TrueAllele® Methods: Statistical Model

System 3, Version 25
September 2008 model

Mark Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA
8 March 2016

Overview

This document provides scientific background and mathematical formulas for statistical modeling in the TrueAllele system. The document complements previously published descriptions of the hierarchical Bayesian model for genotype separation.

Data

Short tandem repeat (STR) data originate as charge-coupled device (CCD) camera counts that are collected on a genetic analyzer from fluorescently end-labeled DNA fragments as they are separated by size via gel electrophoresis. These multi-spectral CCD signals are isolated by their fluorescent dyes using a color separation matrix to form dye-specific signals via matrix inversion.

Signal analysis identifies a DNA data peak corresponding to a particular DNA fragment. Using allelic and internal size ladders, the analysis determines a DNA peak's *size* (bp) and allele length (repeats), as well as the DNA *quantity* measured in relative fluorescent units (rfu). A data vector records the DNA quantity (including zero) for every fragment size.

We model the quantitative data at STR locus l (of L loci) using several variables. Data vector $\mathbf{d}_l$ forms a pattern that maps DNA product lengths into their observed quantitative peak heights.

We linearly model the data vector $\mathbf{d}_l$ using a truncated ($\geq \mathbf{0}$) multivariate normal distribution N_+ of the mean vector μ_l and covariance matrix Σ_l as

$$\mathbf{d}_l \sim N_+(\mu_l, \Sigma_l)$$

We write the peak data covariance matrix Σ_l as

$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2$$

where σ^2 is amplification dispersion, τ^2 is detection variation, and V_l is a diagonal matrix $diag(\mathbf{d}_l)$ of peak heights.

Data Usage

TrueAllele determines statistical parameters directly from the data, mining DNA evidence for statistical information. The fully Bayesian system does not require calibration (i.e., setting parameters from historical laboratory data unrelated to evidence data).

TrueAllele inputs and uses all the data. There are no thresholds, since uncertainty is determined statistically from the data. There is an optional rfu cutoff, usually set at ten rfu (within the background noise), well below allelic peak events. Should alleles be observed below this level, the cutoff can be lowered or turned off.

Allele dropout occurs when alleles that are present in the genotypes do not appear in the data signal. Bayesian modeling accounts for all genotype possibilities, whether or not component alleles manifest themselves in the genotyping data. TrueAllele assesses allele dropout through a likelihood function, assigning lower probabilities to genotype proposals that have less support in the data. TrueAllele addresses allele drop-in events in a similar way. There are no explicit drop parameters – Bayes theorem with an informative likelihood function addresses data drop phenomena.

Using all the data is thorough and preserves identification information. Eliminating human data decisions (choosing loci, peaks, artifacts) removes human bias from the

interpretation process. Users cannot "control their data." The user supplies the data, makes few assumptions (number of contributors, sampling time, degradation option), and the rest is done automatically by the computer system.

TrueAllele can answer questions about different data combinations. Data from multiple items or amplifications can be used in a joint genotype analysis. Known genotypes (e.g., a victim present in a mixture, ascertained by case context or match statistics) can help reduce problem complexity.

A comparison genotype (e.g., a suspect) cannot be part of interpreting evidence. The computer does not know the "answer" when it separates genotype from evidence. A match comparison is only made afterwards. Guaranteeing that genotype inference is entirely separate from match statistic calculation helps ensure process objectivity.

Mass

DNA is packaged in the cell nucleus. This cell packaging is opened when DNA is extracted from a biological sample and made available for laboratory analysis. The mass, or number of intact DNA molecules examined in a test tube, is modeled as a normal random variable.

The total DNA quantity at locus l is given by mass parameter m_l. The locus mass m_l prior is a (nonnegative) truncated normal distribution on feasible total peak rfu values.

$$m_l \sim N_+\left(5000,\, 5000^2\right)$$

Genotype

Individuals inherit DNA from two parents. Therefore, at a given genetic locus on an autosomal chromosome, a cell has two alleles (STR length variants), one from each parent. This pair of alleles is called a *genotype*. A genotype is represented as a vector of all possible allele sizes, with each vector entry containing a number of alleles at that particular size.

With K contributors to the data, we represent the k^{th} contributor genotype parameter at locus l as a vector $\mathbf{g}_{k,l}$, where the DNA length entries contain allele counts

that sum to 1. A heterozygote genotype vector $\mathbf{g}_{k,l}$ contains two 1/2 entries, while a homozygote has a single 1 entry; all other vector entries are 0.

