Forensic Genetics

Summer Institute in Statistical Genetics, July 20-22, 2016

Bruce Weir: bsweir@uw.edu

Contents

Sources of Data

Phenotype	Mendel’s peas
	Blood groups
DNA	Restriction sites, RFLPs
	Length variants, VNTRs, STRs
	SNPs
	Nucleotide sequences

Mendel's Data

Dominant Form		Recessive Form	
Seed characters			
5474	Round	1850	Wrinkled
6022	Yellow	2001	Green
Plant characters			
705	Grey-brown	224	White
882	Simply inflated	299	Constricted
428	Green	152	Yellow
651	Axial	207	Terminal
787	Long	277	Short

ABO System

Human ABO blood groups discovered in 1900. ABO gene on human chromosome 9 has 3 alleles: A, B, O. Six genotypes but only four phenotypes (blood groups):

Genotypes	Phenotype
AA, AO	A
BB, BO	B
AB	AB
OO	O

Charlie Chaplin and ABO Testing

Relationship	Person	Blood Group	Genotype
Mother	Joan Berry	A	AA or AO
Child	Carol Ann Berry	B	BB or BO
Alleged Father	Charles Chaplin	O	OO

The obligate paternal allele was B, so the true father must have been of blood group B or AB.

Berry v. Chaplin, 74 Cal. App. 2d 652

Electrophoretic Detection

Charge differences among alleles ("allozymes") of soluble proteins lead to separation on electrophoretic gels. Protein loaded at one end of a slab gel and an electric current is passed through the gel. Allozymes migrate according to their net charge: separation of alleles depends on how far they migrate in a given amount of time.

This techniques was the first to allow large-scale collection of genetic marker data. The data in this case reflected variation in the amino acid sequences of soluble proteins.

Alec Jeffreys

For forensic applications, the work of Alec Jeffreys with on Restriction Fragment Length Polymorphisms (RFLPs) or Variable Number of Tandem Repeats (VNTRs) also used electrophoresis. Different alleles now represented different numbers of repeat units and therefore different length molecules. Smaller molecules move faster through a gel and so move further in a given amount of time.

Initial work was on mini-satellites, where repeat unit lengths were in the tens of bases and fragment lengths were in thousands of bases. Jeffrey's multi-locus probes detected regions from several pats of the genome and resulted in many detectable fragments per individual. This gave high discrimination but difficulty in assigning numerical strength to matching profiles.

Jeffreys et al. 1985. Nature 316:76-79 and 317: 818-819.

Single-locus Probes

Next development for gel-electrophoresis used probes for single mini-satellites. Only two fragments were detected per individual, but there was difficulty in determining when two profiles matched.

The technology also required “large” amounts of DNA and was not suitable for degraded samples.

PCR-based STR Markers

The ability to increase the amount of DNA in a sample by the Polymerase Chain Reaction (PCR) was of substantial benefit to forensic science. The typing technology changed to the use of capillary tube electrophoresis, where the time taken by a DNA molecule to pass a fixed point was measured and used to infer the number of repeat units in an allele.

A good source is “Following multiplex PCR amplification, DNA samples containing the length-variant STR alleles are typically separated by capillary electrophoresis and genotyped by comparison to an allelic ladder supplied with a commercial kit. ”
Butler JM. Short tandem repeat typing technologies used in human identity testing. BioTechniques 43:Sii-Sv (October 2007) doi 10.2144/000112582

STR markers: CTT set

(http://www.cstl.nist.gov/biotech/strbase/seq_info.htm)

			Usual No.
Locus	Structure	Chromosome	of repeats
CSF1PO	$[\text{AGAT}]_n$	5q	6–16
TPOX	$[\text{AATG}]_n$	2p	5–14
TH01*	$[\text{AATG}]_n$	11p	3–14

* "9.3" is $[\text{AATG}]_6\text{ATG}[\text{AATG}]_3$

Length variants detected by capillary electrophoresis.

