Getting Past First Bayes with DNA Mixtures

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

ABSTRACT

DNA mixtures are a prevalent form of biological evidence. A mixture contains DNA from two or more contributors. There are usually multiple genotype explanations for the observed STR data. Forensic scientists must understand genotype mixture inference in order to give accurate DNA mixture testimony in court.

Fortunately, Bayes theorem provides a robust framework for genotype inference and match. Over 250 years ago, the Rev. Thomas Bayes showed us how to update our belief in hypotheses (probability) by examining how well those hypotheses explain observed data (likelihood). Bayes has us use all the data, and consider all hypotheses.

Bayesian genotype inference (for each contributor at every genetic locus) begins with a *prior* belief that the chance of observing an allele pair before seeing data is proportional to its population prevalence. Careful examination of STR data then uses a *likelihood* function to concentrate probability on those genotype values that best explain the laboratory data. This objectively inferred genotype associates a *posterior* probability with every allele pair, multiplying prior and likelihood.

A DNA match statistic assesses the strength of match between evidence and reference genotypes, relative to coincidence. This Bayesian likelihood ratio (LR) weighs two competing hypotheses – either the reference individual contributed DNA to the evidence, or he did not – based on the observed STR data.

Bayesian beginners often make mistakes. They may fail to use all peak data or not consider all genotype hypotheses. They can confuse likelihood (chance of data given hypothesis) with probability (chance of hypothesis given data). A beginner will apply complex formulas when a simple ratio would suffice. They may change their assumptions in mid-step, and suggest meaningless comparisons.

On April 12, 2013, the National Institute of Standards and Technology (NIST) Applied Genetics Group gave a full day webinar on DNA mixture interpretation. The NIST group presented genotype and LR results from Bayesian software. Since their expertise lies elsewhere, they made many beginner errors and never got past first Bayes. Errors that appear harmless in an academic setting can prove fatal in a court of law, where accuracy is paramount and cross-examination unforgiving.

This poster reviews the basic principles of Bayesian DNA mixture interpretation. The NIST webinar errors are used as teaching points to help beginners avoid common mistakes. The corrections we provided NIST highlight key interpretation steps. With some Bayesic training, DNA analysts can accurately testify about mixture results, and get past first Bayes.