The genotype *prior* probability $\Pr\{\mathbf{g}_{k,l} = x\}$ at allele pair $x = [i \;\; j]$ is a product of population allele frequencies $\{f_i\}$.

$$\mathbf{g}_{k,l} \sim \begin{cases} f_i^2, & i = j \\ 2 f_i f_j, & i \neq j \end{cases}$$

TrueAllele's *likelihood* function assesses a genotype candidate value to determine how well it explains the observed data. The likelihood is larger when the quantitative data is better accounted for by a predicted peak height pattern based on the allele pair value. For the i^{th} data observation $d_{l,i}$ at locus *l*, the likelihood function for a genotype $\mathbf{g}_{k,l}$ is the probability $\Pr\{d_{l,i} \mid \mathbf{g}_{k,l} = x, \ldots\}$ of the data conditioned on genotype value *x*, where "..." denotes the other model variable values, given by the data distribution $\mathbf{d}_l \sim N_+(\mu_l, \Sigma_l)$.

Combining the prior genotype probability together with *I* independent genetic data observations, we can compute the *posterior* genotype probability using Bayes theorem as the product of prior probability and joint likelihood functions. The probability mass function (pmf) $q(x)$ of genotype $\mathbf{g}_{k,l}$ is the joint probability distribution

$$\Pr\{\mathbf{g}_{k,l} = x \mid d_{l,1}, d_{l,2}, \ldots, d_{l,i}, \ldots\} \propto \Pr\{\mathbf{g}_{k,l} = x\} \cdot \prod_{i=1}^{I} \Pr\{d_{l,i} \mid \mathbf{g}_{k,l} = x, \ldots\}$$

over all the relevant random variables.

Mixture Weight

A mixture contains DNA from two or more people. The relative amount of DNA from a person contained in the mixture is a *mixture weight* value between zero and one. The sum of the mixture weights over all the people contributing the mixture is one.

The mixture weight parameter at locus l is a vector $\mathbf{w}_l$ whose K contributor components sum to 1, so that $\sum_{k=1}^{K} w_{k,l} = 1$. A hierarchical model of mixture weight at every

locus provides a better fit to the data. We therefore draw each individual locus weight $\mathbf{w}_l$ as a hierarchical prior from a common DNA template mixture weight $\mathbf{w}$ using a truncated (simplex) multivariate normal distribution as

$$\mathbf{w}_l \sim N_{[0,1]^{K-1}}\left(\mathbf{w},\, \psi^2 \cdot I\right)$$

The mixture weight covariance is an identity matrix scaled by a mixture variance ψ^2.

The template mixture weight $\mathbf{w}$ is assigned a uniform prior probability over the K contributor simplex.

$$\mathbf{w} \sim Dir(\mathbf{1})$$

The mixture variance ψ^2 has an inverse gamma prior probability distribution.

$$\psi^{-2} \sim Gam(1/2,\, 1/200)$$

Genotype Combination

Genotypes are combined in a mixture by adding together contributor vectors, with each contributor weighted by its mixture weight. The sum is a genotype vector that describes the total number of alleles in the sample at each fragment size.

A quantitative linear model of data pattern $\mathbf{d}_l$ at locus l has an expected vector value μ_l given by the weighted genotype sum

$$\mu_l = m_l \cdot \sum_{k=1}^{K} w_{k,l} \cdot \mathbf{g}_{k,l}$$

Amplification Variance

The polymerase chain reaction (PCR) is an imperfect copying mechanism. A PCR cycle does not automatically double the number of copies of a particular DNA fragment. Rather, the number of fragment copies randomly increases each round by a factor between one and two. This random branching process follows the mathematics of a Poisson counting process, which can be modeled as a positive-valued distribution having a variance that scales with fragment quantity y as $\sigma^2 \cdot y$.

The data variation parameter σ^2 has an inverse gamma prior probability distribution.

$$\sigma^{-2} \sim Gam(10, 20)$$

Background Variance

Instrument noise arises from a genetic analyzer's laser signal, optical path, CCD camera, and other sources. This background noise is independent of the PCR process, and can be modeled as a normal distribution having a fixed variance parameter.

