Efficient construction of match strength distributions for uncertain multi-locus genotypes

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

July 20, 2018

Contact Information:

Dr. Mark W. Perlin
Chief Scientific Officer
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
(412) 683-3004
(412) 683-3005 FAX
perlin@cybgen.com

Abstract

Natural variation in biological evidence leads to uncertain genotypes. Forensic comparison of a probabilistic genotype with a person's reference gives a numerical strength of DNA association. The distribution of match strength for all possible references usefully represents a genotype's potential information. But testing more genetic loci exponentially increases the number of multi-locus possibilities, making direct computation infeasible.

At each locus, Bayesian probability can quickly assemble a match strength random variable. Multi-locus match strength is the sum of these independent variables. A multi-locus genotype's match strength distribution is efficiently constructed by convolving together the separate locus distributions. This convolution construction can accurately collate all trillion trillion reference outcomes in a fraction of a second.

This paper shows how to rapidly construct multi-locus match strength distributions by convolution. Function convergence demonstrates that distribution accuracy increases with numerical resolution. Convolution construction has quadratic computational complexity, relative to the exponential number of reference genotypes. A suitably defined random variable reduces high-dimensional computational cost to fast real-line arithmetic.

Match strength distributions are used in forensic validation studies. They provide error rates for match results. The convolution construction applies to discrete or continuous variables in the forensic, natural and social sciences. Computer-derived match strength distributions elicit the information inherent in DNA evidence, often overlooked by human analysis.

Table of Contents

Introduction

Forensic identification is the science of match. When two objects have identical features, their match statistic increases with the rarity of those features. Feature uncertainty or dissimilarity reduces the match statistic. This balance between numerator similarity and denominator surprise is the likelihood ratio (*LR*), first used forensically for glass evidence [1]. Match strength puts the *LR* on a logarithmic scale, enabling the addition of independent evidence factors [2].

Deoxyribonucleic acid (DNA) testing of biological evidence produces phenotypic data from multiple genetic loci. A definite genotype can be inferred from simple locus data. More often, complex DNA evidence produces an uncertain locus genotype that assigns probability to a hundred possible values. Statistical comparison with a definite reference genotype gives an *LR* that numerically divides evidence genotype probability by population probability, both evaluated at the reference. Adding together independent locus log(*LR*) values yields the total match strength.

An evidence genotype's match strength is mathematically determined at every reference point before any comparison is made. The distribution of match strength values gives insight into genotype uncertainty. A definite genotype concentrates all its probability at maximal strength for the one matching reference. An entirely uninformative genotype collapses to zero match strength. Most genotypes fall in between these two extremes, often showing a bell-shaped distribution of match strength along the real line. For references unlikely to have contributed to the evidence, the uncertain genotype's match strength distribution is centered left of zero (Figure 1). For references likely to have contributed to evidence, the contributor distribution is mainly positive (Figure 2).

Match strength distributions have broad application in forensic science. Non-contributor distributions have been graphed as Tippett plots [3] to assess data quality and compare interpretation methods [4]. Distribution curves predict DNA database search specificity [5] and kinship identification power [6]. The distributions provide *LR* error bounds and tail probabilities [7]. In validation studies, match strength distributions summarize the sensitivity and specificity of statistical methods for interpreting DNA mixtures [8, 9].

There are exponentially many multi-locus genotypes. Listing all combinations, with one value from each locus, forms the multi-locus possibilities. A dozen loci generate a trillion trillion possible genotype outcomes. Brute force *LR* comparison of an uncertain genotype with all these reference possibilities is not feasible. Instead, Monte Carlo simulation samples representative match strengths [4, 10]. Branch-and-bound [11] and importance sampling [5] algorithms can improve simulation performance in some applications. But the genotype space grows exponentially with additional locus tests, and sampling is inexact.

There is an analogous combinatorial explosion in probability theory. When tossing a coin n times, there are 2^n possible outcomes of head (H) and tail (T) sequences. One additional toss doubles the number of H-T sequences. But the interesting information concerns the *number* of heads, not where in the sequence these heads occur. With n tosses, there are just n+1 counting results: 0, 1, 2, …, or n heads. A random variable (RV) [12] summarizes the exponential 2^n number of experiment outcomes as a linear n+1 number of informative results. The binomial probability $Binom(k; n, ½)$ of getting k heads in n tosses of a fair coin is $\frac{n!}{k!(n-k)!}2^{-n}$.

The binomial distribution is constructed as a sum of independent coin tosses [13]. Convolving the distribution of n tosses with another toss forms the expanded binomial distribution $Binom(k; n+1, ½)$ for n+1 tosses [14]. Convolution shifts and adds lists of numbers,

producing (for example) the binomial coefficients $\frac{n!}{k!(n-k)!}$ of Pascal's triangle [15]. In this paper, the concepts of RV and convolution are used to efficiently construct match strength distributions for multi-locus evidence genotypes.

The Methods introduce evidence genotypes and their uncertainty. The match strength RV arises from genotype probability functions, and is efficiently constructed by convolution. The function convergence of binned distributions helps demonstrate their accuracy. Mixing genotype distributions to form composites accelerates validation studies.

The Results lend empirical support using an uncertain genotype derived from a 10% DNA mixture component (Materials). The genotype's match strength distribution is constructed for one locus, and then convolved across many loci. Distribution accuracy is assessed by function convergence at increasingly fine bin resolutions. Efficiency is measured by timing different stages of distribution construction. Genotype sample space size is compared with the number of bin intervals. Bin event occupancy explains why the convolution construction works efficiently. Composite distributions can speed up validation. Match strength error rates are instantly calculated from (single or composite) genotype distributions.

The Conclusions discuss the general applicability of these match strength convolution methods for handling genotype uncertainty.

Methods

Multi-locus genotype

DNA is the linear information molecule that encodes cellular function in a four-letter nucleic acid alphabet [16]. The three billion-letter genome sequence differs between people, with greater genetic similarity in more closely related individuals. Two complete genome copies, maternal and paternal, reside in the nucleus of most human cells. When people deposit their biological material, they can be identified through their DNA.

A short tandem repeat (STR) locus is a highly polymorphic marker that accentuates DNA differences at a particular chromosomal location [17]. STR alleles vary by length, based on the number of tandemly repeated short DNA words. A typical STR locus used in human identification has about 15 different length variants. A *genotype* at a locus is a pair of (maternal and paternal) alleles. With $n = 15$ alleles, there are about $n(n+1)/2 \approx 100$ unordered locus allele pairs. These 10^2 allele pairs form the possible locus genotype values.

Forensic scientists sample from $L = 10$ to 25 autosomal STR loci from genetically independent locations across 22 chromosomes [18]. There are roughly $(10^2)^L = 10^{2L}$ possible multi-locus genotype values. Even a dozen ($L = 12$) loci provide a trillion trillion ($10^{2 \cdot 12} = 10^{24}$) possible genotypes, far more genetic bar codes than the seven billion ($< 10^{10}$) people on earth, and thus useful for forensic identification. The number of genotype values 10^{2L} grows exponentially with the number of tested loci L.

The human population is a small (10^{10}) sampling of multi-locus genotype values from the full (10^{24}) set of genotype possibilities. These genotype values follow a non-uniform *population probability* distribution based on locus allele frequencies [19]. This population distribution corresponds to the *prior probability* of a genotype, before observing phenotypic STR data.

Genotype uncertainty

Multiplex STR data can be generated for L loci in a single tube from one biological specimen. A molecular biology laboratory extracts DNA molecules from the specimen, amplifies the STR alleles using polymerase chain reaction, and detects the relative amount of fluorescently-labeled alleles by DNA size separation [20]. With abundant intact DNA from one person (e.g., a reference sample), the observed allele events directly correspond to the person's genotype.

With most DNA evidence, however, the STR data can support multiple genotype explanations. Having different explanations leads to genotype uncertainty. The uncertainty can arise from mixtures of two or more contributors to a DNA specimen [21], damaged or small amounts of DNA [22], or reconstructed genotypes from relatives [23].

Bayesian probability [24] can model STR mixture data as a weighted linear combination of contributor genotypes [25]. A robust probability model [26] accounts for variance and nuisance parameters from the laboratory experiment (e.g., stutter, imbalance, decay). Markov chain Monte Carlo (MCMC) numerically solves high dimensional probability models through statistical sampling [27]. The result is genotype separation [28], producing a *posterior genotype probability* distribution for each person who contributed DNA to the biological specimen.

Genotype probability functions

Genotype uncertainty can be expressed in the standard mathematical language of probability, RV's and their distributions [10, 29]. Let ω_l be a genotype allele pair value for one person at one

locus. Then $\omega = (\omega_1, ..., \omega_L)$ is a person's multi-locus genotype value comprised of allele pairs at all L loci.

Sample space Ω is the set of all genotype outcomes ω for one contributor to DNA evidence. There are natural probability measures on Ω. Prior probability $p(\omega)$ is the chance of observing genotype ω before examining evidence, based on population probability. Function p maps Ω into the unit interval $\mathbb{I} = [0,1]$, a subset of the real numbers $\mathbb{R}$.

STR data introduces a likelihood function λ from Ω into $\mathbb{R}$, where $\lambda(\omega)$ is the conditional probability of observing the data, given a genotype $\omega \in \Omega$. Posterior probability $q(\omega)$ is the chance that a contributor has genotype ω, after observing the STR data. Function q maps genotypes Ω into interval $\mathbb{I}$. This probability mass function (pmf) is calculated from Bayes theorem as $q(\omega) \propto \lambda(\omega) \cdot p(\omega)$, the normalized product of likelihood and prior.

The *LR* of genotype $\omega \in \Omega$ is the posterior-to-prior probability ratio $q(\omega)/p(\omega)$. Bayes theorem can re-express $q(\omega)/p(\omega)$ as a ratio of two likelihoods – the chance $\lambda(\omega)$ of observing the data assuming genotype ω, versus the total data probability $\sum_{\omega \in \Omega} \lambda(\omega) \cdot p(\omega)$ when the genotype is unknown [30].

The logarithm of the *LR*, or the "weight of evidence" [2], measures the strength of association between the evidence genotype and reference ω, relative to coincidence. This *match strength* is a real-valued function s from genotypes Ω into numbers $\mathbb{R}$, where $s(\omega)$ is $\log_{10}[q(\omega)/p(\omega)]$. The number resides on an additive scale in base ten "ban" units.

Match strength distributions

The power set $\mathcal{F}$ contains all subsets of the finite genotype set Ω. $\mathcal{F}$ is a sigma field [12], closed under set union, intersection and complement. When assigning prior measure p, the triple (Ω, $\mathcal{F}$, p) [29] forms a *prior probability space* for Ω. The *non-contributor RV* $\mathbf{X}$ is a function from Ω to $\mathbb{R}$, where genotype ω is mapped into match strength $s(\omega)$.

The *non-contributor distribution function* $F_{\mathbf{X}}$ is the set probability $\Pr\{\mathbf{X} < x\}$ [12]. This cumulative distribution function (cdf) gives the prior probability p of the genotype subset $\{\omega \in \Omega \mid \mathbf{X}(\omega) < x\}$ having match strength $s(\omega)$ less than x ban. A typical non-contributor $F_{\mathbf{X}}$ cdf is shown in Figure 1.

