

# DNA Identification: Mixture Weight & Inference

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

TrueAllele® Lectures  
Fall, 2010



Cybergenetics © 2003-2010

## Mixture Weight: Uncertain Quantity

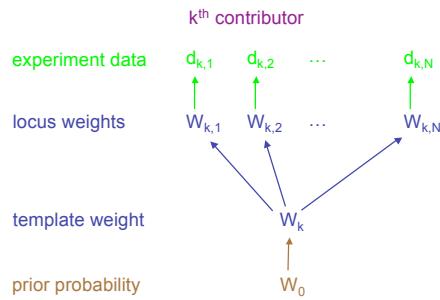
Infer mixture weight from  
STR experiments:

- quantitative peak data
- contributor genotypes

$\Pr(W=w | \text{data}, G_1=g_1, G_2=g_2, \dots)$   
hierarchical Bayesian model

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. Journal of Forensic Sciences. 2011;56(November):in press.

## Mixture Weight Model



## Experiment Estimate

$$w_k = \frac{\text{sum of peak heights from } k^{\text{th}} \text{ contributor}}{\text{sum of peak heights from all contributors}}$$

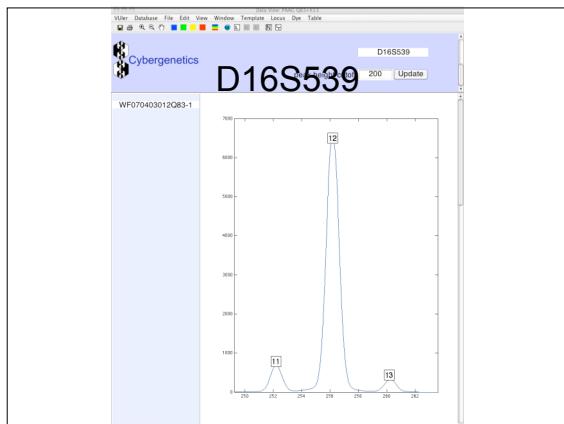
---

---

---

---

---



---

---

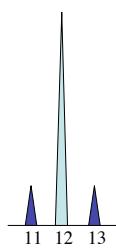
---

---

---

## Three Alleles

Allele	Quantity
11	500
12	6,750
13	250



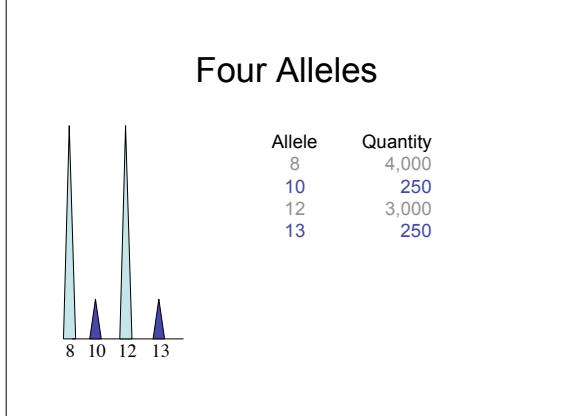
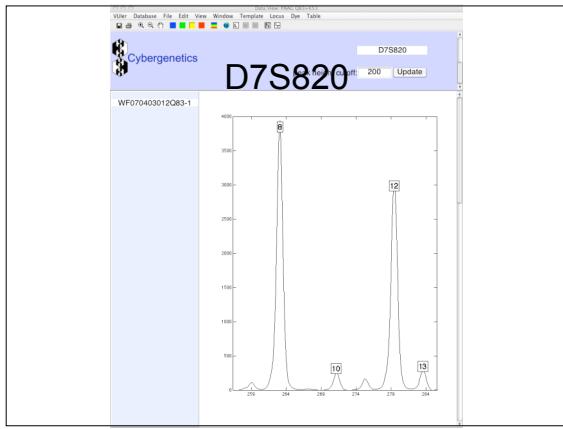
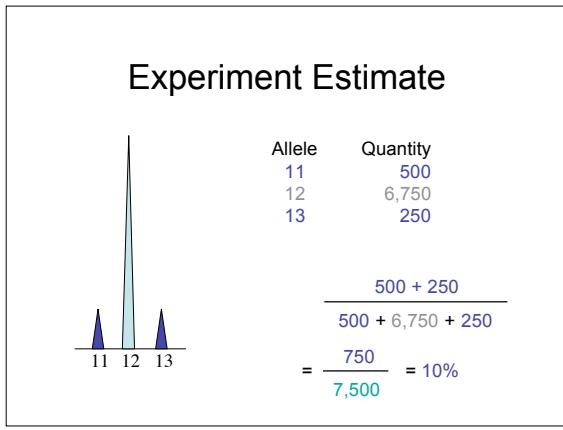
---

