NIST National Institute of Standards and Technology U.S. Department of Commerce

NEAFS
Probabilistic DNA Mixture Interpretation Workshop
Allentown, PA
September 25-26, 2015

Likelihood Ratios for Mixtures: Continuous Approach

Simone Gittelson, Ph.D., simone.gittelson@nist.gov
Michael Coble, Ph.D., michael.coble@nist.gov

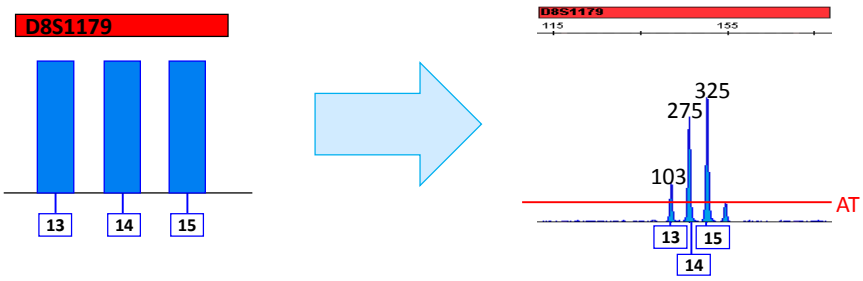
Acknowledgement

I thank Michael Coble, Bruce Weir and John Buckleton for their helpful discussions.

Disclaimer

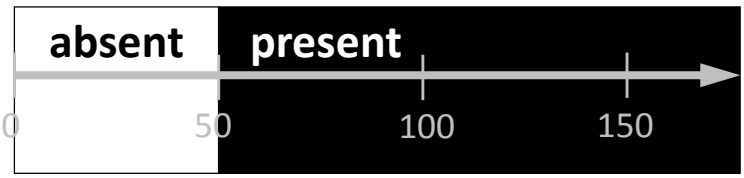
Points of view in this presentation are mine and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

From Semi-continuous to Continuous



Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



- The observed peaks as a **continuous** variable:

Discrete vs. Continuous

The observed peak as a **discrete** variable:

$Pr(peak\ absent) = \bar{C}[(1-p_{15}) + p_{15}D]$

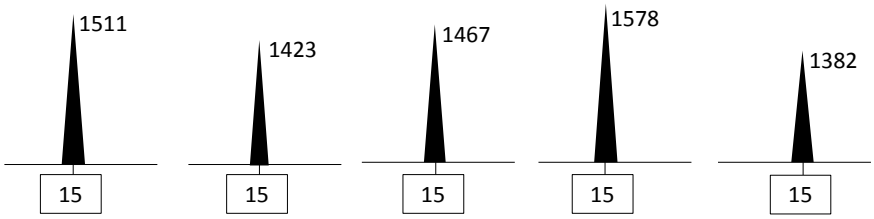
no drop-in and donor has no 15
or
no drop-in, donor has 15 and drop-out

$Pr(peak\ present) = p_{15}\bar{D} + (1-p_{15})Cp_{15}$

donor has 15 and no drop-out
or
donor has no 15 and drop-in of 15

Discrete vs. Continuous

The observed peaks as a **continuous** variable:

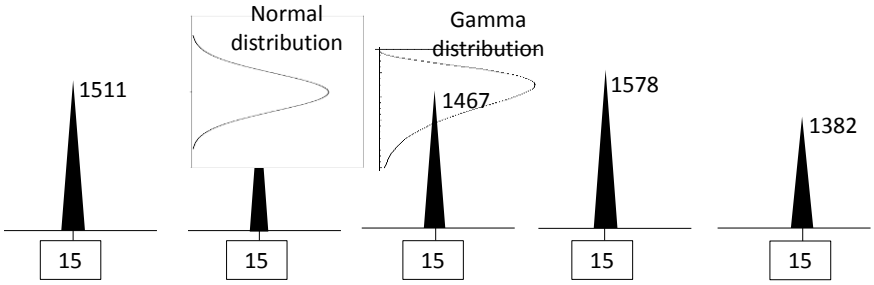


$Pr(peak\ is\ 1511rfu)$ $Pr(peak\ is\ 1467rfu)$ $Pr(peak\ is\ 1382rfu)$

$Pr(peak\ is\ 1423rfu)$ $Pr(peak\ is\ 1578rfu)$

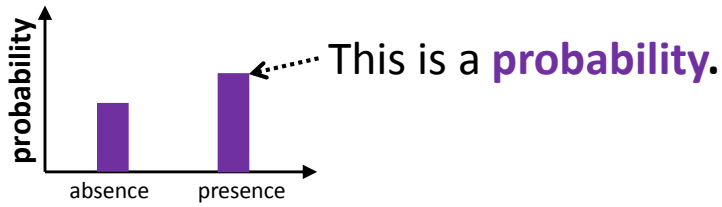
Discrete vs. Continuous

The observed peaks as a **continuous** variable:

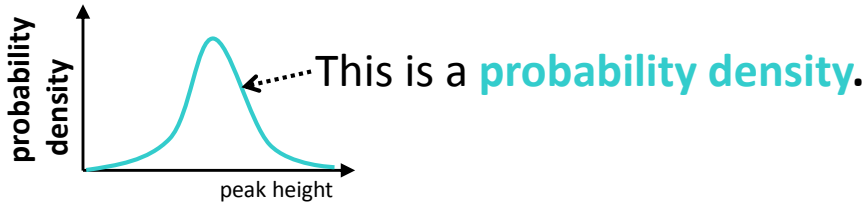


Discrete vs. Continuous

- The observed peaks as a **discrete** variable:



- The observed peaks as a **continuous** variable:



Likelihood Ratio: from semi-continuous to continuous

$$LR = \frac{\sum_{genotype\ sets|H_p} weight \times genotype\ match\ prob.}{\sum_{genotype\ sets|H_d} weight \times genotype\ match\ prob.}$$

stays the same

(Hardy-Weinberg Law, NRC II recommendation 4.1, NRC II recommendation 4.2)

Likelihood Ratio: from semi-continuous to continuous

$$LR = \frac{\sum_{genotype\ sets|H_p} weight \times genotype\ match\ prob.}{\sum_{genotype\ sets|H_d} weight \times genotype\ match\ prob.}$$

changes

semi-continuous weight: any value between 0 and 1, including 0 and 1, assigned using probabilities of drop-out and drop-in

continuous: a **probability density**, generally assigned based on the results of simulations that attempt to reproduce the observed peak heights

Likelihood Ratio

$$LR = \frac{\sum_{j=1}^{m} \mathrm{f}(G_{CS}|S_j)\Pr(S_j|G_K, H_p, I)}{\sum_{j=1}^{m} \mathrm{f}(G_{CS}|S_j)\Pr(S_j|G_K, H_d, I)}$$

match probabilities

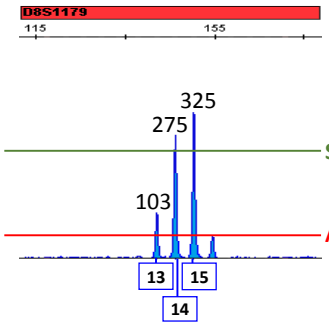
the probability of genotype set S_j given that we have observed genotypes G_K and that the contributors are as specified in H_p or H_d

Likelihood Ratio

$$LR = \frac{\sum_{j=1}^{m} \mathrm{f}(G_{CS}|S_j)\Pr(S_j|G_K, H_p, I)}{\sum_{j=1}^{m} \mathrm{f}(G_{CS}|S_j)\Pr(S_j|G_K, H_d, I)}$$

weights

the probability of obtaining these DNA typing results for genotype set S_j



Probabilistic Genotyping

weight

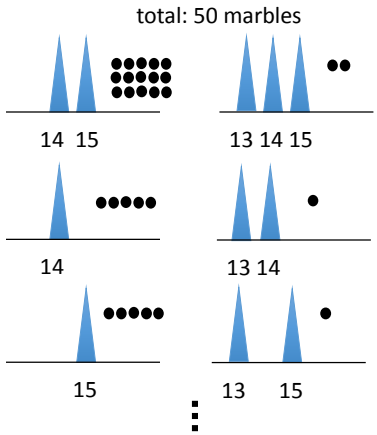
Major	Minor	*w*	*match prob*$_{\cdot major}$	*match prob*$_{\cdot minor}$
14,15	12,13		$2p_{14}p_{15}$	$2p_{12}p_{13}$
14,15	13,13		$2p_{14}p_{15}$	p_{13}^2
14,15	13,14		$2p_{14}p_{15}$	$2p_{13}p_{14}$
14,15	13,15		$2p_{14}p_{15}$	$2p_{13}p_{15}$
14,15	13,16		$2p_{14}p_{15}$	$2p_{13}p_{16}$
14,15	13,17		$2p_{14}p_{15}$	$2p_{13}p_{17}$