A *partition* of the real interval $[a, b]$ is a finite sequence of real numbers $a = x_0 < x_1 < \ldots < x_N = b$ [31]. For any bin resolution $\varepsilon > 0$, the *bin set* B^{ε} is a set of subintervals $[x_n, x_n + \varepsilon)$ of equal length ε covering $[a, b]$, where endpoints $x_n = a + n \cdot \varepsilon$ form a regular partition. For convenience, let a, b and $1/\varepsilon$ be integers. The bin function $\beta_l: \mathbb{R} \rightarrow B^{\varepsilon}$ maps a real number x into a subinterval $\beta_l(x)$ denoted by the bin's left endpoint x_n. Alternatively, β_c can form ε-sized subintervals $[x_n - \varepsilon/2, x_n + \varepsilon/2)$ centered at x_n midpoints.

The probability mass function (pmf) $f_{\mathbf{X}}^{\varepsilon}$ is a discrete density function with bin resolution ε. Evaluated at bin x_n, $f_{\mathbf{X}}^{\varepsilon}\left(x_n\right)$ has value $F_{\mathbf{X}}(x_{n+1}) - F_{\mathbf{X}}(x_n)$, the chance $\Pr\{x_n \leq \mathbf{X} < x_{n+1}\}$ that match strength $\mathbf{X}$ falls in bin x_n. A typical non-contributor pmf $f_{\mathbf{X}}^{\varepsilon}$ is shown in Figure 1.

The *posterior probability space* (Ω, $\mathcal{F}$, q) assigns posterior measure q to genotype set Ω. The *contributor RV* $\mathbf{Y}$ maps this genotype probability space Ω into $\mathbb{R}$ via the match strength

function s. The cdf $F_{\mathbf{Y}}(x) = \Pr\{\mathbf{Y} < x\}$ is the *contributor distribution function*. The discrete pmf $f_{\mathbf{Y}}^{\varepsilon}(x_n)$ maps a match strength bin of B^{ε} into its genotype set probability $\Pr\{x_n \le \mathbf{Y} < x_{n+1}\}$. Figure 2 shows a typical contributor cmf $\mathrm{F}_{\mathbf{Y}}$ and its pmf $f_{\mathbf{Y}}^{\varepsilon}$.

Convolution construction

At one locus l, constructing the locus pmf $f_{\mathbf{X}_l}^{\varepsilon}$ for the non-contributor RV $\mathbf{X}_l$ is straightforward. For each genotype ω_l in the small finite set Ω_l, match strength $s(\omega_l)$ is calculated. Function s is defined for ω_l whenever $p(\omega_l) > 0$ and $q(\omega_l) > 0$. The $s(\omega_l)$ number resides in bin $x_n = \beta(s(\omega_l))$, for some integer n. Genotype ω_l's prior probability amount $p(\omega_l)$ is added to bin x_n. Binning the $(\beta(s(\omega_l)), p(\omega_l))$ pairs for all genotypes $\omega_l \in \Omega_l$ forms non-contributor locus pmf $f_{\mathbf{X}_l}^{\varepsilon}$. Figure 3 shows a locus pmf construction for the genotype rows of Table 1.

The total match strength $\mathbf{X}$ is the sum $\sum_{l=1}^{L} \mathbf{X}_l$ of the independent locus match strengths $\mathbf{X}_l$, since independent factors multiply, and logarithms add the factors. From elementary probability theory [13], the pmf $f_{\mathbf{X}}$ of a sum of L independent of RVs is the L-fold convolution $f_{\mathbf{X}_1} * f_{\mathbf{X}_2} * \ldots * f_{\mathbf{X}_L}$ of their individual pmfs $f_{\mathbf{X}_l}$. *Convolution* is a fast way of smoothing one function f with another function g to form a new function $h(x) = \sum_{y} f(y) g(x-y)$ [32], as shown in Figure 4.

Sequential convolution constructs pmf $f_{\mathbf{X}}$ by adding one locus at a time. The first locus has pmf $f_{\mathbf{X}_1}$. After constructing *K* loci, $f_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ is extended by convolution with locus pmf $f_{\mathbf{X}_{K+1}}$ to form the multi-locus $f_{\mathbf{X}_1+\ldots+\mathbf{X}_{K+1}}$. That is,

$$f_{\mathbf{X}_1+\ldots+\mathbf{X}_{K+1}} = f_{\mathbf{X}_1} * f_{\mathbf{X}_2} * \ldots * f_{\mathbf{X}_{K+1}} = \left(f_{\mathbf{X}_1} * f_{\mathbf{X}_2} * \ldots * f_{\mathbf{X}_K} \right) * f_{\mathbf{X}_{K+1}} = f_{\mathbf{X}_1+\ldots+\mathbf{X}_K} * f_{\mathbf{X}_{K+1}}$$

Convolving all *L* loci constructs $f_{\mathbf{X}}$. The cumulative sum of $f_{\mathbf{X}}$ is cdf $F_{\mathbf{X}}$.

Distribution convergence

Distribution function $F_{\mathbf{X}}^{\varepsilon}$ becomes more exact with smaller ε (Figure 5). Genotype set Ω is finite, so there is a smallest match strength distance $d = \min_{\omega,\omega' \in \Omega} \left| s(\omega') - s(\omega) \right|$ between genotypes. At resolution $\varepsilon_0 = d/2$ ban, $F_{\mathbf{X}}^{\varepsilon_0}$ has at most one genotype event ω in each bin interval. Thus the binned $F_{\mathbf{X}}^{\varepsilon_0}$ and exact $F_{\mathbf{X}}$ distributions both fully resolve the events, assuring eventual convergence of $F_{\mathbf{X}}^{\varepsilon}$ to the limit $F_{\mathbf{X}}$.

The largest vertical difference $\max_{x \in B^{\varepsilon} \cup B^{\varepsilon'}} \left| F_{\mathbf{X}}^{\varepsilon'}(x) - F_{\mathbf{X}}^{\varepsilon}(x) \right|$ between two functions $F_{\mathbf{X}}^{\varepsilon}$ and $F_{\mathbf{X}}^{\varepsilon'}$ is the standard L^{∞} supremum norm $\left\| F_{\mathbf{X}}^{\varepsilon'} - F_{\mathbf{X}}^{\varepsilon} \right\|_{\infty}$ [31]. As ε decreases, function distance $\left\| F_{\mathbf{X}}^{\varepsilon'} - F_{\mathbf{X}}^{\varepsilon} \right\|$ measures the Cauchy convergence [31] of $F_{\mathbf{X}}^{\varepsilon}$ to $F_{\mathbf{X}}$.

Composite distribution

Combining a set of genotype probability distributions produces a new aggregate distribution. A *composite mixture distribution* F_N averages together N individual evidence distributions [33]. Suppose F_n is the distribution function of the n^{th} individual genotype RV $\mathbf{X}_n$. Equally weighting individual genotype components, the composite RV $\mathbf{X}_N$ has the mixture cdf $F_N = \frac{1}{N}\sum_{n=1}^{N} F_n$. A composite mixture pmf f_N is similarly formed as $\frac{1}{N}\sum_{n=1}^{N} f_n$ from individual genotype pmfs f_n.

Materials

Statistical software

The fully Bayesian TrueAllele® Casework system (Cybergenetics, Pittsburgh, PA) separates STR mixture data to produce a genotype for each DNA contributor [25]. Genotype uncertainty is represented as prior and posterior probability. The computer constructs non-contributor **X** and contributor **Y** match strength distributions by convolution, draws their pmf curves, and calculates tail probabilities. It summarizes genotype information, providing a Kullback-Leibler (KL) statistic $\mathbb{E}[\mathbf{Y}]$ that predicts LR values [34].

TrueAllele can compare a separated evidence genotype with a reference genotype, relative to a population, to calculate *LR* match strength. The match module accounts for population co-ancestry via its coefficient θ. The match calculation can substitute one population prior for another by Bayesian rearrangement. Locus log(*LR*) values are bounded below by –2 ban, based on validation studies [35].

STR mixture data

Two-person STR mixture data was available from a previous study [21]. The samples were amplified using PowerPlex16® (Promega, Madison, WI), a multiplex kit containing 15 independent STR locus tests. Readout from an ABI310® (Applied Biosystems, Foster City, CA) capillary sequencer produced .fsa electronic data files. The population frequencies used were from the FBI's expanded Caucasian allele database [36].

This study used data from the ten 250 pg samples. The mixture ratios were 1:9 (B3, F3, I3, M3), 3:7 (C3, E3, J3, L3), and 5:5 (D3, K3). The results here focus on the minor M3 12.67% component. It contained 30 pg of DNA (12% of 250 pg), amounting to 5 cells (6 pg DNA per cell). The non-overlapping minor data peaks heights were all under 50 relative fluorescent units (RFU). The minor genotype had a *KL* of 7.8364 ban. Comparison with the known reference gave log(*LR*) values of 5.7291 ($\theta = 0$) and 5.4989 ($\theta = 0.01$) ban.

Results

Locus binning construction

Figure 3 steps through the single locus construction of non-contributor pmf $f^{\varepsilon}_{\mathbf{X}_l}$. Table 1 lists 13 ω_l allele pairs at locus CSF1PO for mixture sample M3's minor genotype. Each row shows the genotype variable's prior and posterior probabilities, and the posterior-to-prior *LR*, with its base ten match strength logarithm. The interval partition uses centered bins, rounding log(*q*/*p*) match strengths to the nearest x_n point at resolution $\varepsilon = 1/4$.

The first table row represents genotypes having zero posterior probability, putting a total 0.1561 prior probability dose into bin –2 (Figure 3, blue bar 1). The second row for allele pair 8,12 adds more probability $p(8,12) = 0.0209$ to the same bin –2 (green bar 2). For the third genotype 9,12 (blue bar 3), the *LR* is 0.0239, which has log strength –1.6215, corresponding to the centered bin –1.5 representing subinterval [–1.5–ε, –1.5+ε). This genotype deposits a prior probability of $p(9,12) = 0.0188$ into bin –1.5 (blue bar 3). Genotype 12,12's log(LR) is match strength $s(12,12) = -1.5042$, and so its prior probability $p(12,12) = 0.0894$ is added to bin –1.5 (green bar 4).

Genotype binning of prior probability into match strength bins continues until all 13 values have been added to form pmf $f^{\varepsilon}_{\mathbf{X}_l}$ (all bars).

Multi-locus convolution

Figure 6 shows the sequential convolution of individual locus pmfs $f_{\mathbf{X}_l}$ to form the multi-locus pmf $f_{\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3+\mathbf{X}_4}$. The first row is for locus D2S1338 of M3's minor genotype. Locus pmf $f_{\mathbf{X}_1}$ is shown on a focused [–2, 2] ban locus-level scale (left), and also on a broader [–10, 5] ban multi-locus scale (right). The bin resolution is 1/4 ban.

The second row adds a second locus TPOX. On the left is locus pmf $f_{\mathbf{X}_2}$. The right plot convolves $f_{\mathbf{X}_1}$ (above) with $f_{\mathbf{X}_2}$ (left) to form (light green arrows) the multi-locus pmf $f_{\mathbf{X}_1+\mathbf{X}_2}$ (right). The two locus combination shows more locus pair genotype events (as bars) for $f_{\mathbf{X}_1+\mathbf{X}_2}$ than for either of the single genotype locus pmfs $f_{\mathbf{X}_1}$ or $f_{\mathbf{X}_2}$.