“CTT” Data - Forensic Frequency Database

CSF1P0		TPOX		TH01	
11	12	8	11	7	8
11	13	8	8	6	7
11	12	8	11	6	7
10	12	8	8	6	9
11	12	8	12	9	9.3
10	12	9	11	6	7
10	13	8	11	6	6
11	12	8	8	6	9.3
9	10	8	9	7	9.3
11	12	8	8	6	8
11	13	8	11	7	9
11	12	8	11	6	9.3
10	11	8	8	7	9.3
10	10	8	11	7	9.3
9	10	8	8	6	9.3
11	12	9	11	9	9.3
9	11	9	11	9	9.3
11	12	8	8	6	7
10	10	9	11	6	9.3
10	13	8	8	8	9.3

Sequencing of STR Alleles

"STR typing in forensic genetics has been performed traditionally using capillary electrophoresis (CE). Massively parallel sequencing (MPS) has been considered a viable technology in recent years allowing high-throughput coverage at a relatively affordable price. Some of the CE-based limitations may be overcome with the application of MPS ... generate reliable STR profiles at a sensitivity level that competes with current widely used CE-based method."

Zeng XP, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16:38-47.

Single Nucleotide Polymorphisms (SNPs)

“Single nucleotide polymorphisms (SNPs) are the most frequently occurring genetic variation in the human genome, with the total number of SNPs reported in public SNP databases currently exceeding 9 million. SNPs are important markers in many studies that link sequence variations to phenotypic changes; such studies are expected to advance the understanding of human physiology and elucidate the molecular bases of diseases. For this reason, over the past several years a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for SNP analysis, yielding a large number of distinct approaches. ”

Kim S. Misra A. 2007. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289-320.

Phase 3 1000Genomes Data

- 84.4 million variants

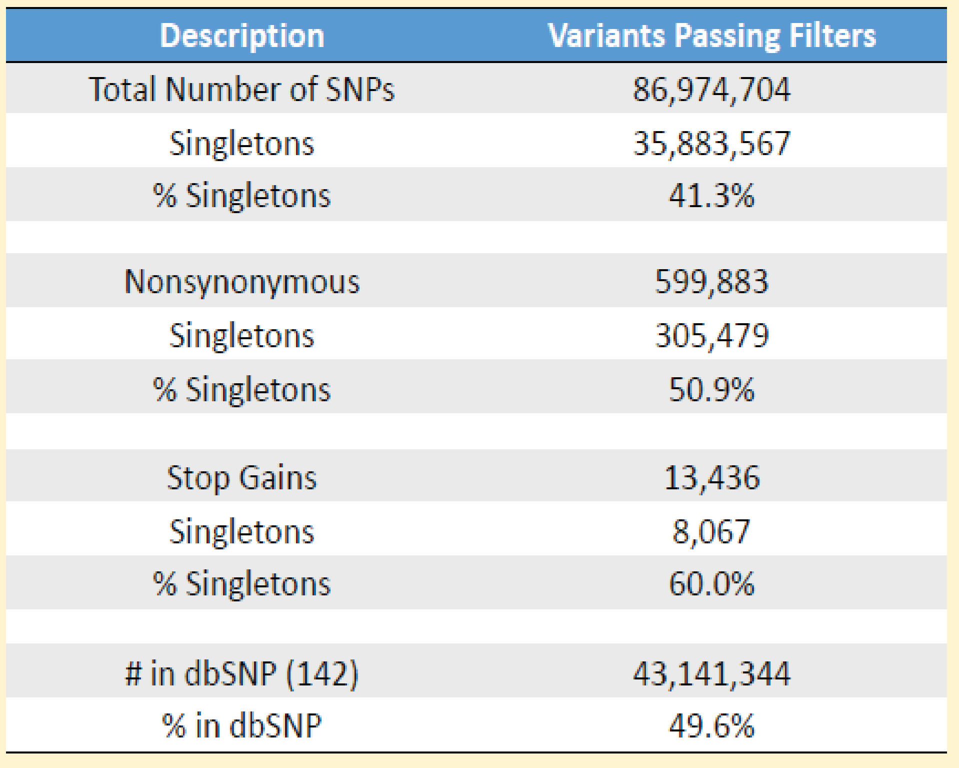
- 2504 individuals

- 26 populations

www.1000Genomes.org

Whole-genome Sequence Studies

One current study is the NHLBI Trans-Omics for Precision Medicine (TOPMed) project. www.nhlbiwgs.org
In the first data freeze of Phase 1 of this study:



Description	Variants Passing Filters
Total Number of SNPs	86,974,704
Singletons	35,883,567
% Singletons	41.3%
Nonsynonymous	599,883
Singletons	305,479
% Singletons	50.9%
Stop Gains	13,436
Singletons	8,067
% Singletons	60.0%
# in dbSNP (142)	43,141,344
% in dbSNP	49.6%

Abecasis et al. 2016. ASHG Poster

Probability Theory

We wish to attach probabilities to different kinds of events (or hypotheses or propositions):

- Event A: the next card is an Ace.
- Event R: it will rain tomorrow.
- Event C: the suspect left the crime stain.