Cybergenetics

PRIOR PROBABILITY

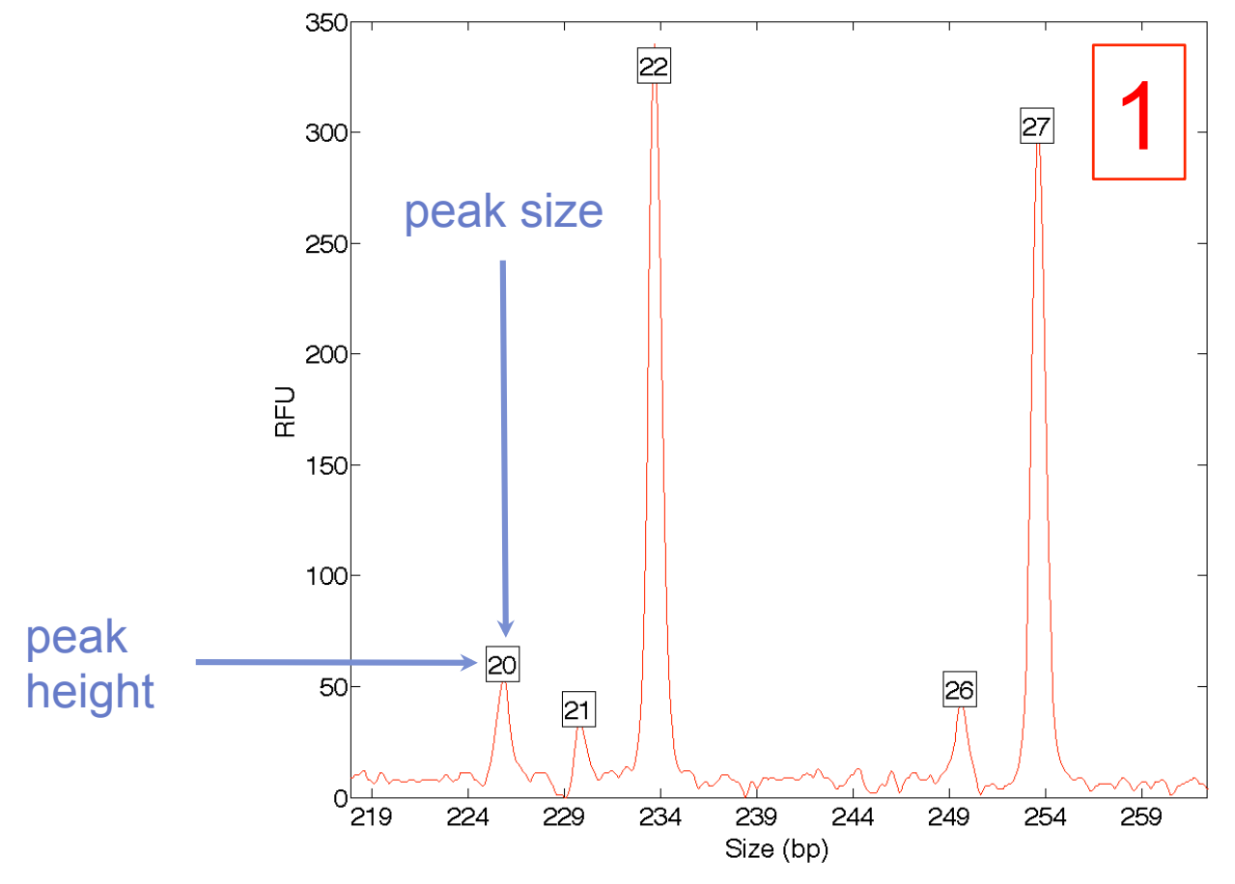


Figure 1. DNA mixture data.

Simple DNA evidence can be analyzed simply, often by just inspecting the short tandem repeat (STR) data. This is because the one or two allele peaks provide overwhelming evidence for but a single genotype possibility.

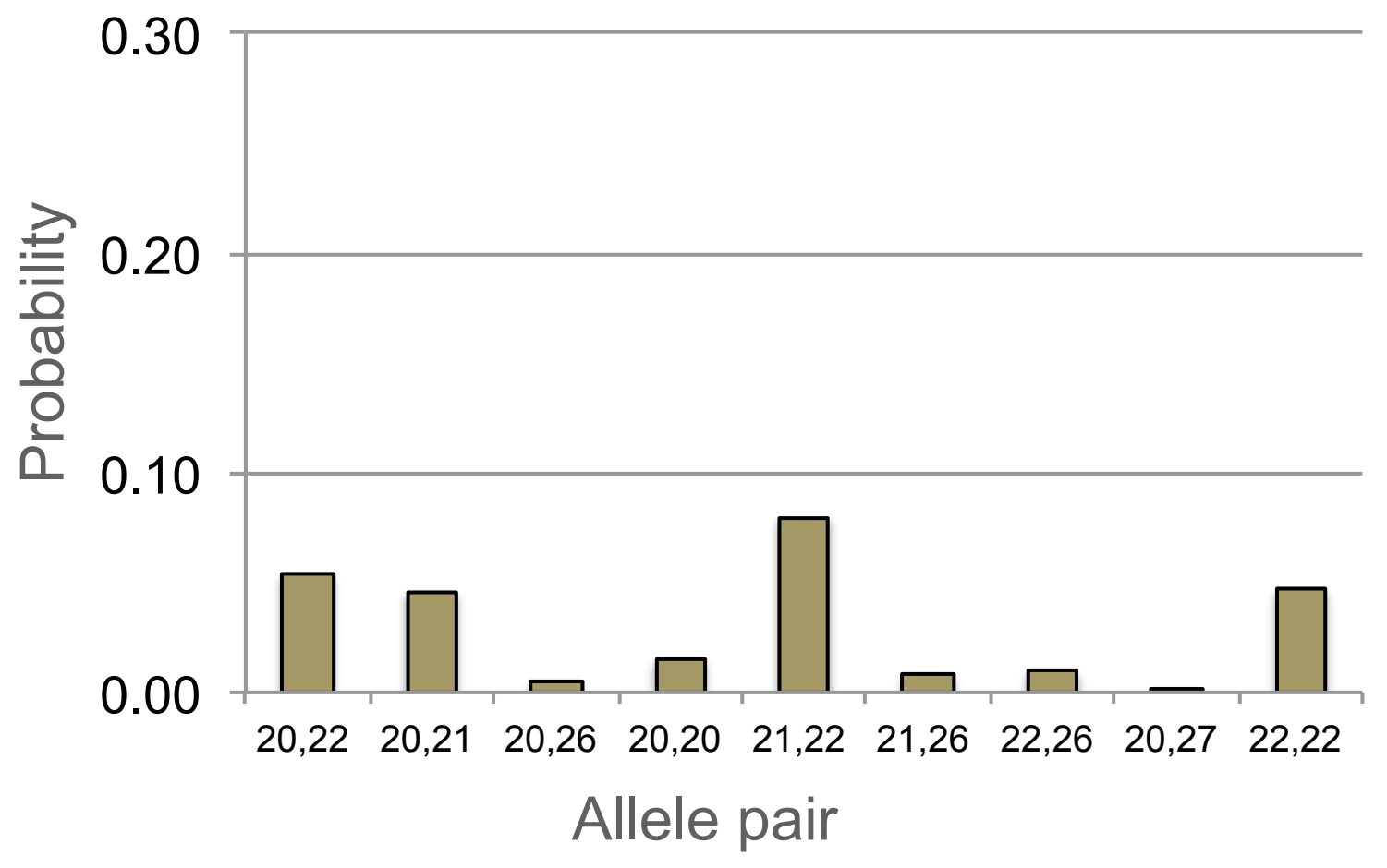
Mixtures are a more complex form of DNA evidence, for which there may exist multiple genotype possibilities (Fig 1). To address this uncertainty, probability is required.