The third row shows pmf $f_{\mathbf{X}_3}$ for locus D3S1358 on the left. Combining with $f_{\mathbf{X}_1+\mathbf{X}_2}$ (above right) forms (green arrows) the convolution $f_{\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3}$ (right). The triple locus combined pmf is jagged, but now developing shape.

The fourth row combines the FGA locus pmf $f_{\mathbf{X}_4}$ (left) with $f_{\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3}$ (above right) to form (dark green arrows) the quadruple convolution $f_{\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3+\mathbf{X}_4}$ (right). Convolving more loci has made this pmf smoother than its multi-locus precursors (right column). Each non-contributor locus $\mathbf{X}_K$ adds exclusionary power, pushing $f_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ further to the left (right column).

Adding more loci to match strength $\sum_{l=1}^{K}\mathbf{X}_l$ continues these trends (Figure 7). With five loci, $f_{\mathbf{X}_1+\ldots+\mathbf{X}_5}$ has a unimodal shape (green). At ten loci, a smooth bell-shaped curve emerges for $f_{\mathbf{X}_1+\ldots+\mathbf{X}_{10}}$, further shifted to the left (blue). Combining all fifteen loci, $f_{\mathbf{X}_1+\ldots+\mathbf{X}_{15}}$ shows the distribution of match strength for non-contributor multi-locus genotypes (black). Increasing convolution with more loci smooths the curve, pushing the pmf leftward toward greater exclusionary power.

Cumulative distribution convergence

Locus cdf $F^{\varepsilon}_{\mathbf{X}_l}$ is the cumulative sum of locus pmf $f^{\varepsilon}_{\mathbf{X}_l}$. As ε decreases, cdf $F^{\varepsilon}_{\mathbf{X}_l}$ converges to $F_{\mathbf{X}_l}$. Figure 5 shows this convergence for the minor M3 genotype at locus CSF1PO.

Setting $\varepsilon_k = 2^{-k}$ ban, increasingly fine resolutions discretize cdf $F_{\mathbf{X}_l}$ for $k = 0, 1, 2, 3$. Moving from $\varepsilon_0 = 1$ to $\varepsilon_1 = 1/2$ ban refines the partitioning of function $F^{\varepsilon}_{\mathbf{X}_l}$ on the interval [–2, 1] (Figure 5, $k = 0, 1$). Further cdf $F^{\varepsilon_2}_{\mathbf{X}_l}$ step function refinement continues with $\varepsilon_2 = 1/4$ ban, corresponding to the pmf $f^{\varepsilon_2}_{\mathbf{X}_l}$ histogram binning shown in Figure 3. Resolutions beyond $\varepsilon_3 = 1/8$ ban do not change $F^{\varepsilon}_{\mathbf{X}_l}$, so $F^{\varepsilon_3}_{\mathbf{X}_l}$ has converged to the limit distribution $F_{\mathbf{X}_l}$.

Multi-locus cdf $F^{\varepsilon}_{\mathbf{X}}$ combines the individual $F^{\varepsilon}_{\mathbf{X}_l}$ locus distributions through the match strength sum $\mathbf{X}_1 + \ldots + \mathbf{X}_L$. Figure 8 shows $F^{\varepsilon_k}_{\mathbf{X}}$ for a series of increasingly fine bin resolutions $\varepsilon_k = 10^{-k}$ ban, as k progresses from 0 to 3. At $k = 0$, ε_k is 1 ban, and the step function $F^{\varepsilon_0}_{\mathbf{X}}$ has clear one ban increments. At $k = 1$, ε_k is 1/10 ban, and the steps of $F^{\varepsilon_1}_{\mathbf{X}}$ are still visible. Once $k = 2$, the $\varepsilon_k = 1/100$ bin resolution is no longer visible for $F^{\varepsilon_2}_{\mathbf{X}}$. Beyond that resolution, as shown for $\varepsilon_k = 1/1000$ ban at $k = 3$, $F^{\varepsilon_k}_{\mathbf{X}}$ looks the same as $F^{\varepsilon_2}_{\mathbf{X}}$.

Binned distribution accuracy

The convergence of binned $F^{\varepsilon_k}_{\mathbf{X}}$ functions, as resolution k increases, measures their accuracy. Since there are finitely many genotypes, $F^{\varepsilon_k}_{\mathbf{X}}$ must eventually reach the distribution limit $F_{\mathbf{X}}$. The goal is a bin resolution ε_k that provides sufficient accuracy in reasonable time.

The maximum probability difference $\left\| F_{\mathbf{X}}^{\varepsilon_k} - F_{\mathbf{X}}^{\varepsilon_6} \right\|$ between $F_{\mathbf{X}}^{\varepsilon_k}$ at bin resolution $\varepsilon_k = 10^{-k}$ ban, and $F_{\mathbf{X}}^{\varepsilon_6}$ at $\varepsilon_6 = 10^{-6}$ ban, was measured for the minor M3 genotype. These function distances are listed in Table 2 for non-contributor **X** and contributor **Y** match strengths. As bin resolution k becomes finer, the cdf differences get smaller.

Figure 9 plots the non-contributor cdf differences on a logarithmic scale (blue cross line). The negative linear slope indicates exponential improvement with increasing k. At $\varepsilon_2 = 10^{-2}$ ban, the maximum cdf difference is 8.215×10^{-4}, or under one in a thousand (Table 2). With $k = 3$ and $\varepsilon_3 = 10^{-3}$ ban, the difference is 8.533×10^{-5}, or under one in ten thousand. At either resolution, $k = 2$ or $k = 3$, the $F_{\mathbf{X}}^{\varepsilon_k}$ probability error is negligible.

Computer calculation time is shown in Table 2, for both **X** and **Y**. The non-contributor times for constructing distribution $F_{\mathbf{X}}^{\varepsilon_k}$ are plotted on a logarithmic scale in Figure 9 (red plus line) as k varies. For $k = 0$, 1, 2 and 3, the time is under 1/100 sec (Table 2). That time increases to over 1/10 sec for $k \geq 4$. A practical choice of bin resolution is thus at $k = 3$ for $\varepsilon_3 = 10^{-3}$ ban, used in the remainder of this paper, where the probability function deviation is under 10^{-4} and the computer time is under 10^{-2} sec.

Computational complexity analysis

The divide-and-conquer convolution algorithm for computing $F_{\mathbf{X}}^{\varepsilon}$ has quadratic computational complexity $O(L^2)$ in the number of tested loci L. There are three main algorithmic steps.

(a) Constructing each locus pmf function $f^{\varepsilon}_{\mathbf{X}_l}$ uses a fixed bin resolution ε for a relatively constant number of locus genotypes Ω_l. So each locus function $f^{\varepsilon}_{\mathbf{X}_l}$ incurs a constant $O(1)$ construction cost. Across L loci, the cost adds up to $O(L)$.

(b) The pmf $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ of the first $K<L$ loci is sequentially convolved with the $(K+1)^{st}$ locus pmf $f^{\varepsilon}_{\mathbf{X}_{K+1}}$ to form the pmf $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_{K+1}} = f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K} * f^{\varepsilon}_{\mathbf{X}_{K+1}}$ of the first $K+1$ loci. This pairwise convolution combines $O(K)$ bins from the first K loci, with $O(1)$ bins from the next locus, to augment the number of bins to $O(K+1)$. The $O(K)$ stepwise cost is bounded above by $O(L)$. Iterating over L loci, the cost tallies to $O(L^2)$.

(c) Cumulative summation of pmf $f^{\varepsilon}_{\mathbf{X}}$ to form cdf $F^{\varepsilon}_{\mathbf{X}}$ visits all bins. After convolving L loci, there are $O(L)$ bins. So there is an $O(L)$ summation cost.

Since step (b) dominates the $F^{\varepsilon}_{\mathbf{X}}$ formation cost, the process has quadratic cost $O(L^2)$.

Empirical efficiency measurements

Empirical timings on the minor M3 genotype concur with the algorithmic complexity analysis. The computing times for building the locus pmfs $f^{\varepsilon}_{\mathbf{X}_l}$, and convolving them to form the multi-locus mass functions $f_{\mathbf{X}_1+\ldots+\mathbf{X}_L}$, are listed in Table 3 for both non-contributor **X** and contributor **Y**. The locus breakdown gives the incremental costs of sequentially constructing $f^{\varepsilon}_{\mathbf{X}}$.

Figure 10 plots the computing times for building the locus pmfs (blue cross) and convolving the multi-locus pmfs (red plus). The $f^{\varepsilon}_{\mathbf{X}_l}$ locus build time is relatively constant (blue

cross line & Table 3). The $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ multi-locus convolution time increases linearly with each additional locus K (red plus line & Table 3). A constant locus build time across L loci has $O(L)$ cost, while a linearly increasing multi-locus convolution time for L loci has $O(L^2)$ cost.

Construction method comparison

The quadratic number of convolved bin events is far less than the exponential number of multi-locus genotypes. Table 4 lists the numbers of single genotypes $\#\Omega_K$ and multi-locus genotypes $\#\Omega_1\times\ldots\times\Omega_K$ considered at each locus K for M3's minor genotype. The straight line (red plus) plotted on Figure 11's logarithmic scale shows the *exponential* rise of multi-locus genotype counts. At 15 loci, there are 1.355×10^{23} genotypes. Constructing a match strength distribution directly from the genotype sample space Ω can be exponentially expensive.

The convolution method instead used discrete bins on a bounded interval [–35, 35], with resolution $\varepsilon = 10^{-3}$ ban. Bins are filled by genotype events at each locus (Table 4, Bins column), and expanded with each multi-locus convolution step (X & Y columns). The multi-locus bin growth is *linear* (Figure 11, blue cross), as seen by the logarithmic curve plotted on a logarithmic scale. At 15 loci, 52,406 of the total 70,000 $f^{\varepsilon}_{\mathbf{X}}$ bins were occupied. Clearly, 7×10^4 numeric bins are far fewer than 1.355×10^{23} multi-locus genotypes.

Match strength bin occupancy

Numeric convolution operates on one-dimensional bins, not multi-dimensional genotypes. For the M3 minor genotype mixture data, 10^{23} genotypes were represented in 10^5 bins. This space

efficiency stems from using an RV **X** that reduces the exponential genotype space **Ω** to a bounded interval on the real line $\mathbb{R}$. The convolution operation layers quantitative real-valued locus pmfs atop one another, shifting bins and adding probabilities.

Match strength bins are efficiently reused, as measured by bin occupancy (Table 5). On reaching the STR kit's 15 loci, non-contributor RV **X** had 75% bin occupancy, while contributor **Y**'s bins were 70% occupied.

Re-convolving the loci a second time shows the bin behavior out to 30 loci (Figure 12). The respective **X** (blue cross) and **Y** (cyan plus) bin occupancy rates remain level beyond 15 loci. On average, one $\varepsilon = 10^{-3}$ ban interval for **X** numerically collects 1.936×10^{18} multi-locus genotypes per milliban bin (i.e., 1.355×10^{23} genotypes / 7×10^{4} bins).

Composite genotype distribution

A composite mixture distribution aggregates multiple genotypes into one combined distribution. For example, validation studies can examine a system's specificity as a histogram of evidence genotype match strength [8, 9]. A costly Monte Carlo approach compares each uncertain genotype against thousands of reference genotypes to calculate their match statistics, and then bins the results and collates a composite histogram. Numerical aggregation of match strength distributions can be far more efficient.