The data variation parameter τ^2 has an inverse gamma prior probability distribution.

$$\tau^{-2} \sim Gam(10, 500)$$

PCR Stutter

The DNA polymerase enzyme can drop or add a repeated STR unit when replicating an STR fragment. The Markov chain process forms a random pattern of fragment lengths centered about the primary allele length. This PCR *stutter* pattern is far more pronounced with the mono- or di-nucleotide repeat loci used in genetics, and attenuated somewhat with the tetra- or penta-nucleotide repeats used in forensics.

The stutter amount increases with the number of repeats, and can be modeled as a regression line. Let x be the number of repeat units, and y the stutter proportion. Then the linear model relating increasing stutter amount to repeat length at a locus is:

$$y \sim N\left(a + bx, \sigma_S^2\right)$$

Prior probabilities for the PCR stutter model parameters are:

$$a \sim N(0,1)$$
$$b \sim N\left(0,10^{-6}\right)$$
$$\sigma_S^{-2} \sim Gam\left(0.5, 0.5 \cdot 10^{-2}\right)$$

The stutter proportion is constrained to lie between 0% and 15%.

Relative Amplification

PCR amplifies shorter DNA fragments more efficiently than longer ones. This *relative amplification* displaces allele mass away from longer alleles toward short ones.

The allele mass rebalancing increases with the size difference between alleles, and can be modeled as normally distributed variation in allele height. Let Δx be the difference in repeat units, and Δy the difference in allele peak heights. Then the linear model relating allele height difference to size difference at a locus is:

$$\Delta y \sim N\left(c \cdot \Delta x, \sigma_R^2\right)$$

Prior probabilities for the relative amplification model parameters are:

$$c \sim N\left(0, 10^{-4}\right)$$
$$\sigma_R^{-2} \sim Gam\left(0.5, 0.5 \cdot 10^{-6}\right)$$

Differential Degradation

Polymerase requires a connected DNA fragment in order to make a copy. One or more breaks in a DNA sequence will prevent PCR copying. The chance of having no breaks in a fragment (unimpeded copying) follows an exponential decay curve in the fragment length variable, with a decay rate proportional to the density of DNA breaks.

Since TrueAllele models the DNA mass and variation of each experiment separately, no additional modeling is needed when DNA degradation or inhibition is the same for all contributors. However, when there is differential degradation between the different contributors, the decay rate of each contributor's DNA can be determined by logarithmic modeling of the exponential process.

Let x be allele size, y contributor allele amount, and y_{eff} the effective contributor amount following DNA degradation. Then the linear model relating effective allele amount to allele size for a contributor at a locus is:

$$\log\left(\frac{y_{eff}}{y}\right) \sim N\left(-\lambda \cdot x, \sigma_D^2\right)$$

Prior probabilities for the differential degradation model parameters are:

$$\lambda \sim N_+\left(0, 10^{-6}\right)$$
$$\sigma_D^{-2} \sim Gam\left(0.5, 0.5 \cdot 10^{-2}\right)$$

Hierarchical Modeling

TrueAllele models variables hierarchically, subdividing them by experiment. Thus one parameter can expand into many parameters, one for each STR locus experiment, and another one for the group. This expansion of variables permits modeling that is more customized to the data, yielding more accurate answers.

For example, contributor mixture weights are determined for each locus experiment as the set of variables $\{\mathbf{w}_l\}$, and also for the DNA template as group variable $\mathbf{w}$.

$$\mathbf{d}_l \sim N_+\left(m_l \cdot \sum_{k=1}^{K} w_{k,l} \cdot \mathbf{g}_{k,l}, \Sigma_l\right)$$

$$\mathbf{w}_l \sim N_{[0,1]^{K-1}}\left(\mathbf{w}, \psi^2 \cdot I\right)$$

$$\mathbf{w} \sim Dir(\mathbf{1})$$

Statistical Computing

The joint probability distribution is fully specified as the product of the likelihood and prior distributions. Using a Metropolis-Hastings sampler, we iteratively draw from the posterior probability distributions of $\{\mathbf{g}_{k,l}\}$, $\{\mathbf{w}_l\}$, $\{m_l\}$, $\mathbf{w}$, σ^2, τ^2, ψ^2 and other variables using Markov chain Monte Carlo (MCMC) computer methods.

Once beyond the initial burn in phase, the Markov chain samples from the joint posterior probability distribution. Marginalizing these posterior samples to each genotype

random variable $\mathbf{g}_{k,l}$ for contributor k at locus l, we obtain the desired posterior probability functions $q(x)$ for the genotypes.