Concept of weights

the probability of obtaining these DNA typing results for a particular genotype set

Concept of weights

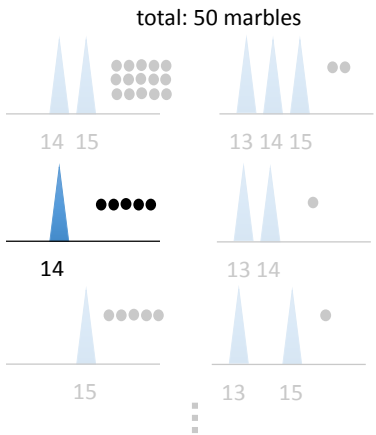
single source:



genotype = {14,15}

Concept of weights

single source:



genotype = {14,15}

$$\text{weight} = \frac{5}{50}$$

Concept of weights

how large or small a weight is depends on:

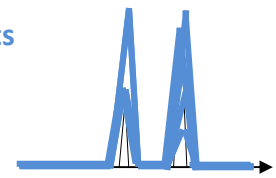
- possibility of allele drop-out
- possibility of allele drop-in
- amount of template DNA
- possibility of degradation
- possibility of oversized stutter peaks

⋮

Continuous Model: short explanation

Simulations attempt to reproduce the observed peak heights by varying a set of parameters.

simulated peak heights



observed peak heights

The better S_j and the set of parameters explain G_{CS}, the greater the probability density $f\left(G_{CS}|S_j\right)$ assigned to that genotype set.

Continuous model: longer explanation

1) Randomly choose a genotype set and values for all of the parameters that are necessary to model the expected peak heights.

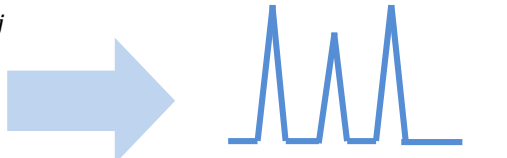
genotype set S_j
parameter 1
parameter 2
parameter 3
⋮

Markov Chain Monte Carlo (MCMC) sampling

Continuous model: longer explanation

2) Model the expected peak heights given the chosen genotype set S_j and parameter values.

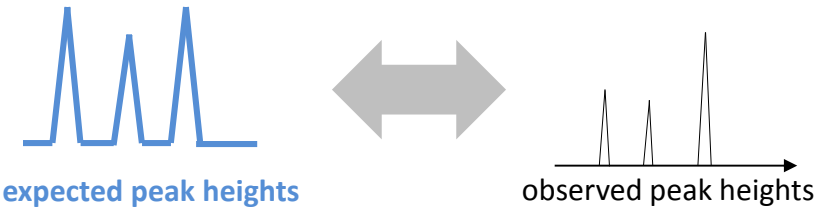
genotype set S_j
parameter 1
parameter 2
parameter 3
⋮



expected peak heights

Continuous model: longer explanation

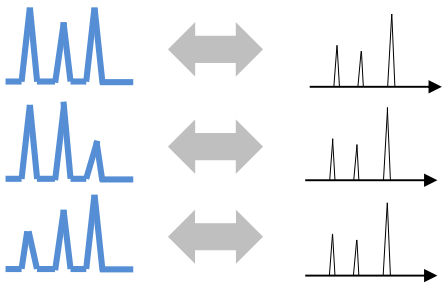
3) Compare the expected peak heights with the observed peak heights. How similar are they?



Assign a probability density for the observed peak heights given the expected peak heights.

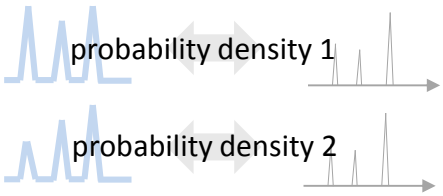
Continuous model: longer explanation

4) Repeat steps 1) through 3) a large number of times. A large number of simulations are performed that randomly vary the genotype set and the parameter values.