Probabilities

Assign probabilities to events: Pr(A) or p_A or even p means "the probability that event A is true." All probabilities are conditional, so should write Pr$(A|E)$ for "the probability that A is true given that E is known."

No matter how probabilities are defined, they need to follow some mathematical laws in order to lead to consistent theories.

First Law of Probability

$$\begin{aligned} 0 \leq \ \mathrm{Pr}(A|E) & \ \leq 1 \\ \mathrm{Pr}(A|A) & = 1 \end{aligned}$$

If A is the event that a die shows an even face (2, 4, or 6), what is E? What is $\mathrm{Pr}(A|E)$?

Second Law of Probability

If A, B are mutually exclusive given E

$$\begin{aligned} \text{Pr}(A \text{ or } B|E) &= \text{Pr}(A|E) + \text{Pr}(B|E) \\ \text{so } \text{Pr}(\bar{A}|E) &= 1 - \text{Pr}(A|E) \end{aligned}$$

($\bar{A}$ means not-A).

If A is the event that a die shows an even face, and B is the event that the die shows a 1, verify the Second Law.

Third Law of Probability

$$\text{Pr}(A \text{ and } B|E) \;=\; \text{Pr}(A|B,E) \times \text{Pr}(B|E)$$

If A is event that die shows an even face, and B is the event that the die shows a 1, verify the Third Law.

Independent Events

Events A and B are independent if knowledge of one does not affect probability of the other:

$$\begin{aligned} \mathrm{Pr}(A|B) &= \mathrm{Pr}(A) \\ \mathrm{Pr}(B|A) &= \mathrm{Pr}(B) \end{aligned}$$

Therefore, for independent events

$$\mathrm{Pr}(A \text{ and } B) = \mathrm{Pr}(A)\,\mathrm{Pr}(B)$$

This may be written as

$$\mathrm{Pr}(AB) = \mathrm{Pr}(A)\,\mathrm{Pr}(B)$$

Law of Total Probability

Because B and $\bar{B}$ are mutually exclusive and exhaustive:

$$\Pr(A) \;=\; \Pr(A|B)\Pr(B) + \Pr(A|\bar{B})\Pr(\bar{B})$$

If A is the event that die shows a 3, B is the event that the die shows an even face, and $\bar{B}$ the event that the die shows an odd face, verify the Law of Total Probability.

IF B_1, B_2, B_3 are mutually exclusive and exhaustive:

$$\begin{aligned}\Pr(A) \;=\; & \Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2) \\ & + \Pr(A|B_3)\Pr(B_3)\end{aligned}$$

Odds

The odds $O(A)$ of an event A are the probability of the event being true divided by the probability of the event not being true:

$$O(A) = \frac{\mathrm{Pr}(A)}{\mathrm{Pr}(\bar{A})}$$

This can be rearranged to give

$$\mathrm{Pr}(A) = \frac{O(A)}{1 + O(A)}$$

Odds of 10 to 1 are equivalent to a probability of 10/11.

Bayes' Theorem

The third law of probability can be used twice to reverse the order of conditioning:

$$
\begin{aligned}
\text{Pr}(E|A) &= \frac{\text{Pr}(E \text{ and } A)}{\text{Pr}(A)} \\
&= \frac{\text{Pr}(A|E)\,\text{Pr}(E)}{\text{Pr}(A)}
\end{aligned}
$$

Odds Form of Bayes' Theorem

From the third law of probability

$$\begin{aligned}
\Pr(E|A) &= \Pr(A|E)\Pr(E)/\Pr(A) \\
\Pr(\bar{E}|A) &= \Pr(A|\bar{E})\Pr(\bar{E})/\Pr(A)
\end{aligned}$$

Taking the ratio of these two equations:

$$\frac{\Pr(E|A)}{\Pr(\bar{E}|A)} = \frac{\Pr(A|E)}{\Pr(A|\bar{E})} \times \frac{\Pr(E)}{\Pr(\bar{E})}$$

Posterior odds $=$ likelihood ratio $\times$ prior odds.