At each genetic locus there are a dozen or so possible alleles, hence about a hundred feasible allele pairs. Some of these genotypes are more probable than others, because their alleles are more prevalent in the population.

The *prior probability* of a genotype (for a contributor at a locus) is what we believe the allele pair will be before ever seeing the STR data. This amounts to just the population genotype probability distribution, based on allele prevalence (Fig 2).

Allele Pair	Prior
20, 22	0.0543
20, 21	0.0461
20, 26	0.0058
20, 20	0.0156
21, 22	0.0800
21, 26	0.0085
22, 26	0.0100
20, 27	0.0008
22, 22	0.0471
	0.2682

Figure 2. Prior genotype probability (brown).



DATA & LIKELIHOOD

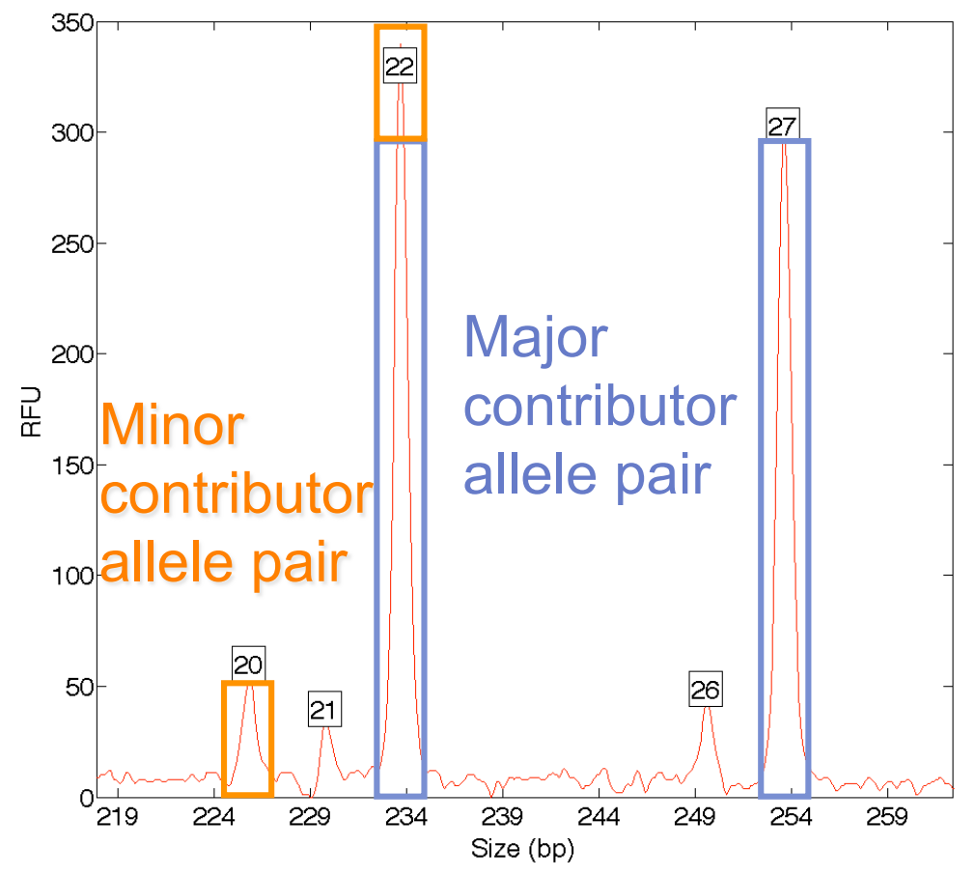


Figure 3. A good explanation of the data has a higher likelihood.