TrueAllele separated the 250 pg two person samples (B3–M3) into their component genotypes. The ten minor contributor genotypes were combined into a composite distribution. Each genotype's match strength cdf F_n^{ε} was computed at high resolution ($\varepsilon = 10^{-3}$ ban), and then

averaged into an aggregate distribution F_N^{ε}. Taking one-ban cdf differences $F_N^{\varepsilon}(m+1) - F_N^{\varepsilon}(m)$ constructed a $f_N^{1}(m)$ pmf histogram. Figure 13 shows the resulting non-contributor specificity (red left) and contributor sensitivity (blue right) histograms. The average time to construct all genotype cdfs, form a composite mixture, and construct a histogram was 0.703 sec.

Match strength error frequency

Cumulative distributions $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$, and their associated histograms $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, offer a frequency perspective on match strength. For any evidence genotype, the distributions reveal how frequently a match event would occur at that magnitude, relative to all possible reference genotypes. One application is determining false positive error rate – how often a non-contributor would adventitiously match as strongly as the defendant [37].

Statisticians call false positive *LR* error the *probability of misleading evidence* (*PME*) [7]. Comparing an evidence genotype with a reference ω yielding match strength $x = s(\omega)$ ban, the *PME* is $\Pr\{\mathbf{X} \geq x\}$, or $1 - \Pr\{\mathbf{X} < x\}$, which equals $1 - F_{\mathbf{X}}(x)$. This cdf value is the tail probability of pmf $f_{\mathbf{X}}$ beyond x.

Comparing M3's minor genotype with its known contributor ω gives an *LR* of 536 thousand, for a log(*LR*) of 5.7291. Evaluating non-contributor distribution $F_{\mathbf{X}}$ (Figure 1, Cumulative) at this match strength gives a *PME* of $1 - F_{\mathbf{X}}(5.791)$, which is 1.3662×10^{-7}. Thus, based on the evidence genotype, the chance that a non-contributor matches the DNA evidence as strongly as does the reference is one in 7.32 million.

A validation study's composite F_N can estimate an ensemble *PME* based on a set of representative genotypes. One may ask how often a suspect's 5.7291 ban match would occur in a 250 pg two-person minor mixture. Binning at 1 ban resolution, the composite non-contributor f_N^1 histogram (Figure 13, red left) provides an answer. Calculating either cdf $1-F_N^1(5)$ at bin 5, or the equivalent pmf right tail bin sum $\sum_{m\geq 5} f_N^1(m)$, gives an ensemble *PME* frequency of 1.0367 x 10^{-7}. This is a one in 9.65 million validation probability estimate that the evidence would match a non-contributor as strongly as it does the suspect.

Conclusions

Forensic interpretation is an information science [38]. The computer can organize a sample space of possible outcomes, describe the prior probability of each outcome, and compute the outcome's posterior probability from available evidence data. Constructing probability spaces and random variables from these elements provides a detailed match strength analysis of an evidence item. All reference outcomes are accounted for, so no comparison reference is needed at this pre-match stage. The match strength distributions are useful for quantifying potential identification information, preparing database searches, assessing data or methods, performing validation studies, and calculating LR error.

All forensic information should be extracted from evidence data. In the DNA mixture example, the minor contributor contained five cells, with all non-overlapping peaks having low heights under 50 RFU. Crime labs typically discard such DNA data as uninterpretable, inconclusive, too low, or too complex. They do not use the evidence for database searches or match comparisons. Yet the match strength RV distributions were informative. Eventual

comparison with the true contributor reference gave a match statistic of 536 thousand. That level of DNA association could help convict or acquit a defendant.

The mathematical probability framework led to efficient algorithms for constructing match strength distributions and calculating LR error. The independence of additive locus RVs permitted rapid and accurate joint RV construction by convolution. The sample space contained exponentially many multi-locus genotypes. The match strength RV mapped these multi-dimensional outcomes into uni-dimensional numbers, collecting and preserving match information in quadratic time.

The match strength RV approach is quite general. While genotypes have a discrete representation, other variables (e.g., glass index of refraction) are continuous. The RV distribution approach extends to continuous variables and any dimensionality, with integration replacing summation [12]. LR associations are used in fields beyond forensic science, for example, in artificial intelligence [39], medical diagnosis [40] and legal reasoning [41]. Rapid construction of match strength distributions may offer insights, applications and efficiencies in handling uncertainty in such areas.

Uncertainty is prevalent in the natural and social sciences. Bayesian probability modeling helps extract information from real world data to harness that uncertainty [42]. Advance knowledge of the full range of possible outcomes aids decision-making, whether in forensic biology or diagnostic medicine. This paper showed how probabilistic RV analysis can preserve and use identification information, even when the evidence data are thought to be uninterpretable and no reference is available for comparison.

Acknowledgements

The author would like to thank William Allan, Ria David, John Donahue, Jennifer Hornyak, Matthew Legler, Erin Monko, Eitan Perlin and Mark Wilson for their helpful comments on the manuscript. Margaret Kline and Alexander Sinelnikov prepared the DNA mixture samples for an earlier study.

References

[1] D.V. Lindley, A problem in forensic science, Biometrika, 64 (1977) 207-213.
[2] I.J. Good, Probability and the Weighing of Evidence, Griffin, London, 1950.
[3] C.F. Tippett, V.J. Emerson, M.J. Fereday, F. Lawton, A. Richardson, L.T. Jones, S.M. Lampert, The evidential value of the comparison of paint flakes from sources other than vehicles, Journal of the Forensic Science Society, 8 (1968) 61-65.
[4] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, J. Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Science International: Genetics, 2 (2008) 91-103.
[5] M. Kruijver, Efficient computations with the likelihood ratio distribution, Forensic Science International: Genetics, 14 (2015) 116-124.
[6] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with applications to kinship problems, Int J Legal Med, 128 (2014) 415-425.
[7] R. Royall, On the probability of observing misleading evidence, Journal of the American Statistical Association, 95 (2000) 760-768.
[8] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci Int Genet, 6 (2012) 749-761.
[9] M.W. Perlin, J. Hornyak, G. Sugimoto, K. Miller, TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors, J Forensic Sci, 60 (2015) 857-868.
[10] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with applications to mixtures, Int J Legal Med, 130 (2015) 39-57.
[11] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, T. Egeland, Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Science International: Genetics, 9 (2014) 93-101.
[12] A.N. Kolmogorov, Foundations of the Theory of Probability, Chelsea Publishing Company, New York, 1956.
[13] W. Feller, An Introduction to Probability Theory and Its Applications, Third ed., John Wiley & Sons, New York, 1968.
[14] A. De Moivre, The Doctrine of Chances: A Method of Calculating the Probability of Events in Play, W. Pearson, London, 1718.
[15] B. Pascal, Treatise on the Arithmetical Triangle, Guillaume Desprez, Paris, 1665.
[16] J.D. Watson, F.H. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, Nature, 171 (1953) 737-738.
[17] J. Weber, P. May, Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction, Am. J. Hum. Genet., 44 (1989) 388-396.
[18] M.J. Ludeman, C. Zhong, J.J. Mulero, R.E. Lagace, L.K. Hennessy, M.L. Short, D.Y. Wang, Developmental validation of GlobalFiler PCR amplification kit: a 6-dye multiplex assay designed for amplification of casework samples, Int J Legal Med, (2018).
[19] D.L. Hartl, A.G. Clark, Principles of Population Genetics, Fourth ed., Sinauer Associates, Sunderland, MA, 2006.
[20] C.J. Fregeau, R.M. Fourney, DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification, Biotechniques, 15 (1993) 100-119.

[21] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS ONE, 4 (2009) e8327.
[22] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci Intl, 112 (2000) 17-40.
[23] E. Essen-Möller, The evidential value of similarity as proof of paternity, fundamental principles, Transactions of the Anthropological Society in Vienna, 68 (1938) 9-53.
[24] T. Bayes, R. Price, An essay towards solving a problem in the doctrine of chances, Phil. Trans., 53 (1763) 370-418.
[25] M.W. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, J Forensic Sci, 46 (2001) 1372-1377.
[26] A. Gelman, J.B. Carlin, H.S. Stern, D. Rubin, Bayesian Data Analysis, Chapman & Hall/CRC, Boca Raton, FL, 1995.
[27] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J Chem Phys, 21 (1953) 1087–1092.
[28] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele® DNA mixture interpretation, J Forensic Sci, 56 (2011) 1430-1447.
[29] R.B. Ash, Basic Probability Theory, John Wiley & Sons, New York, 1970.
[30] M.W. Perlin, Explaining the likelihood ratio in DNA mixture interpretation, in: Promega's Twenty First International Symposium on Human Identification, San Antonio, TX, 2010.
[31] W. Rudin, Principles of Mathematical Analysis, Third ed., McGraw Hill, New York, 1976.
[32] A.V. Oppenheim, R.W. Shafer, Digital Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.
[33] W. Feller, On a general class of "contagious" distributions, The Annals of Mathematical Statistics, 14 (1943) 389–399.
[34] S. Kullback, R.A. Leibler, On information and sufficiency, Ann Math Stat, 22 (1951) 79-86.
[35] M.W. Perlin, K. Dormer, J. Hornyak, T. Meyers, W. Lorenz, How inclusion interpretation of DNA mixture evidence reduces identification information (A123), in: AAFS 65th Annual Scientific Meeting, American Academy of Forensic Sciences, Washington, DC, 2013, pp. 93.
[36] T.R. Moretti, L.I. Moreno, J.B. Smerick, M.L. Pignone, R. Hizon, J.S. Buckleton, J.A. Bright, A.J. Onorato, Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States, Forensic Sci Int Genet, 25 (2016) 175-181.
[37] M.W. Perlin, Error in the likelihood ratio: false match probability, in: American Academy of Forensic Sciences 69th Annual Meeting, AAFS, New Orleans, 2017.
[38] M.W. Perlin, Forensic science in the information age, Forensic Magazine, 9 (2012) 17-21.
[39] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Second ed., Morgan Kaufmann, San Francisco, 1988.
[40] J.R. Thornbury, D.G. Fryback, W. Edwards, Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information, Radiology, 114 (1975) 561-565
[41] J.B. Kadane, D.A. Schum, A Probabilistic Analysis of the Sacco and Vanzetti Evidence, John Wiley & Sons, New York, 1996.
[42] J.B. Kadane, Principles of Uncertainty, Chapman & Hall, Boca Raton, FL, 2011.

Figure Captions

Figure 1. *Non-contributor distribution*. (Cumulative) An uncertain genotype's cdf $F_{\mathbf{X}}$ for non-contributor RV **X** shows cumulative probability (y-axis) relative to logarithmic match strength (x-axis). (Probability) At bin resolution $\varepsilon = 10^{-3}$ ban, the corresponding pmf $f_{\mathbf{X}}^{\varepsilon}$ gives the probability in each bin.

Figure 2. *Contributor distribution*. (Cumulative) An uncertain genotype's cdf $F_{\mathbf{Y}}$ for contributor RV **Y** shows cumulative probability (y-axis) relative to logarithmic match strength (x-axis). (Probability) At bin resolution $\varepsilon = 10^{-3}$ ban, the corresponding pmf $f_{\mathbf{Y}}^{\varepsilon}$ gives the probability in each bin.