AIDS Example

Suppose the event E of AIDS occurs 1 in 10,000 people chosen at random.

Suppose a test procedure has two outcomes: A (positive) and B (negative). The probability of a positive result is 0.99 if the person has AIDS, and 0.05 if the person does not have AIDS. What is the probability that a person has AIDS if she tests positive?

AIDS Example

The problem is to determine Pr$(E|A)$ when Pr$(A|E)$ is known. This requires Bayes' theorem, and the term Pr(A) follows from the Law of Total Probability.

$$\begin{aligned}
\text{Pr}(E) &= \\
\text{Pr}(\bar{E}) &= \\
\text{Pr}(A|E) &= \\
\text{Pr}(A|\bar{E}) &= \\
\text{Pr}(A) &= \\
\text{Pr}(E|A) &=
\end{aligned}$$

Birthday Problem

Forensic scientists in Arizona looked at the 65,493 profiles in the Arizona database and reported that two profiles matched at 9 loci out of 13. They reported a "match probability" for those 9 loci of 1 in 754 million. Are the numbers 65,493 and 754 million inconsistent?
(Troyer et al., 2001. Proc Promega 12th Int Symp Human Identification.)

To begin to answer this question suppose that every possible profile has the same profile probability P and that there are N profiles in a database (or in a population). The probability of at least one pair of matching profiles in the database is one minus the probability of no matches.

Birthday Problem

Choose profile 1. The probability that profile 2 does not match profile 1 is $(1-P)$. The probability that profile 3 does not match profiles 1 or 2 is $(1-2P)$, etc. So, the probability P_M of at least one matching pair is

$$\begin{aligned} P_M &= 1 - \{1(1-P)(1-2P)\cdots[1-(N-1)P]\} \\ &\approx 1 - \prod_{i=0}^{N-1} e^{-iP} \approx 1 - e^{-N^2P/2} \end{aligned}$$

If $P = 1/365$ and $N = 23$, then $P_M = 0.51$. So, approximately, in a room of 23 people there is greater than a 50% probability that two people have the same birthday.

Birthday Problem

If $P = 1/(754 \text{ million})$ and $N = 65,493$, then $P_M = 0.98$ so it is highly probable there would be a match. There are other issues, having to do with the four non-matching loci, and the possible presence of relatives in the database.

If $P = 10^{-16}$ and $N = 300$ million, then $P_M =$ is essentially 1. It is almost certain that two people in the US have the same rare DNA profile.

Statistics

- Probability: For a given model, what do we expect to see?

- Statistics: For some given data, what can we say about the model?

- Example: A marker has an allele A with frequency p_A.

  - Probability question: If $p_A = 0.5$, and if alleles are independent, what is the probability of AA?

  - Statistics question: If a sample of 100 individuals has 23 AA's, 48 Aa's and 29 aa's, what is an estimate of p_A?

Binomial distribution

Imagine tossing a coin n times, when every toss has the same chance p of giving a head:

The probability of x heads in a row is

$$p \times p \times \ldots \times p \;=\; p^x$$

The probability of $n - x$ tails in a row is

$$(1-p) \times (1-p) \times \ldots \times (1-p) \;=\; (1-p)^{n-x}$$

The number of ways of ordering x heads and $n - x$ tails among n outcomes is $n!/[x!(n-x)!]$.

Binomial distribution

Combining the probabilities of x successive heads, $n-x$ successive trials, and the number of ways of ordering x heads and $n-x$ tails: the binomial probability of x successes (heads) in n trials (tosses) is

$$\mathrm{Pr}(x|p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Binomial distribution

The probabilities of x heads in $n = 4$ tosses of a coin when the chance of a head is 1/2 at each toss:

No. heads x	Probability Pr$(x \vert p)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

Note that $0! = 1$ and $p^0 = 1$.

Binomial distribution

Find the binomial probabilities, for a sample of size $n = 4$ alleles, when the chance that each allele is of type A is 1/10.