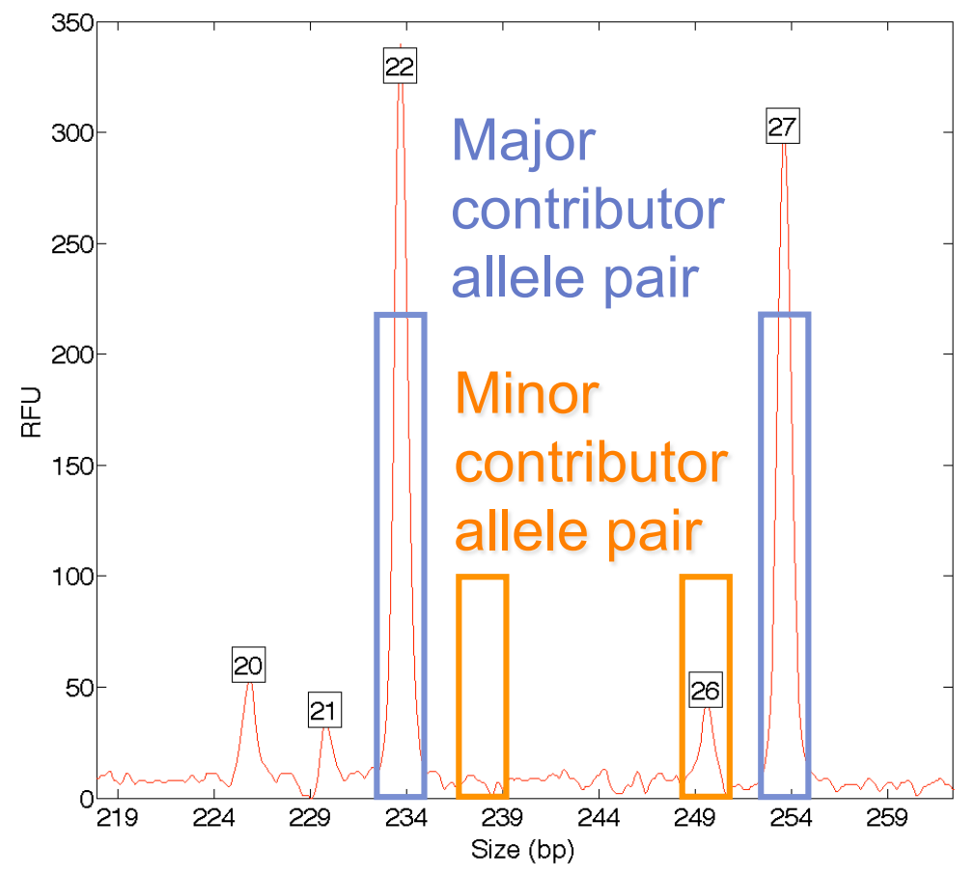


Figure 4. A poor explanation of the data has a lower likelihood.