Figure 3. *Locus construction*. Constructing non-contributor pmf $f_{\mathbf{X}_l}^{\varepsilon}$ for locus CSF1PO at bin resolution $\varepsilon = 1/4$ ban. The l^{th} colored bar represents the $(\beta_c(s(\omega_l)), p(\omega_l))$ strength-probability pair for the corresponding locus genotype ω_l in Table 1. For a genotype match strength $s(\omega_l)$, prior probability amount $p(\omega_l)$ is added to match strength bin $\beta_c(s(\omega_l))$. Accumulating binned probability values over all genotypes builds the locus pmf.

Figure 4. *Function convolution*. Convolving a jagged function *f* (top, dark blue) with a blurring function *g* (middle, green) to form a smooth function *h* (bottom, light blue). Here *f* is a partial pmf convolution $f_{\mathbf{X}_1+\ldots+\mathbf{X}_4}^{\varepsilon}$ of four loci, *g* is a binomial distribution ($n = 10$, $p = 0.5$), and *h* is their convolution. The bin resolution is $\varepsilon = 1/4$ ban.

Figure 5. *Bin resolution.* Locus CSF1PO cdf $F^{\varepsilon}_{\mathbf{X}_l}$ is shown for increasingly fine bin resolution ε values. For illustration, the resolutions are set at $\varepsilon_k = 2^{-k}$ for $k = 0, 1, 2, 3$.

Figure 6. *Sequential convolution*. Sequential convolution builds pmf $f^{\varepsilon}_{\mathbf{X}}$ at four loci. The left column shows individual locus pmf $f^{\varepsilon}_{\mathbf{X}_K}$ bar charts for locus $K = 1, 2, 3, 4$. The right column shows K-fold partially convolved pmf $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ bar charts for locus $K = 1, 2, 3, 4$. The convolution process combines partial convolution $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ (right column, row K), with locus pmf $f^{\varepsilon}_{\mathbf{X}_{K+1}}$ (left, row $K+1$), to extend (green arrows) the multi-locus convolution to $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_{K+1}}$ (right, row $K+1$). The bin resolution is $\varepsilon = 1/4$ ban.

Figure 7. *Further convolution*. Sequential convolution incrementally constructs the multi-locus pmf $f^{\varepsilon}_{\mathbf{X}_1+\ldots+\mathbf{X}_K}$ as K increases from 5 (green) to 10 (blue) to 15 (black) loci. The bin resolution is $\varepsilon = 1/10$ ban.

Figure 8. *Distribution resolution*. Joint cdf $F^{\varepsilon}_{\mathbf{X}}$ is shown at increasingly fine bin resolutions. For illustration, the resolutions $\varepsilon_k = 10^{-k}$ are set at $k = 0, 1, 2$ and 3.

Figure 9. *Accuracy vs. efficiency*. Assessing joint cdf $F^{\varepsilon}_{\mathbf{X}}$ accuracy and efficiency for bin resolutions $\varepsilon_k = 10^{-k}$ where $k = 0, 1, \ldots, 6$. *Accuracy* is logarithmically plotted (blue cross) as the

maximum cdf difference between $F_{\mathbf{X}}^{\varepsilon_k}$ and $F_{\mathbf{X}}^{\varepsilon_6}$ for increasing resolution *k*. *Efficiency* is logarithmically plotted (red plus) as the time (sec) computing $F_{\mathbf{X}}^{\varepsilon_k}$ for increasing *k*.

Figure 10. *Building vs. convolving.* Computing time (sec) for pmf $f_{\mathbf{X}}^{\varepsilon}$ at each locus *l*, as locus number increases. The timings are shown in two parts, *building* (blue cross line) a locus $f_{\mathbf{X}_K}^{\varepsilon}$ pmf, and *convolving* (red plus line) a partial joint $f_{\mathbf{X}_1+\ldots+\mathbf{X}_{K-1}}^{\varepsilon}$ with a new locus $f_{\mathbf{X}_K}^{\varepsilon}$ to form the augmented pmf $f_{\mathbf{X}_1+\ldots+\mathbf{X}_K}^{\varepsilon}$. The bin resolution is $\varepsilon = 10^{-3}$ ban.

Figure 11. *Bins vs. genotypes*. Counting $f_{\mathbf{X}}^{\varepsilon}$ computation size with increasing locus number. Shown is the number (blue cross line) of ε-bins in bin set B^{ε} for the real interval [*a*, *b*] after processing *K* loci. Also shown is the number (red plus line) of genotype *K*-tuples in the locus product $\Omega_1 \times \ldots \times \Omega_K$ after processing *K* loci. The bin resolution is $\varepsilon = 10^{-3}$ ban.

Figure 12. *Bin occupancy.* The percentage of occupied bins when increasing from 1 to 30 loci for non-contributor pmf $f_{\mathbf{X}}^{\varepsilon}$ (blue cross line), and contributor pmf $f_{\mathbf{Y}}^{\varepsilon}$ (cyan plus line). The bin resolution is $\varepsilon = 10^{-3}$ ban.

Figure 13. *Composite frequency*. Validation plots computed as composite f_N^{ε} pmfs for non-contributor **X** specificity (red left histogram), and contributor **Y** sensitivity (blue right histogram) genotype mixture distributions. The histogram bin resolution shown is $\varepsilon = 1$ ban.

Tables

Table 1. Constructing non-contributor pmf $f^{\varepsilon}_{\mathbf{x}_l}$ at locus CSF1PO with bin resolution ε = 1/4 ban.

For each locus genotype ω_l row, the columns show the numbered allele pair, prior $p(\omega_l)$ and posterior $q(\omega_l)$ genotype probabilities, $LR\ q(\omega_l)/p(\omega_l)$, logarithmic match strength $s(\omega_l)$, and the rounded bin's center point $\beta_c(s(\omega_l))$. Rows are sorted by ascending LR.

Genotype	Allele Pair	Prior	Posterior	LR	Strength	Bin
1	many	0.1561	0.0000	0.0000	-2.0000	-2.00
2	8 12	0.0209	0.0002	0.0096	-2.0000	-2.00
3	9 12	0.0188	0.0004	0.0239	-1.6215	-1.50
4	12 12	0.0894	0.0028	0.0313	-1.5042	-1.50
5	11 13	0.0295	0.0018	0.0627	-1.2028	-1.25
6	12 13	0.0332	0.0015	0.0451	-1.3460	-1.25
7	11 11	0.0703	0.0080	0.1138	-0.9439	-1.00
8	7 10	0.0171	0.0029	0.1718	-0.7651	-0.75
9	11 12	0.1586	0.0236	0.1485	-0.8282	-0.75
10	10 13	0.0299	0.0114	0.3805	-0.4197	-0.50
11	10 10	0.0725	0.0946	1.3045	0.1154	0.00
12	10 12	0.1610	0.2743	1.7037	0.2314	0.25
13	10 11	0.1428	0.5785	4.0517	0.6076	0.50

Table 2. Accuracy and efficiency at different bin resolutions $\varepsilon_k = 10^{-k}$ ban. The accuracy of non-contributor cdf $F_{\mathbf{X}}^{\varepsilon_k}$ is measured by its maximum probability difference $\left\| F_{\mathbf{X}}^{\varepsilon_k} - F_{\mathbf{X}}^{\varepsilon_6} \right\|$ from the micro-ban resolution cdf $F_{\mathbf{X}}^{\varepsilon_6}$. Efficiency is measured by the computer time (sec) needed to construct $F_{\mathbf{X}}^{\varepsilon_k}$. Accuracy and efficiency are also shown for contributor distribution $F_{\mathbf{Y}}^{\varepsilon_k}$.

	X		Y	
Resolution	**Difference**	**Time**	**Difference**	**Time**
0	4.0542E-02	2.2753E-03	1.3910E-01	2.2050E-03
1	1.8200E-03	3.4260E-03	1.6871E-02	3.3593E-03
2	8.2153E-04	5.2983E-03	3.5183E-03	5.2863E-03
3	8.5332E-05	9.4160E-03	1.9434E-04	9.0350E-03
4	2.8206E-06	1.1621E-01	1.8599E-05	1.0425E-01
5	1.1386E-06	1.6800E+00	3.2775E-06	1.6754E+00
6		3.4960E+01		3.1200E+01

Table 3. Locus efficiency breakdown. At each incremental locus step, there is a time cost (sec) for building the locus pmf, and a cost for convolving with the preceding convolved loci. These timings are shown for the non-contributor **X** and contributor **Y** distributions.

	X		Y	
Locus	**Build**	**Convolve**	**Build**	**Convolve**
1	2.705E-04	2.229E-04	1.979E-04	1.206E-05
2	3.478E-04	4.609E-04	4.347E-04	1.402E-04
3	2.838E-04	6.287E-04	9.809E-04	2.394E-04
4	2.937E-04	8.052E-04	3.696E-04	3.753E-04
5	1.249E-04	9.678E-04	1.449E-04	4.980E-04
6	1.551E-04	1.294E-03	1.921E-04	6.309E-04
7	2.306E-04	1.604E-03	3.400E-04	8.046E-04
8	2.629E-04	1.939E-03	3.378E-04	1.004E-03
9	1.527E-04	2.225E-03	2.997E-04	1.186E-03
10	2.455E-04	2.564E-03	2.604E-04	1.404E-03
11	2.734E-04	2.996E-03	4.023E-04	1.692E-03
12	4.453E-04	3.596E-03	9.856E-04	2.053E-03
13	3.719E-04	4.118E-03	6.317E-04	2.401E-03
14	1.326E-04	4.544E-03	2.488E-04	2.691E-03
15	4.421E-04	5.128E-03	8.595E-04	3.098E-03

Table 4. Counting genotypes and numeric bins. Shown at each incremental locus step K are the number of locus genotypes $\#\Omega_K$ and number of partial product genotypes $\#\Omega_1 \times \ldots \times \Omega_K$. Also shown is the number of locus bins, and occupied partial multi-locus bins for the non-contributor **X** and contributor **Y** distributions.

Locus	Genotypes	Product	Bins	X	Y
1	11	1.100E+01	8	8	8
2	45	4.950E+02	40	310	303
3	39	1.931E+04	37	5,104	4,985
4	56	1.081E+06	49	10,323	9,851
5	13	1.405E+07	12	13,474	12,864
6	23	3.232E+08	22	17,081	16,105
7	40	1.293E+10	37	20,595	19,516
8	46	5.948E+11	42	24,036	22,231
9	25	1.487E+13	21	27,189	25,279
10	36	5.353E+14	36	30,653	28,099
11	55	2.944E+16	46	36,453	33,304
12	83	2.444E+18	77	41,675	38,446
13	60	1.466E+20	56	45,347	42,030
14	12	1.759E+21	11	48,550	45,233
15	77	1.355E+23	69	52,406	48,941

Table 5. Bin occupancy rate. As loci are incrementally convolved, the total number of numeric bins increase. The number of bins occupied by a genotype event, and the occupancy rate (fraction of total occupied) are shown for non-contributor **X** and contributor **Y** distributions.