No. A's	Probability
0	
1	
2	
3	
4	

TRANSFER EVIDENCE

Relevant Evidence

Rule 401 of the US Federal Rules of Evidence:

“Relevant evidence” means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.

Single Crime Scene Stain

Suppose a blood stain is found at a crime scene, and it must have come from the offender. A suspect is identified and provides a blood sample. The crime scene sample and the suspect have the same (DNA) “type.”

The prosecution subsequently puts to the court the proposition (or hypothesis or explanation):

H_p: The suspect left the crime stain.

The symbol H_p is just to assist in the formal analysis. It need not be given in court.

Transfer Evidence Notation

G_S, G_C are the DNA types for suspect and crime sample. $G_S = G_C$. I is non-DNA evidence.

Before the DNA typing, probability of H_p is conditioned on I.

After the typing, probability of H_p is conditioned on G_S, G_C, I.

Updating Uncertainty

Method of updating uncertainty, or changing $\Pr(H_p|I)$ to $\Pr(H_p|G_S, G_C, I$
uses Bayes' theorem:

$$\begin{aligned}\Pr(H_p|G_S, G_C, I) &= \frac{\Pr(H_p, G_S, G_C|I)}{\Pr(G_S, G_C|I)} \\ &= \frac{\Pr(G_S, G_C|H_p, I)\Pr(H_p|I)}{\Pr(G_S, G_C|I)}\end{aligned}$$

We can't evaluate $\Pr(G_S, G_C|I)$ without additional information, and we don't know $\Pr(H_p|I)$.

Can proceed by introducing alternative to H_p.

First Principle of Evidence Interpretation

To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.

The simplest alternative explanation for a single stain is:

H_d: Some other person left the crime stain.

Updating Odds

From the odds form of Bayes' theorem:

$$\frac{\text{Pr}(H_p|G_S, G_C, I)}{\text{Pr}(H_d|G_S, G_C, I)} = \frac{\text{Pr}(G_S, G_C|H_p, I)}{\text{Pr}(G_S, G_C|H_d, I)} \times \frac{\text{Pr}(H_p|I)}{\text{Pr}(H_d|I)}$$

i.e. Posterior odds = LR × Prior odds

where

$$\text{LR} = \frac{\text{Pr}(G_S, G_C|H_p, I)}{\text{Pr}(G_S, G_C|H_d, I)}$$

Questions for a Court to Consider

The trier of fact needs to address questions of the kind

- What is the probability that the prosecution proposition is true given the evidence,
  $\Pr(H_p|G_C, G_S, I)$?

- What is the probability that the defense proposition is true given the evidence,
  $\Pr(H_d|G_C, G_S, I)$?

Questions for Forensic Scientist to Consider

The forensic scientist must address different questions:

- What is the probability of the DNA evidence if the prosecution proposition is true,
  $\Pr(G_C, G_S | H_p, I)$?

- What is the probability of the DNA evidence if the defense proposition is true,
  $\Pr(G_C, G_S | H_d, I)$?

Important to articulate H_p, H_d. Also important not to confuse the difference between these two sets of questions.

Second Principle of Evidence Interpretation

Evidence interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition.'

This question is answered for alternative explanations, and the ratio of the probabilities presented. It is not necessary to use the words "likelihood ratio". Use phrases such as:
'The probability that the crime scene DNA type is the same as the suspect's DNA type is one million times higher if the suspect left the crime sample than if someone else left the sample.'

Third Principle of Evidence Interpretation

Evidence interpretation is conditioned not only on the alternative propositions, but also on the framework of circumstances within which they are to be evaluated.

The circumstances may simply be the population to which the offender belongs so that probabilities can be calculated. Forensic scientists must be clear in court about the nature of the non-DNA evidence I, as it appeared to them when they made their assessment. If the court has a different view then the scientist must review the interpretation of the evidence.

Example

“In the analysis of the results I carried out I considered two alternatives: either that the blood samples originated from Pengelly or that the ... blood was from another individual. I find that the results I obtained were at least 12,450 times more likely to have occurred if the blood had originated from Pengelly than if it had originated from someone else.”

Example

Question: “Can you express that in another way?”