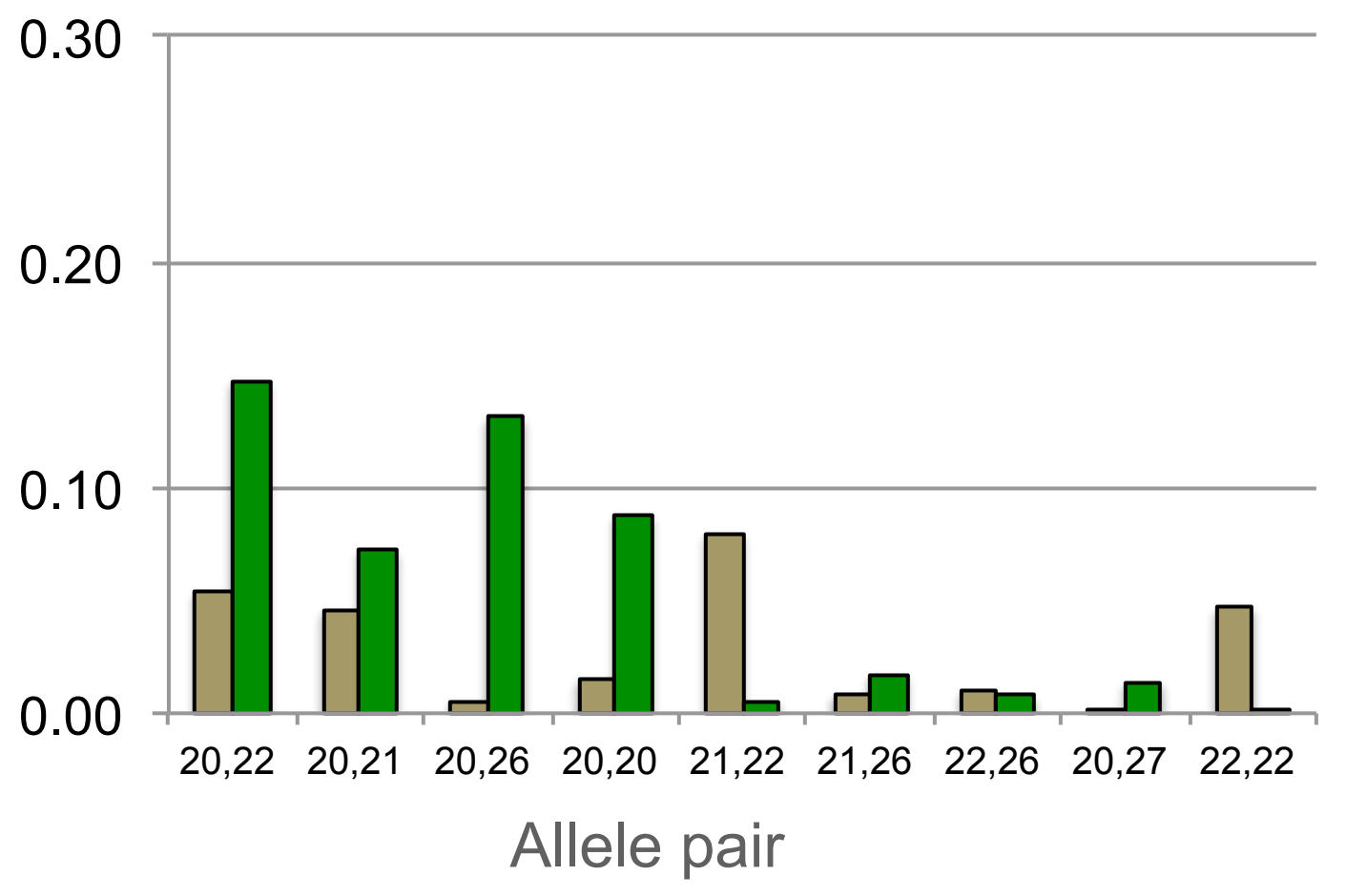
STR data changes our prior belief in the genotype. This change is moderated through a *likelihood* function that expresses how well a genotype hypothesis explains the observed data.

Good data explanations impart higher likelihood to hypothesized genotypes (Fig 3), while poor explanations confer a lower likelihood to a genotype (Figs 4 & 5).

2

Allele Pair	Prior	Likelihood
20, 22	0.0543	0.1474
20, 21	0.0461	0.0722
20, 26	0.0058	0.1309
20, 20	0.0156	0.0882
21, 22	0.0800	0.0056
21, 26	0.0085	0.0176
22, 26	0.0100	0.0077
20, 27	0.0008	0.0142
22, 22	0.0471	0.0010
	0.2682	*0.4848*

Figure 5. Genotype likelihood (green).



POSTERIOR PROBABILITY

Figure 6. The Reverend Thomas Bayes, born 1701 (London, England), died 1761 (Tunbridge Wells, Kent).

Bayes Theorem (Fig 6) is the mathematical rule for updating a belief based on evidence. Starting with some prior probability in a hypothesis, and multiplying that by the likelihood of the hypothesis (based on data), Bayes computes the *posterior probability* of the hypothesis (Fig 7).

As shown in the table, the products of priors and likelihoods generally do not add up to 1 (*Prior*Like*). Therefore, Bayes renormalizes these products, dividing by the sum of the products (*Prior*Like* total). This renormalization ensures that the resulting values sum to 1, producing a probability distribution (*Posterior*).