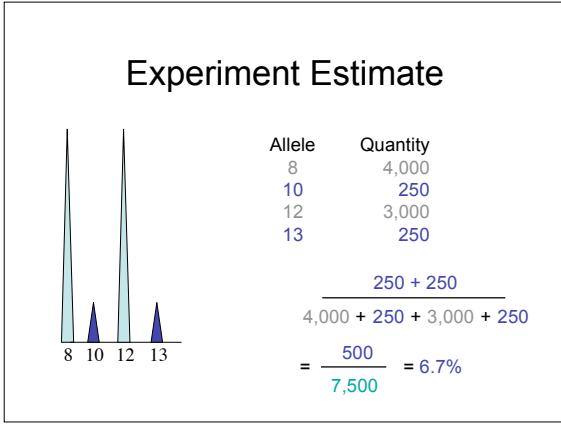
---

---

---

---





**Overlapping Alleles**

Mark W. Perlin, <sup>1</sup>Ph.D., M.D., Ph.D. and Bruce Szydlo, <sup>2</sup>Ph.D.

**Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples**

---

REPRINTED FROM: Perlin MW, Szydlo B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci 1998;43(5):861-870.

**ABSTRACT:** The analysis of PCR-amplified STR typing products from mixed DNA samples has become a routine forensic procedure. In this article we present a mathematical approach to the resolution of mixed DNA samples. This approach is based on the concept of linear mixture analysis, which is used to calculate the proportion of each individual's DNA in a mixed sample. The approach is based on the premise that each individual's DNA is represented by a unique set of allelic proportions. These proportions are determined by the relative amounts of each individual's DNA in the mixed sample. The approach is based on the premise that each individual's DNA is represented by a unique set of allelic proportions. These proportions are determined by the relative amounts of each individual's DNA in the mixed sample.

**KEYWORDS:** forensic science, DNA typing, STR, DNA mixture analysis, linear mixture analysis, mixed DNA samples

In forensic science, DNA samples are often found that contain DNA from more than one individual. This can occur due to a variety of reasons, such as sexual assault, homicide, or other criminal acts. The analysis of these mixed DNA samples is a complex process and requires specialized techniques. One common technique is called "linear mixture analysis," which involves the calculation of the proportion of each individual's DNA in the mixed sample. This proportion is determined by the relative amount of each individual's DNA in the mixed sample. The proportion is calculated by dividing the total amount of DNA in the mixed sample by the sum of the amounts of DNA from each individual. This proportion is then used to determine the likelihood of a particular individual being the source of the mixed DNA sample.

Linear mixture analysis is a mathematical approach to the resolution of mixed DNA samples. It is based on the premise that each individual's DNA is represented by a unique set of allelic proportions. These proportions are determined by the relative amounts of each individual's DNA in the mixed sample. The approach is based on the premise that each individual's DNA is represented by a unique set of allelic proportions. These proportions are determined by the relative amounts of each individual's DNA in the mixed sample.

Linear mixture analysis is a mathematical approach to the resolution of mixed DNA samples. It is based on the premise that each individual's DNA is represented by a unique set of allelic proportions. These proportions are determined by the relative amounts of each individual's DNA in the mixed sample.