		X		Y	
Loci	**Total**	**Occupied**	**Rate**	**Occupied**	**Rate**
1	4,000	8	0.20%	8	0.20%
2	8,000	310	3.88%	303	3.79%
3	14,000	5,104	36.46%	4,985	35.61%
4	18,000	10,323	57.35%	9,851	54.73%
5	22,000	13,474	61.25%	12,864	58.47%
6	26,000	17,081	65.70%	16,105	61.94%
7	30,000	20,595	68.65%	19,516	65.05%
8	34,000	24,036	70.69%	22,231	65.39%
9	38,000	27,189	71.55%	25,279	66.52%
10	42,000	30,653	72.98%	28,099	66.90%
11	50,000	36,453	72.91%	33,304	66.61%
12	58,000	41,675	71.85%	38,446	66.29%
13	62,000	45,347	73.14%	42,030	67.79%
14	66,000	48,550	73.56%	45,233	68.53%
15	70,000	52,406	74.87%	48,941	69.92%

Figure 1

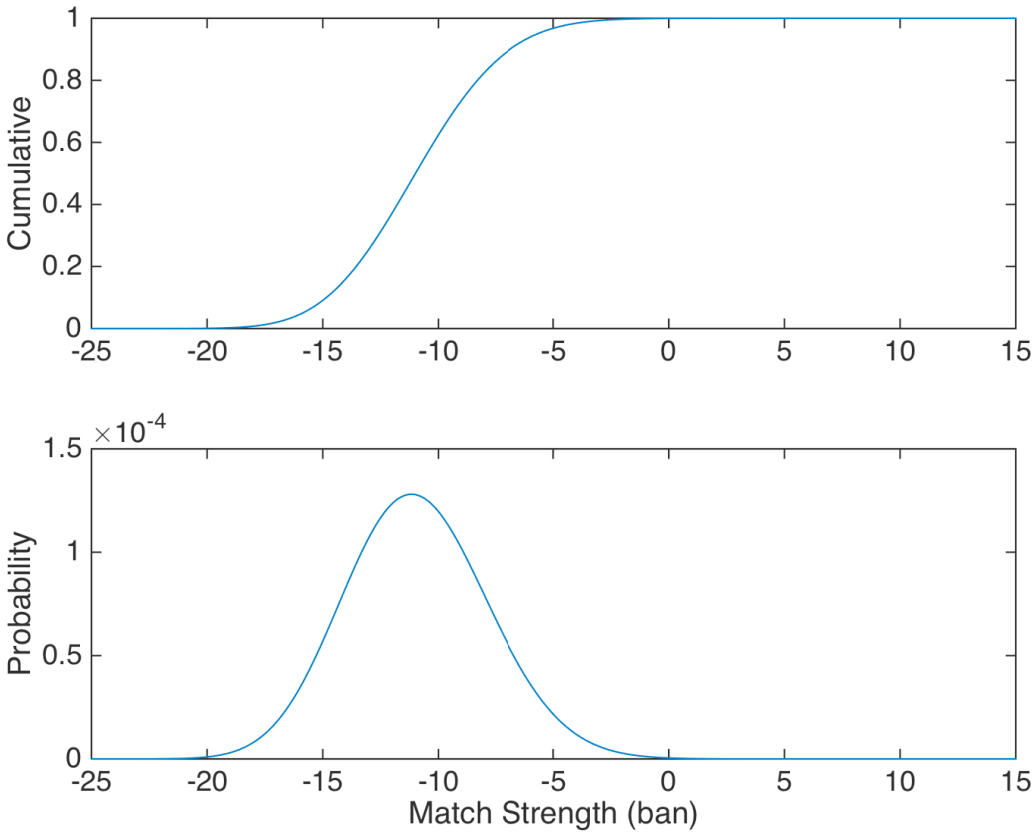
Cumulative
1
0.8
0.6
0.4
0.2
0
-25
-20
-15
-10
-5
0
5
10
15
×10^-4
Probability
1.5
1
0.5
0
Match Strength (ban)

Figure 2

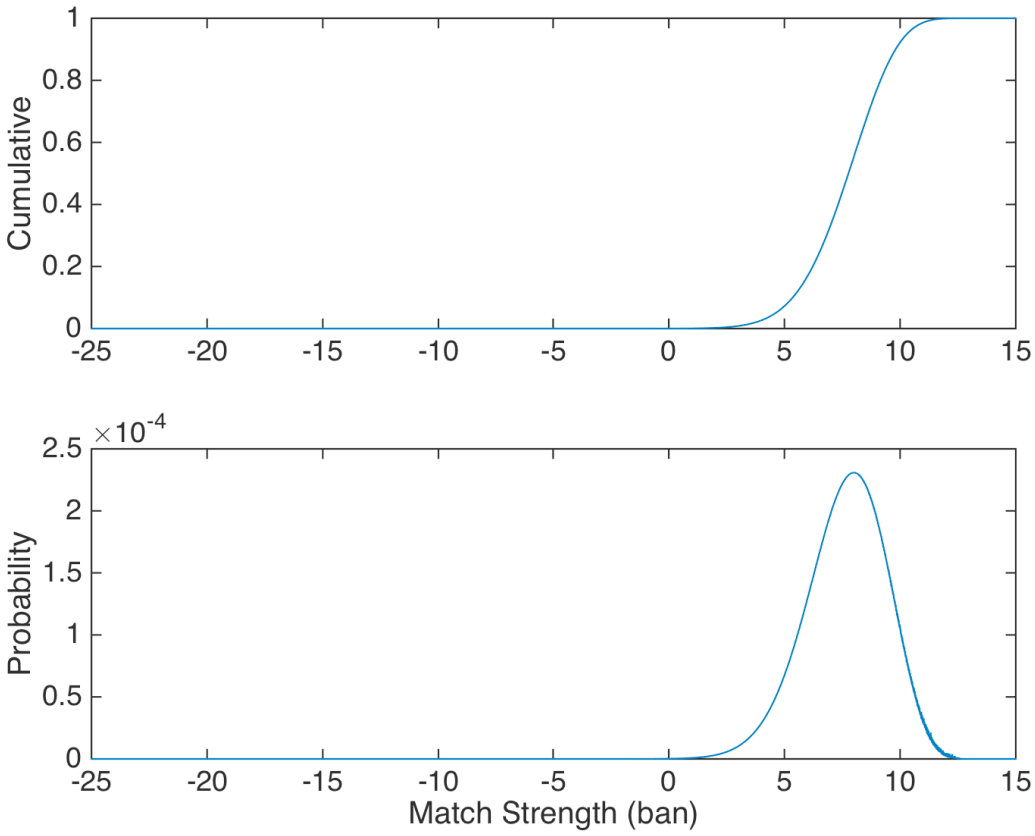
Cumulative
1
0.8
0.6
0.4
0.2
0
-25
-20
-15
-10
-5
0
5
10
15
Probability
×10^-4
2.5
2
1.5
1
0.5
0
Match Strength (ban)

Figure 3

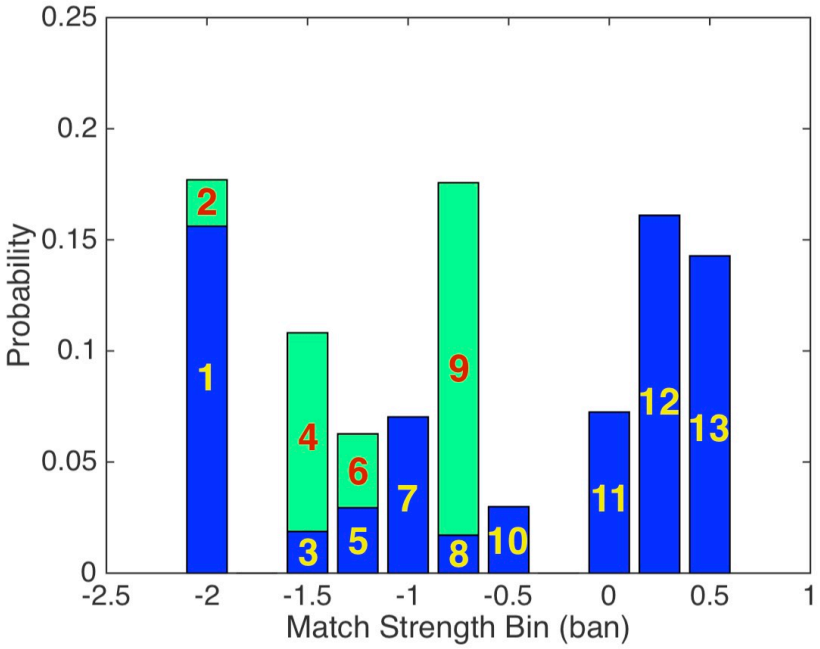
0.25
0.2
0.15
0.1
0.05
0
Probability
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
Match Strength Bin (ban)
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 4

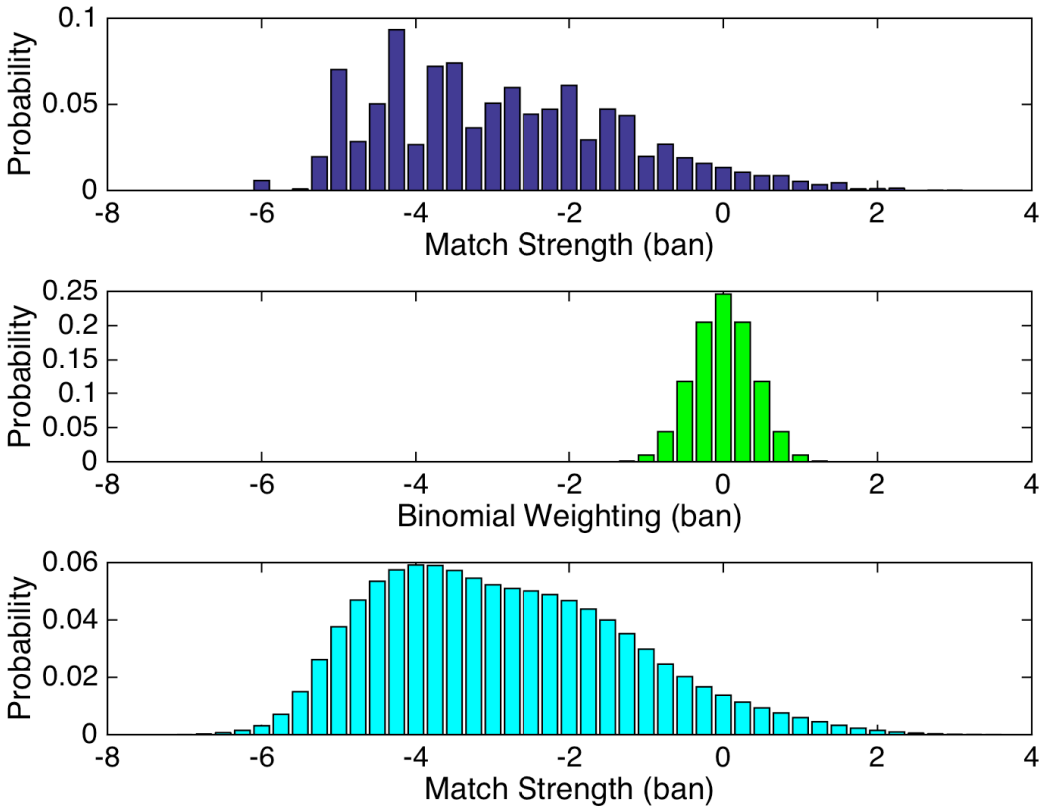
Probability
0.1
0.05
0
-8
-6
-4
-2
0
2
4
Match Strength (ban)
Probability
0.25
0.2
0.15
0.1
0.05
0
Binomial Weighting (ban)
Probability
0.06
0.04
0.02
Match Strength (ban)

