DNA Identification: Quantitative Data Modeling

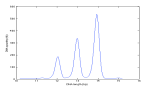
Mark W Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

TrueAllele® Lectures
Fall, 2010

Cybergenetics

Quantitative Data

- PCR is a linear process
- peak heights reflect the underlying DNA quantity
- use quantitative peak heights to explain the observed data



Genotype Model: 1 + 1 = 2

victim genotype

second genotype?
12, 14 ?

+

genotype pattern

=

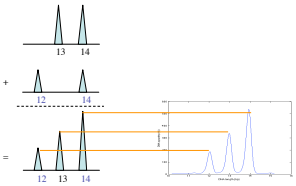
Consider all possible allele pair values by trying out each candidate

Compare Model to Data

victim genotype

second genotype?
12, 14 ?

genotype pattern



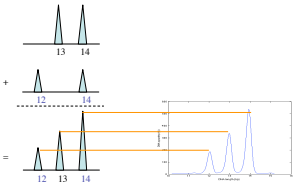
data

Likelihood Function

victim genotype

second genotype?
12, 14 ?

genotype pattern



data

Deviation small small small

Likelihood large x large x large = large

12, 14 is very likely

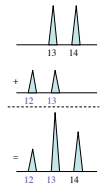
$\prod Pr(data_{peak}|Q=x,\ldots)$
joint likelihood function

Genotype: Alternative Value

victim genotype

second genotype?
12, 13 ?

genotype pattern



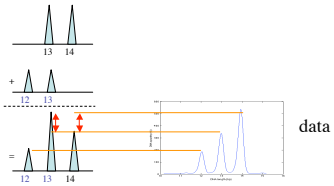
Consider a different allele pair value by trying out another candidate

Compare Model to Data

victim genotype

second genotype?
12, 13 ?

genotype pattern



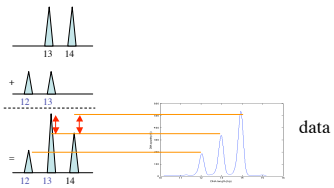
data

Likelihood Function

victim genotype

second genotype?
12, 13 ?

genotype pattern



data

Deviation small large large

Likelihood large x small x small = small

12, 13 is less likely

$\prod \Pr(data_{peak} | Q=x,\ldots)$
joint likelihood function

All Genotype Possibilities

allele pair	likelihood	probability
12, 12	small	5%
12, 13	small	5%
12, 14	large	90%
13, 13	very small	0
13, 14	very small	0
14, 14	very small	0
12, 15	0	0
13, 16	0	0
sum =		100%

prior × likelihood ⇒ posterior

Genotype inference

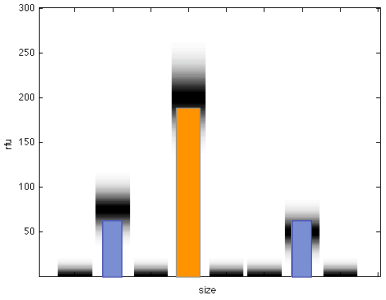
$Pr(Q=x|data,\ldots) \propto \prod Pr(data|Q=x,\ldots) \times Pr(Q=x)$

posterior probability	joint likelihood function	prior probability

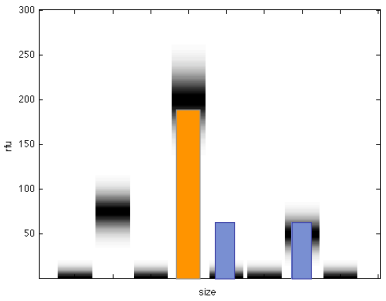
Try out all value possibilities;
better fit's more likely it.

$\prod Pr(data_{locus}|Q=x,\ldots)$
joint likelihood function

Genotype probability with data uncertainty



Genotype alternative value



Bayesian probability

- Assess ALL genotype patterns to find the probability of each allele pair.
- Similarly compute the data variance.
- Small data variation is RESTRICTIVE: only few genotype values are possible. (more certain)
- Large data variation is PERMISSIVE: many genotype values are possible. (less certain)

Likelihood ratio match statistic reflects genotype uncertainty

$$LR = \frac{Pr(Q=s|data)}{Pr(Q=s)}$$

Genotype certainty concentrates probability on just a few good bets, and focuses LR. (more info)

Genotype uncertainty diffuses probability across many candidates, and reduces LR. (less info)

Mixture weight inference

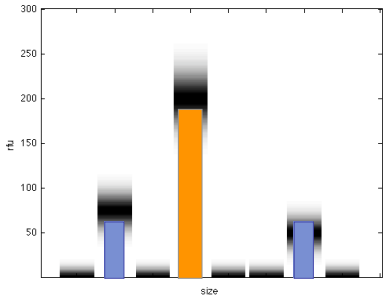
$$Pr(W=w|data,\ldots) \propto \prod Pr(data|W=w,\ldots) \times Pr(W=w)$$

posterior probability | joint likelihood function | prior probability

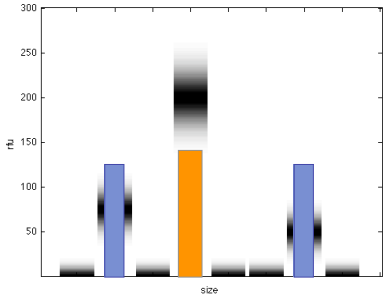
Try out all value possibilities; better fit's more likely it.

$\prod Pr(data_{locus}|W=w,\ldots)$
joint likelihood function

Mixture weight probability with data uncertainty



Mixture weight alternative



Data variance inference

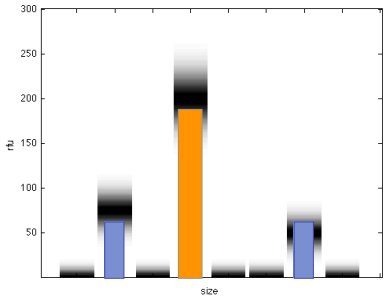
$Pr(V=v|data,\ldots) \propto \prod Pr(data|V=v,\ldots) \times Pr(V=v)$

posterior probability — joint likelihood function — prior probability

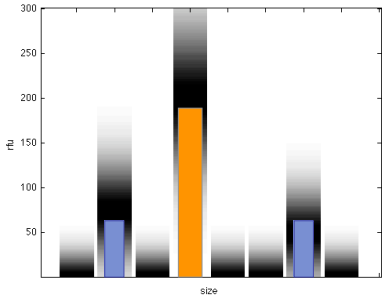
Try out all value possibilities; better fit's more likely it.

$\prod Pr(data_{peak}|V=v,\ldots)$
joint likelihood function

Data variance probability of data peak uncertainty



Data variance alternative



Quantitative data modeling

- genotype is main variable of interest
- genotype gives identification LR
- mixture weight is explanatory variable
- data variance, stochastic effects
- identification information
  preserved by quantitative modeling