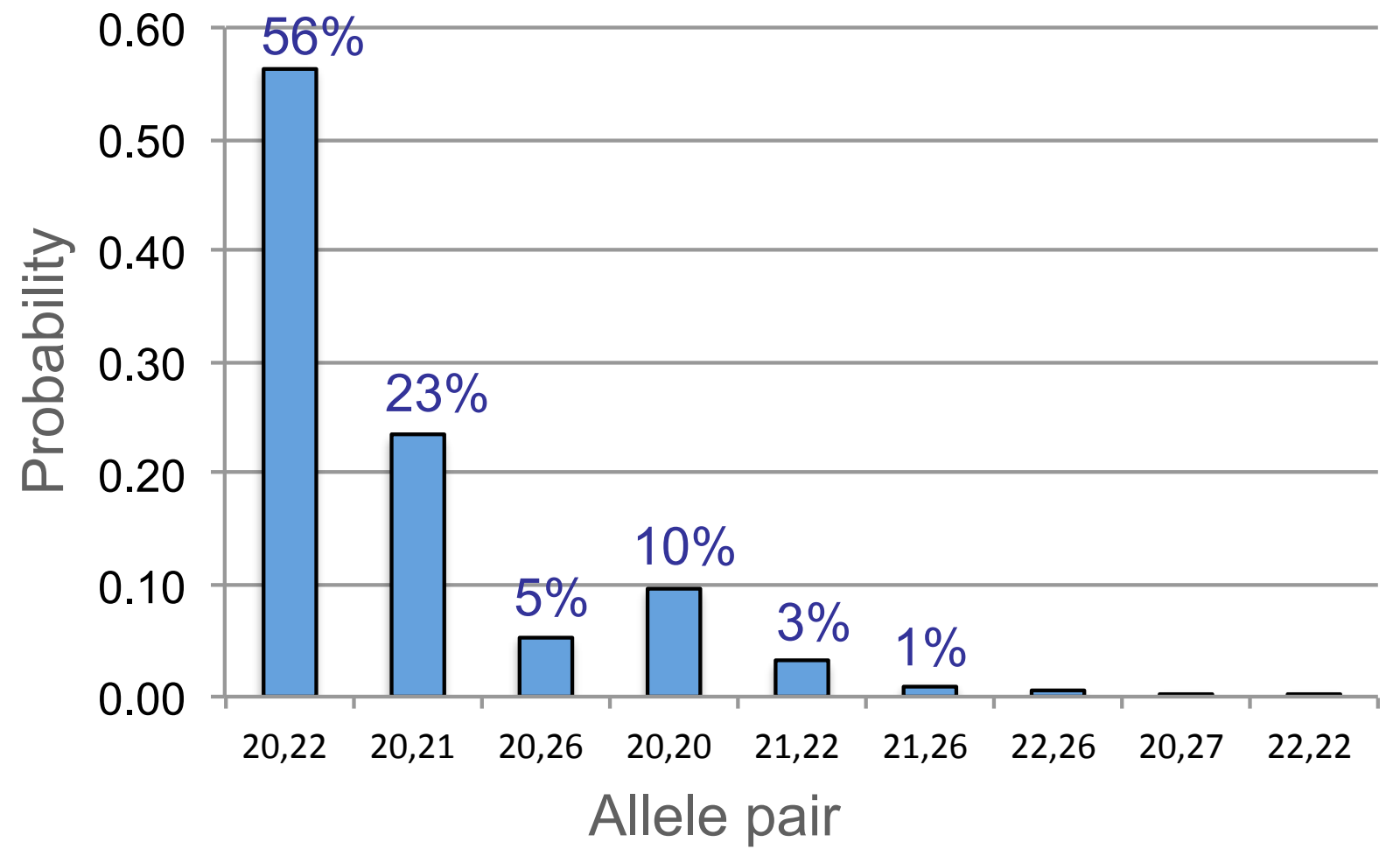
Bayes Theorem considers all hypotheses, and is written mathematically as:

$$posterior_i = \frac{prior_i \times likelihood_i}{\sum_j prior_j \times likelihood_j}$$

3

Allele Pair	Prior	Likelihood	Prior*Like	Posterior
20, 22	0.0543	0.1474	0.008004	0.5636
20, 21	0.0461	0.0722	0.003328	0.2344
20, 26	0.0058	0.1309	0.000759	0.0535
20, 20	0.0156	0.0882	0.001376	0.0969
21, 22	0.0800	0.0056	0.000448	0.0315
21, 26	0.0085	0.0176	0.000150	0.0105
22, 26	0.0100	0.0077	0.000077	0.0054
20, 27	0.0008	0.0142	0.000011	0.0008
22, 22	0.0471	0.0010	0.000047	0.0033
	0.2682	*0.4848*	*0.014200*	*1.0000*

Figure 7. Posterior genotype probability (blue).