Continuous model: longer explanation

3) Compare the expected peak heights with the observed peak heights. How similar are they?
4) Repeat steps 1) through 3) a large number of times. A large number of simulations are performed that randomly vary the genotype set and the parameter values.

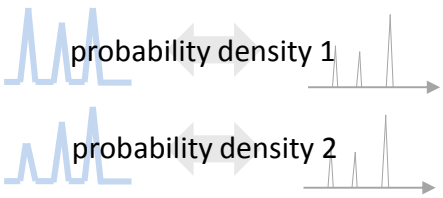


If *probability density 2* $\geq$ *probability density 1*, save S_j and the parameter values of simulation 2.

If *probability density 2* < *probability density 1*, save S_j and the parameter values of simulation 2 only some of the times. The probability of saving the values is equal to $\frac{probability\ density\ 2}{probability\ density\ 1}$.

Continuous model: longer explanation

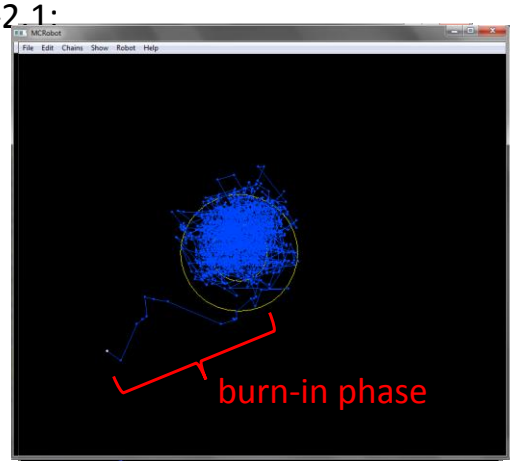
3) Compare the expected peak heights with the observed peak heights. How similar are they?
4) Repeat steps 1) through 3) a large number of times. A large number of simulations are performed that randomly vary the genotype set and the parameter values.



Metropolis-Hastings algorithm

Continuous model: longer explanation

MCRobot-2_1:



after taking 200 samples

384

Continuous model: longer explanation

5) Assign the weight $f(G_{CS}|S_j)$ based on the saved simulation results.

discard burn-in

saved simulation results:		
Genotype set	*Parameter 1*	*Parameter 2*
~~6,7 and 8,9~~		
~~6,7 and 8,8~~		
~~7,7 and 6,8~~		
6,7 and 8,8		
6,7 and 8,8		
6,8 and 7,8		
6,7 and 8,8		
6,7 and 8,8		
6,8 and 7,8		

Continuous model: longer explanation

5) Assign the weight $f(G_{CS}|S_j)$ based on the saved simulation results.

The weight $f(G_{CS}|S_j)$ is assigned as the proportion of genotype sets S_j among the saved genotype sets.

Example: S_j = 6,7 and 8,8

→ $f(G_{CS}|S_j) = \frac{4}{6}$

saved simulation results:		
Genotype set	*Parameter 1*	*Parameter 2*
~~6,7 and 8,9~~		
~~6,7 and 8,8~~		
~~7,7 and 6,8~~		
6,7 and 8,8		
6,7 and 8,8		
6,8 and 7,8		
6,7 and 8,8		
6,7 and 8,8		
6,8 and 7,8		

Continuous model: longer explanation

simulations

1) Randomly choose a genotype set S_j and values for all of the parameters that are necessary to model the expected peak heights.
2) Model the expected peak heights given the chosen genotype set S_j and parameter values.
3) Compare the expected peak heights with the observed peak heights. How similar are they?
4) Repeat steps 1) through 3) a large number of times. A large number of simulations are performed that randomly vary the genotype set S_j and the parameter values.
5) Assign the weight $f(G_{CS}|S_j)$ based on the saved simulation results.

Summary of main points

- The **peak heights** are a **continuous variable**.
- Mathematically speaking, $f\left(G_{CS}|S_j\right)$ is the **probability density** for the observed peak heights given the genotype set S_j.
- The probability densities $f\left(G_{CS}|S_j\right)$ are assigned based on the results of simulations that attempt to reproduce the observed peak heights by varying a set of parameters.