Figure 5

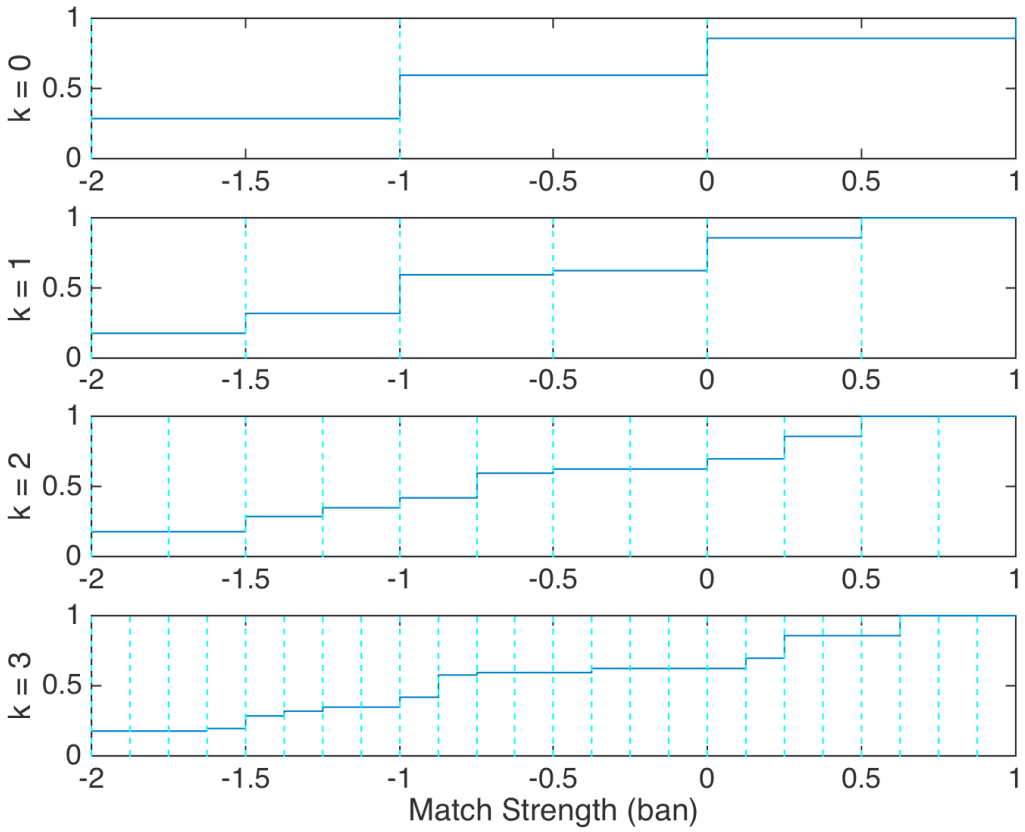

Figure 6

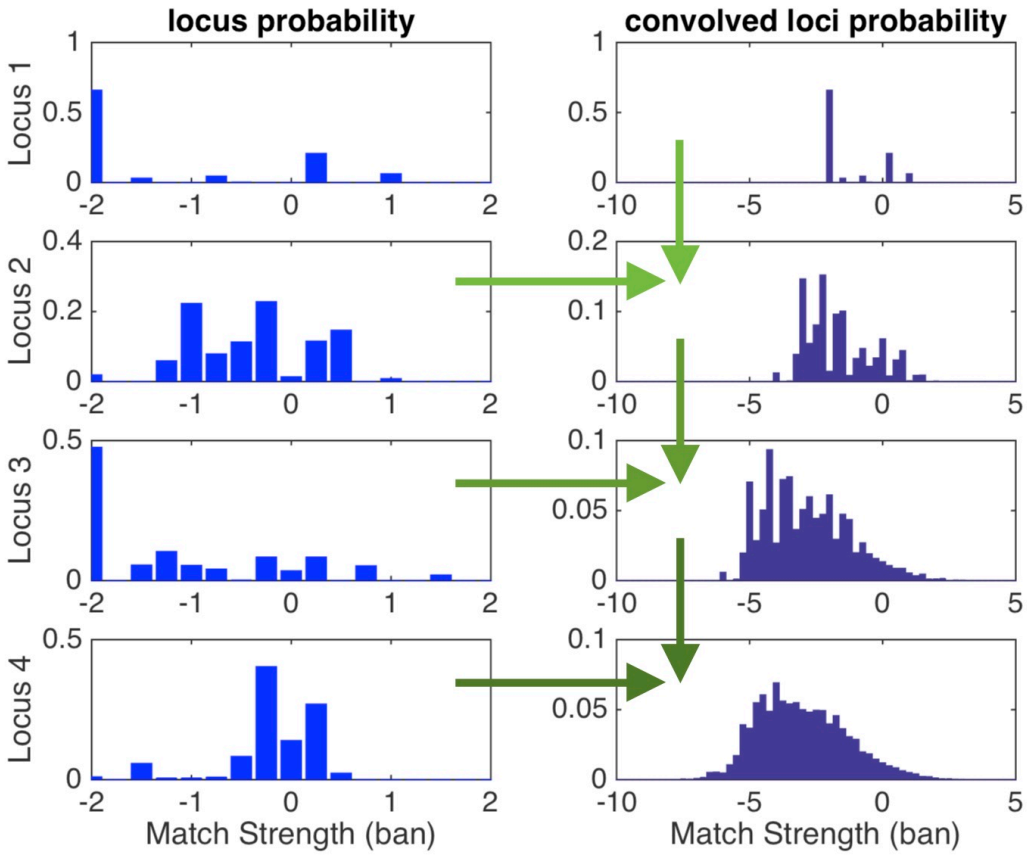
locus probability
convolved loci probability
Locus 1
Locus 2
Locus 3
Locus 4
Match Strength (ban)
Match Strength (ban)

Figure 7

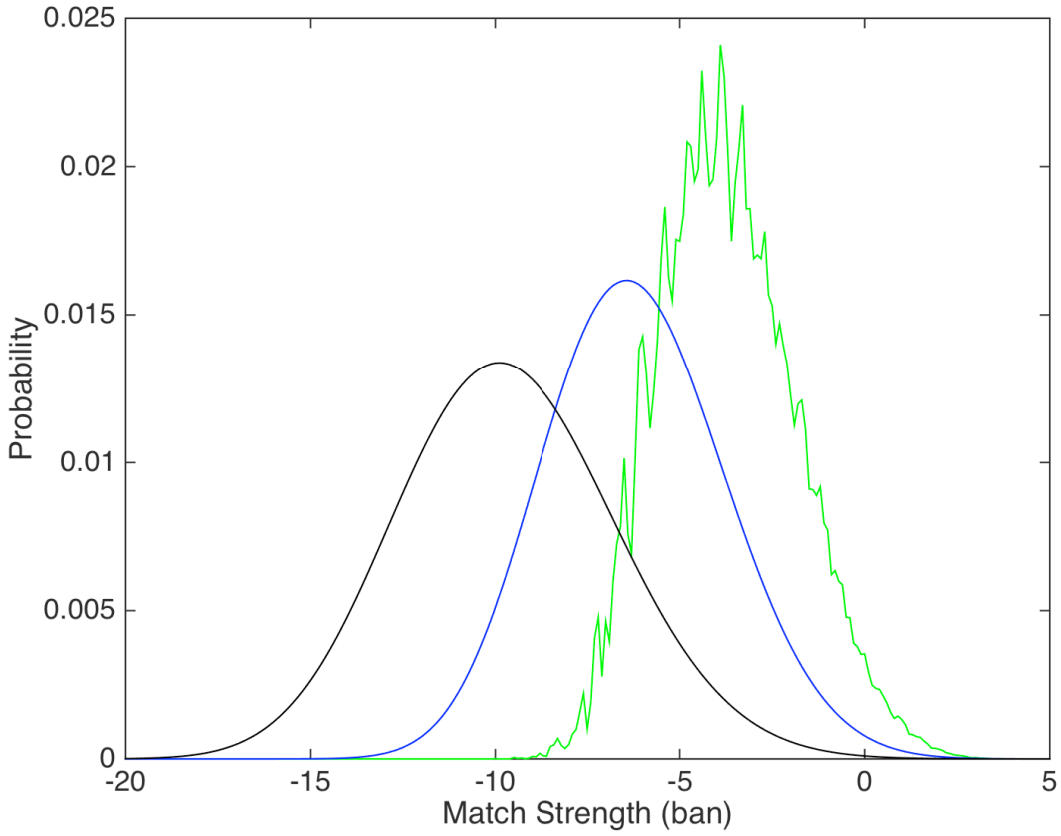
0.025
0.02
0.015
0.01
0.005
0
Probability
-20
-15
-10
-5
0
5
Match Strength (ban)