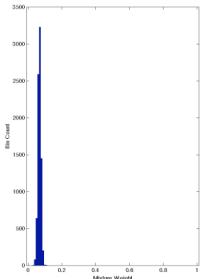
<sup>1</sup>Department of Pathology, University of California San Francisco, San Francisco, CA 94158; <sup>2</sup>Department of Pathology, University of Michigan, Ann Arbor, MI 48105

**Template Average**

mean  $\mu_k = \frac{1}{N} \sum_{n=1}^N w_{k,n}$

variance  $\sigma_w^2 = \frac{1}{N} \sum_{n=1}^N (w_{k,n} - \mu_k)^2$

## Template Mixture Weight Probability Distribution



mean = 6.7%

standard deviation = 0.9%

## Central Limit Theorem

- more data experiments for a template provide greater mixture weight precision
- double the precision by doing four times the number of experiments
- combine evidence from multiple experiments to obtain a more informative result

## Probability Solution

interacting random variables

$w | d, g_1, g_2, \dots$

$g_1 | d, g_2, w, \dots$

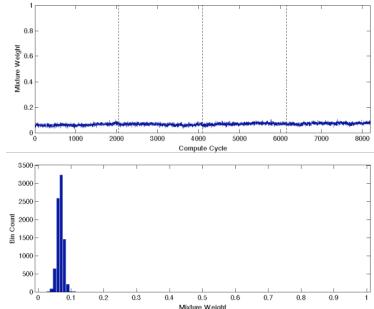
$g_2 | d, g_1, w, \dots$

$z_i | d, g_1, g_2, w, \dots$

find probability distributions by iterative sampling

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *J. American Statist. Assoc.*, 85:398-409.

## Markov Chain Monte Carlo



## Prior Probability

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

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

DNA quantity       $m_l \sim N_+(5000, 5000^2)$

variance parameters       $\sigma^{-2} \sim Gam(10, 20)$   
 $\tau^{-2} \sim Gam(10, 500)$   
 $\psi^{-2} \sim Gam(1/2, 1/200)$

## Joint Likelihood Function

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

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

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

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

## Posterior Probability

$$\Pr\{Q = x | d_{i,1}, d_{i,2}, \dots, d_{i,n} \} \propto \Pr\{Q = x\} \cdot \prod_{j=1}^J \Pr\{d_{i,j} | Q = x, \dots\}$$

genotype

$$\Pr\{W = w | d_1, d_2, \dots, d_J, \dots\} \propto \Pr\{W = w\} \cdot \prod_{j=1}^J \Pr\{d_j | W = w, \dots\}$$

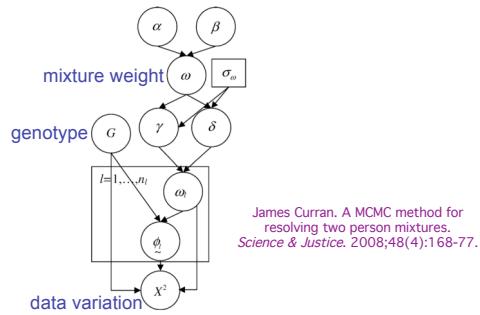
mixture weight

$$\Pr\{\sigma^2 = s^2 | d_1, d_2, \dots, d_J, \dots\} \propto \Pr\{\sigma^2 = s^2\} \cdot \prod_{j=1}^J \Pr\{d_j | \sigma^2 = s^2, \dots\}$$

data variation

$$\Pr\{\tau^2 = t^2 | d_1, d_2, \dots, d_J, \dots\} \propto \Pr\{\tau^2 = t^2\} \cdot \prod_{j=1}^J \Pr\{d_j | \tau^2 = t^2, \dots\}$$

## Generally Accepted Method



## Hierarchical Bayesian Model with MCMC Solution

- standard approach in modern science
- describes uncertainty using probability
- the "new calculus"
- replaces hard calculus with easy computing
- can solve virtually any problem
- well-suited to interpreting DNA evidence