, is heterozygous, or homozygous for allele a, respectively.

If the coverage was large enough (for all contributors), the coverage could be modelled using the CE methods, as the continuous models would provide a satisfactory approximation of the discrete coverage. However, if the DNA of the sample is of low quantity, if the sample is degraded, or if the DNA mixture proportions is skewed (e.g. with DNA from one contributor in low quantity compared to that of the major contributor to the mixture), a continuous model would lead to a poor approximation.

Using the mean structure of the gamma model as a starting point, we modelled the allele coverage by a Poisson-gamma distribution (also known as a negative binomial distribution parameterised by its mean). We made modifications to the mean structure because: (1) The MPS data suffers from large marker imbalances, which was accounted for by multiplying the expected coverage by a marker dependent scaling factor, and (2) parental alleles can produce multiple stutters, and a stutter can have more than one possible parental allele [24].

If thresholds have not been applied, the coverage model presented in this manuscript quite naturally accounted for drop-out by its definition. However, as thresholds limit the amount of information, we took two slightly different approaches to estimate the marker dependent parameters of the model, to ensure we had enough information to estimate all the parameters. The first approach relied on a database of reference samples (a workflow database), while the second used multiple replicates of the sample. Given the estimated parameters, we found the probability of allelic drop-out by calculating the probability of the coverage being smaller than the threshold.

The aim of this manuscript is to present an MPS DNA mixture coverage model, called the Poisson-gamma coverage model, and investigate its behaviour when used to analyse DNA samples. The DNA samples analysed in this manuscript were described in Hussing et al. [25]. They also described the marker imbalances, marker drop-out rate, and other characteristics of the samples.

In order to measure the performance of the presented model, we examined how well the coverage model predicted the probability of drop-out. We compared the observed drop-out with the probability of drop-out predicted under the model using the Brier-score. The observed Brier-score was compared to the expected Brier-score in order to investigate whether the coverage model behaved as expected. We believe that the model and thoughts presented in the manuscript may serve as a foundation for the work going forward in modelling STR DNA mixture samples.

2. Materials and methods

2.1. Experimental data

DNA libraries were build using the ForenSeq™ DNA Signature Prep Kit (Illumina®) Primer Mix A and B. Primer Mix A amplifies markers for human identification (HID), while Primer Mix B amplifies the same HID markers plus ancestry informative markers (AIMs) and markers associated with eye and hair colour. DNA sequencing was performed with the MiSeq FGx (Illumina®) as previously described in [25,26].

DNA was extracted from blood samples and buccal swabs collected on FTA cards from 363 individuals. The samples were amplified and sequenced in duplicate.

Dilution series of DNA were created from four contributors. The DNA was amplified and sequenced in triplicate using Primer Mix A. The amounts of DNA in each series were 1 ng, 500 pg, 250 pg, 125 pg, 62.5 pg, 31.25 pg, 15.63 pg, and 7.86 pg. A consensus DNA profile from each individual was generated based on all experiments with 1 ng input.

Fifteen two person DNA mixture samples (in proportions: 1:1000, 1:100, 1:50, 1:25, 1:12, 1:6, 1:3, 1:1, 3:1, 6:1, 12:1, 25:1, 50:1, 100:1, and 1000:1) were made with a male and a female contributor. The total amount of DNA was 1 ng in all cases. The DNA mixtures were amplified and sequenced in duplicate using Primer Mix B. The profiles of the two contributors were known.

Sequences were identified by their flanking regions using STRait Razor v3.0 [27]. They were subsequently trimmed to include just the STR region and the coverage of every unique sequence found. Thus, if two sequences had the same length, but their sequences were different (sometimes referred to as isoalleles), they were treated as different sequences throughout the manuscript.

2.2. Analytic thresholds

We opted for simple marker dependent thresholds, taking a percentage, p, of the maximum coverage of the marker [21]. That is, the threshold of marker m from sample s is:

$$t_{\text{sm}} = p \cdot \max_a \{y_{\text{sma}}\}, \tag{2.1}$$

where y_{sma} is the coverage of sample s, marker m, and allele a. Thus, given a marker where we have observed three strings a, b, and c, with coverage 100, 90, and 4, respectively, i.e. the maximum coverage on the marker is 100. Assuming that $p = 0.05$, then the threshold $t = 5$, implying that the observation c was classified as noise.

We chose this method for establishing the thresholds because it was the simplest method for defining marker dependent thresholds. This method does not change the modelling approach, the determination of the probability of drop-out (other than slight changes to the probabilities, themselves), or the conclusions.

2.3. The coverage model

The coverage of an allele is not continuous. The coverage of a string is a synonym for the count of the string. Thus, it would be natural to model the coverage as count data, i.e. discrete. The most common, and simplest, choice of model for count data is the Poisson distribution that assumes equal mean and variance, which is unrealistic for MPS STR data. We modelled the coverage by using a Poisson-Gamma distribution (see Appendix A), which includes a dispersion parameter to account for the possible deviations from the assumption of equal mean and variance.

We introduce the mean structure of the coverage model in three stages, by (1) introducing the model corresponding to the situation, where there is only one contributor and no artefact, (2) accounting for multiple contributors, and (3) accounting for stutters.