Figure 8

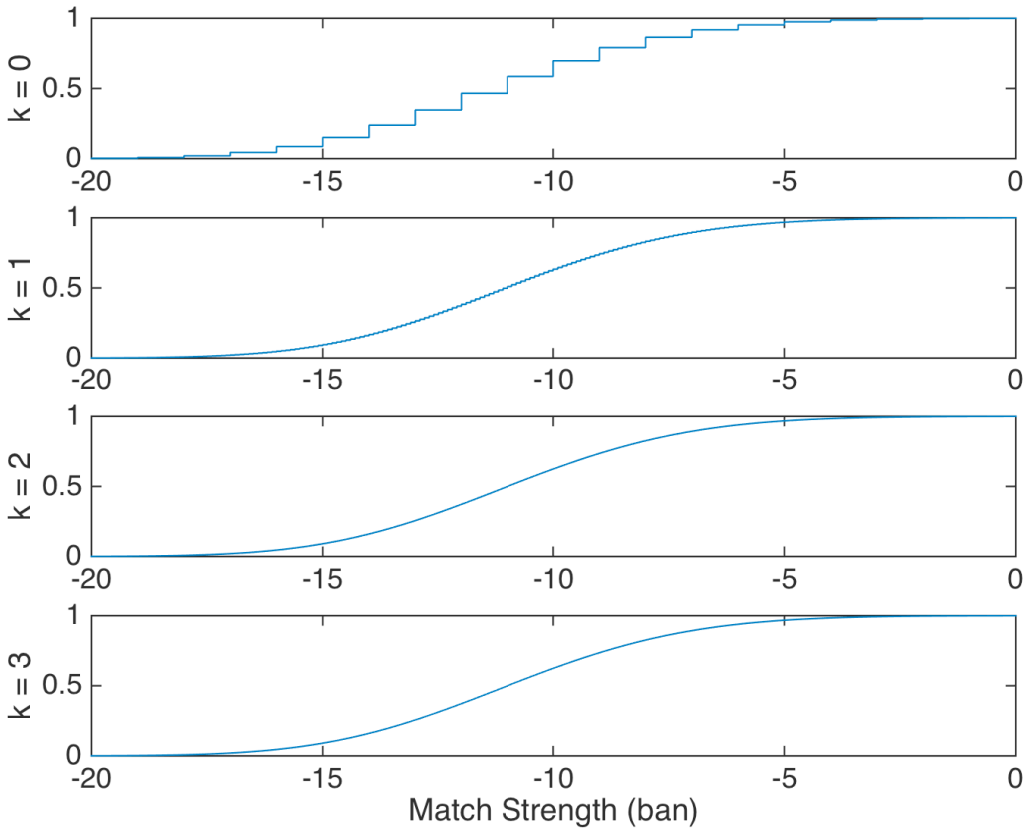

Figure 9

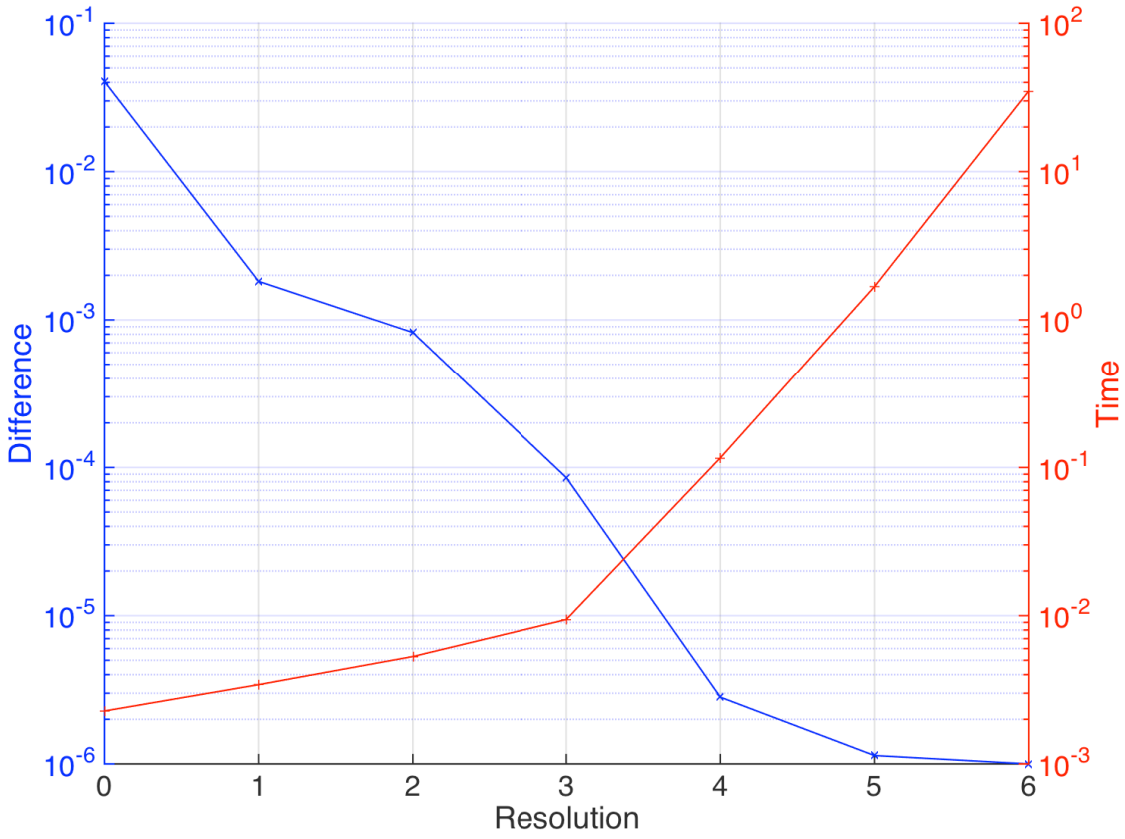
Difference
Time
Resolution
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
1
2
3
4
5
6

Figure 10

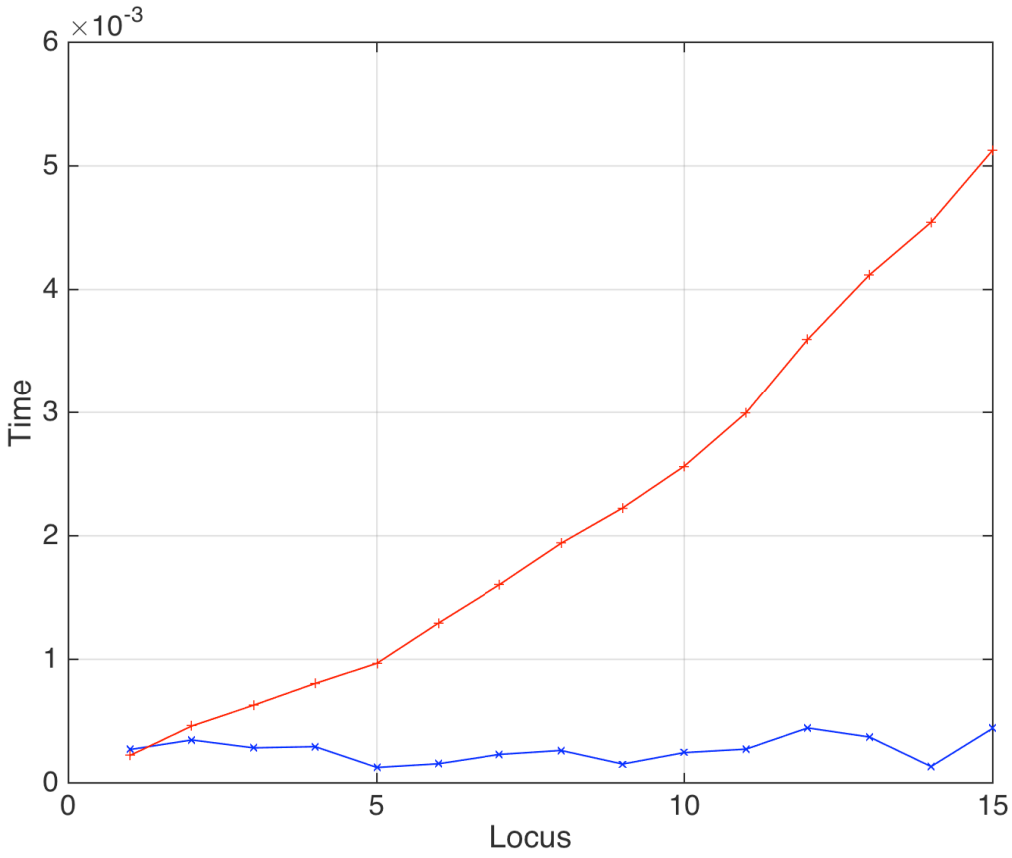

Figure 11

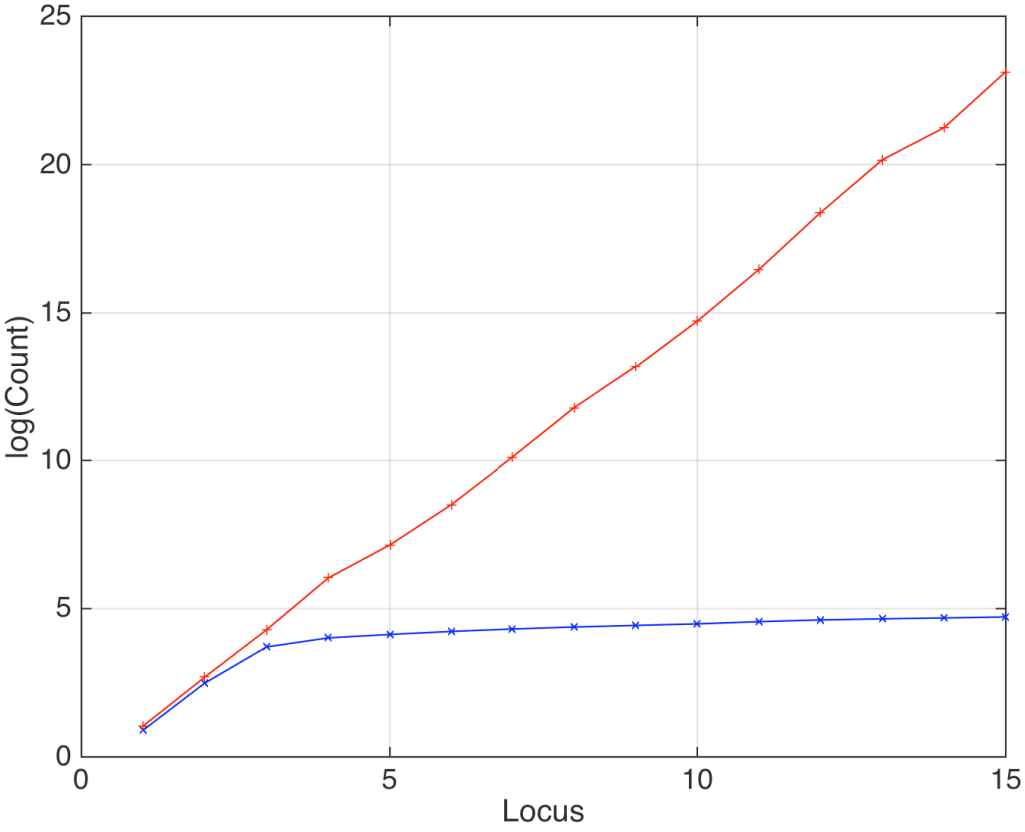

Figure 12

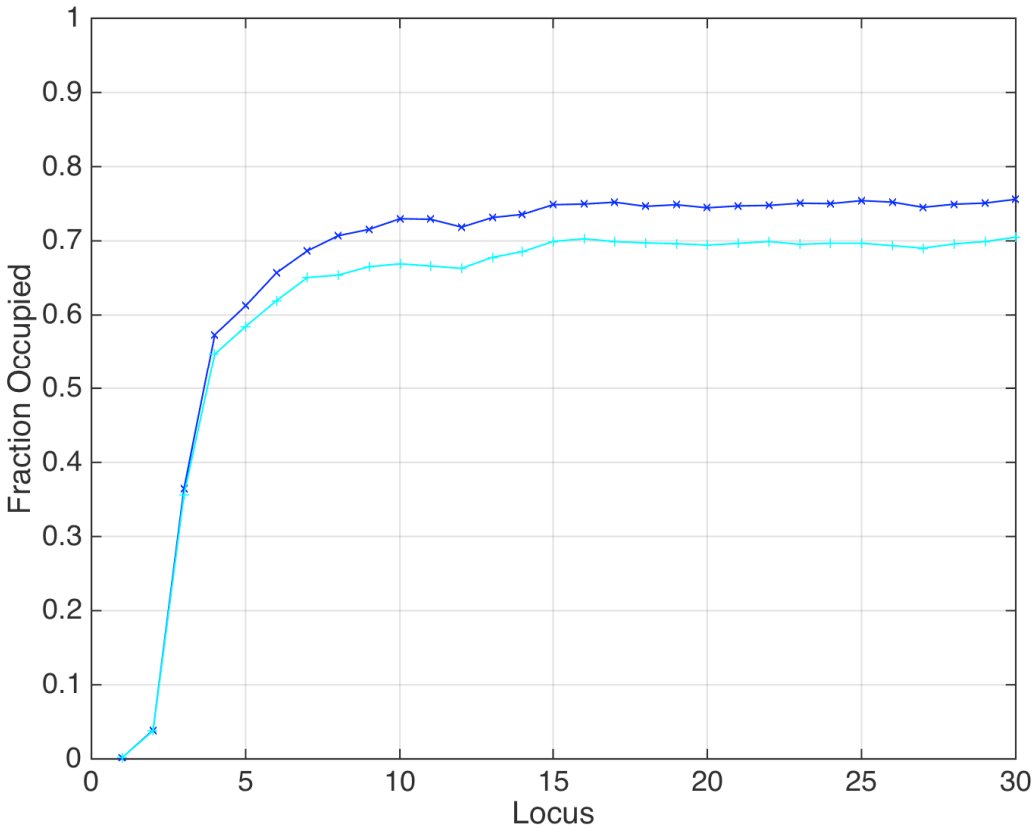
Fraction Occupied
Locus
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30

Figure 13