MATCH INFORMATION

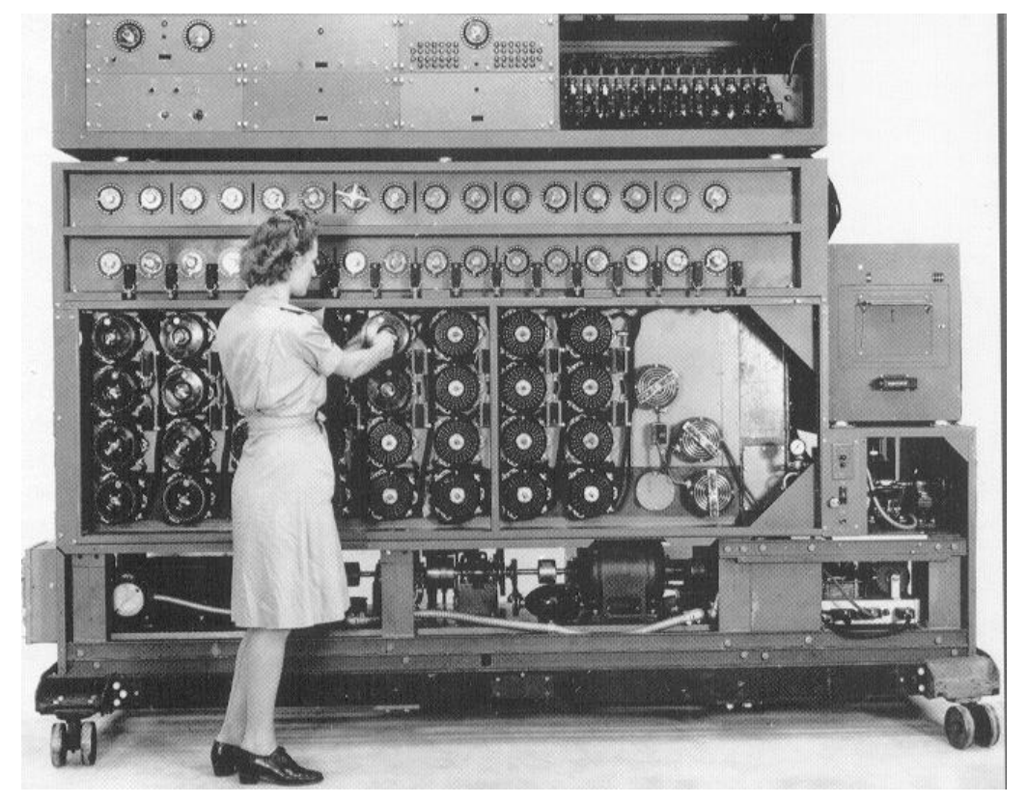


Figure 8. Applying the likelihood ratio for the World War II Bombe computer to crack the German Enigma code.

The *likelihood ratio* (LR) is a simple form of Bayes Theorem that has only one hypothesis (Fig 8). In forensic science, this is the identification hypothesis that a person contributed their DNA to biological evidence.

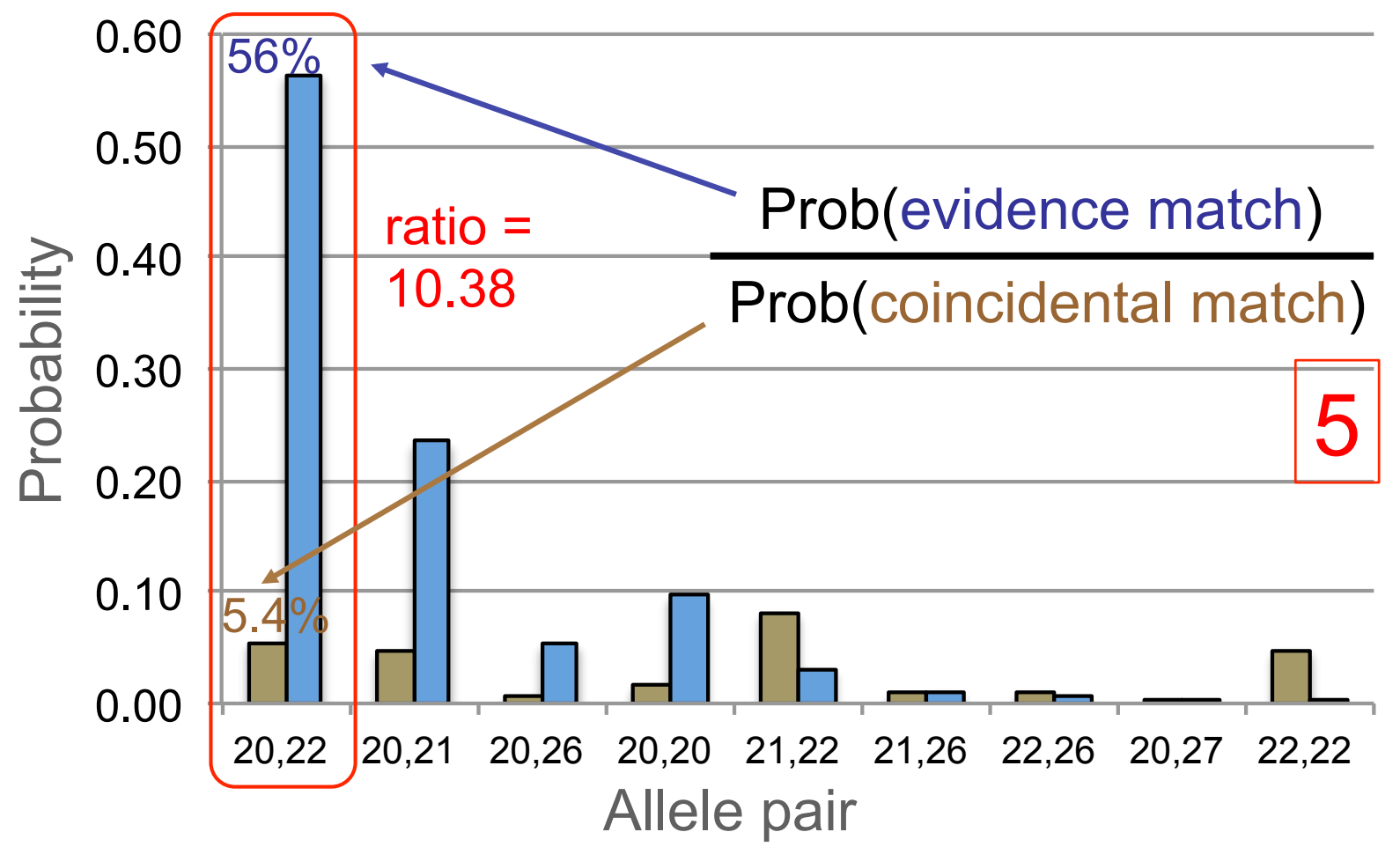
The LR can be written mathematically in different ways. Expressed as information change, the LR gives the hypothesis probability after having seen data, relative to before. In DNA identification, this is simply the ratio of posterior genotype probability to prior population probability (Fig 9).