Match Statistic

The likelihood ratio (*LR*) is the information gained in the hypothesis *H* odds by having observed data

$$LR = \frac{O\left(H \middle| d_Q, d_R, d_S\right)}{O(H)}$$

Here, hypothesis *H* is that the suspect contributed to the DNA evidence, and the DNA data comprises the questioned evidence d_Q, the reference population allele frequencies d_R and suspect profile d_S.

Standard Bayesian rearrangements tell us that the *LR* can also be written as the ratio of conditional probabilities

$$LR = \frac{\Pr\left\{d_Q \middle| H, d_R, d_S\right\}}{\Pr\left\{d_Q \middle| \overline{H}, d_R, d_S\right\}}$$

where $\overline{H}$ is the alternative hypothesis that someone else contributed to the evidence.

Suppose that there is uncertainty in the evidence genotype having pmf $q(x)$ or in the suspect genotype with pmf $s(x)$. Then this genotype uncertainty is expressed in the *LR* as

$$LR = \frac{\sum_{x \in G} \lambda_Q(x) \cdot s(x)}{\sum_{x \in G} \lambda_Q(x) \cdot r(x)}$$

where $\lambda_Q(x)$ is the likelihood function of the evidence genotype *Q* and $r(x)$ is the pmf of reference population genotype.

Bayes theorem lets us rewrite this ratio of likelihood sums as a numerically equivalent sum of posterior genotype probability product ratios. Probability can be more intuitive and easier to explain than likelihood.

$$LR = \sum_{x \in G} \frac{q(x) \cdot s(x)}{r(x)}$$

This genotype probability formulation expresses the LR as a sum of ratios that compare match probability to coincidence.

Co-ancestry Correction

The LR for the hypothesis that a person contributed their DNA to evidence items 1 and 2 is calculated from genotype probability distributions via:

$$LR = \frac{\sum_{x \in G} \lambda_1(x) \cdot \lambda_2(x) \cdot \pi_\theta(x)}{\sum_{x \in G} \sum_{y \in G} \lambda_1(x) \cdot \lambda_2(y) \cdot \pi_\theta(x,y)}$$

The joint prior probability $\pi_\theta(x,y)$ function is just the product of independent population priors $\pi(x)$ and $\pi(y)$ when not accounting for co-ancestry (i.e., $\theta = 0$). However, it is more accurate and conservative to recognize that people in a human population share common ancestors (i.e., $\theta > 0$).

The conditional match formulae for the homozygote and heterozygote cases developed by Balding and Nichols were given in the National Research Council (NRC) II report equations 4.10a and 4.10b, derivable from the probability ratio $\pi_\theta(x,y) / \pi_\theta(x)$. The corresponding joint prior probabilities $\pi_\theta(x,y)$ at a particular value of θ are:

$$\pi_\theta(aa,aa) = \frac{p_a\left[(1-\theta)p_a + \theta\right]\left[(1-\theta)p_a + 2\theta\right]\left[(1-\theta)p_a + 3\theta\right]}{(1+\theta)(1+2\theta)}$$

$$\pi_\theta(ab,ab) = \frac{4p_a\left[(1-\theta)p_a + \theta\right]p_b\left[(1-\theta)p_b + \theta\right](1-\theta)}{(1+\theta)(1+2\theta)}$$

In situations where the genotype allele pair values are not the same, the joint probabilities $\pi_\theta(x,y)$ can be similarly calculated from their Dirichlet distributions, as described in Chapter 4 of Evett and Weir's DNA interpretation textbook.

References

Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int* 64: 125-140.

Ballantyne J, Hanson EK, Perlin MW (2013) DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. *Sci Justice* 53: 103-114

Evett IW, Weir BS (1998) Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sunderland, MA: *Sinauer Assoc*.

National Research Council (1996) Evaluation of Forensic DNA Evidence: Update on Evaluating DNA Evidence. Washington, DC: *National Academies Press*.

Perlin MW, Lancia G, Ng S-K (1995) Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. *Am J Hum Genet* 57: 1199-1210.

Perlin MW, Szabady B (2001) Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *J Forensic Sci* 46: 1372-1377.

Perlin MW, Sinelnikov A (2009) An information gap in DNA evidence interpretation. *PLoS ONE* 4: e8327.

Perlin MW (2010) Explaining the likelihood ratio in DNA mixture interpretation. *Promega's Twenty First International Symposium on Human Identification*. San Antonio, TX.

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, et al. (2011) Validating TrueAllele® DNA mixture interpretation. *J Forensic Sci* 56: 1430-1447.