The logarithm of the LR is a standard additive measure of information that is used in many fields. Thus the log(LR) is a natural way to quantify by how much STR data increases or decreases belief in a person's having contributed their DNA to biological evidence.

4

Allele Pair	Prior	Likelihood	Prior*Like	Posterior	Post/Prior
20, 22	0.0543	0.1474	0.008004	0.5636	10.3800
20, 21	0.0461	0.0722	0.003328	0.2344	5.0843
20, 26	0.0058	0.1309	0.000759	0.0535	9.2180
20, 20	0.0156	0.0882	0.001376	0.0969	6.2111
21, 22	0.0800	0.0056	0.000448	0.0315	0.3944
21, 26	0.0085	0.0176	0.000150	0.0105	1.2394
22, 26	0.0100	0.0077	0.000077	0.0054	0.5422
20, 27	0.0008	0.0142	0.000011	0.0008	1.0000
22, 22	0.0471	0.0010	0.000047	0.0033	0.0704
	0.2682	*0.4848*	*0.014200*	*1.0000*	

Figure 9. Likelihood ratio (blue to brown).



ADVICE FOR BEGINNERS

1 **Use all the data**. Genotypes are inferred from STR pattern data. By Bayes Theorem, all the data must be considered. This means that peak heights must be used, and thresholds should not be applied.

2 **Consider all genotype hypotheses**. The probability of an allele pair depends on all the other allele pairs. By Bayes Theorem, all allele pair candidates must be considered. This means that all alleles and their combinations must be considered, whether or not they have tall peak heights.

3 **Don't confuse likelihood with probability**. Probability quantifies belief, and is something that people can understand. Likelihood is a mathematical tool that aids in calculations, but is often unintuitive. Bayes Theorem uses likelihood to describe how data observations update belief.

4 **Use simple ratios, not complex formulas**. There are many ways to compute a LR for a match to a suspect. Bayes Theorem tells us that the LR is basically the ratio of the posterior to prior genotype probabilities, assessed at the suspect's genotype. But Bayes can also produce alternative formulas which are unnecessarily complex and unintuitive.

5 **Make meaningful comparisons**. DNA match statistics depend on the reference population, which specifies the prior genotype probability. By Bayes Theorem, changing the population can drastically alter the LR denominator. When comparing methods, be sure that the same data and reference populations are used.

REFERENCES

Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE. 2009;4(12):e8327.

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

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci. 2011;56(6):1430-47.

Perlin MW. DNA done right. FHC Experts for Law: Experts Forum Newsletter. 2013 Summer: 4-7.

Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. J Forensic Sci. 2013;58(6):DOI 10.1111/556-4029.12223