Answer: “It could also be said that 1 in 12,450 people would have the same profile . . . and that Pengelly was included in that number . . . very strongly suggests the premise that the two blood stains examined came from Pengelly.”

[Testimony of M. Lawton in R. v Pengelly 1 NZLR 545 (CA), quoted by Robertson & Vignaux, “Interpreting Evidence”, Wiley 1995.]

Likelihood Ratio

$$\text{LR} = \frac{\Pr(G_C, G_S | H_p, I)}{\Pr(G_C, G_S | H_d, I)}$$

Apply laws of probability to change this into

$$\text{LR} = \frac{\Pr(G_C | G_S, H_p, I) \Pr(G_S | H_p, I)}{\Pr(G_C | G_S, H_d, I) \Pr(G_S | H_d, I)}$$

Likelihood Ratio

Whether or not the suspect left the crime sample (i.e. whether or not H_p or H_d is true) provides no information about his genotype:

$$\text{Pr}(G_S|H_p, I) = \text{Pr}(G_S|H_d, I) \quad = \quad \text{Pr}(G_S|I)$$

so that

$$\text{LR} \quad = \quad \frac{\text{Pr}(G_C|G_S, H_p, I)}{\text{Pr}(G_C|G_S, H_d, I)}$$

Likelihood Ratio

$$\text{LR} = \frac{\Pr(G_C|G_S,H_p,I)}{\Pr(G_C|G_S,H_d,I)}$$

When $G_C = G_S$, and when they are for the same person (H_p is true):

$$\Pr(G_C|G_S,H_p,I) = 1$$

so the likelihood ratio becomes

$$\text{LR} = \frac{1}{\Pr(G_C|G_S,H_d,I)}$$

This is the reciprocal of the probability of the *match probability*, the probability of profile G_C, conditioned on having seen profile G_S in a different person (i.e. H_d) and on I.

Likelihood Ratio

$$\text{LR} = \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

The next step depends on the circumstances I. If these say that knowledge of the suspect's type does not affect our uncertainty about the offender's type when they are different people (i.e. when H_d is true):

$$\Pr(G_C|G_S, H_d, I) = \Pr(G_C|H_d, I)$$

and then likelihood ratio becomes

$$\text{LR} = \frac{1}{\Pr(G_C|H_d, I)}$$

The LR is now the reciprocal of the *profile probability* of profile G_C.

Profile and Match Probabilities

Dropping mention of the other information I, the quantity Pr(G_C) is the probability that a person randomly chosen from a population will have profile type G_C. This profile probability usually very small and, although it is interesting, it is not the most relevant quantity.

Of relevance is the match probability, the probability of seeing the profile in a randomly chosen person after we have already seen that profile in a typed person (the suspect). The match probability is bigger than the profile probability. Having seen a profile once there is an increased chance we will see it again. This is the genetic essence of DNA evidence.

Likelihood Ratio

The estimated probability in the denominator of LR is determined on the basis of judgment, informed by I. Therefore the nature of I (as it appeared to the forensic scientist at the time of analysis) must be explained in court along with the value of LR. If the court has a different view of I, then the scientist will need to review the interpretation of the DNA evidence.

Random Samples

The circumstances I may define a population or racial group. The probability is estimated on the basis of a sample from that population. If the probability is written as P, then the likelihood ratio is $1/P$. If P is estimated to be 1 in a million, then LR is 1 million.

When we talk about DNA types, by "selecting a man at random" we mean choosing him in such a way as to be as uncertain as possible about his DNA type.

Convenience Samples

The problem with a formal approach is that of defining the population: if we mean the population of a town, do we mean *every* person in the town at the time the crime was committed? Do we mean some particular area of the town? One sex? Some age range?

It seems satisfactory instead to use a convenience sample, i.e. a set of people from whom it is easy to collect biological material in order to determine their DNA profiles. These people are not a random sample of people, but they have not been selected on the basis of their DNA profiles.

Meaning of Likelihood Ratios

There is a personal element to interpreting DNA evidence, and there is no "right" value for the LR. (There is a right answer to the question of whether the suspect left the crime stain, but that is not for the forensic scientist to decide.)

The denominator for LR is conditioned on the stain coming from an unknown person, and "unknown" may be hard to define. A relative? Someone in that town? Someone in the same racial group? (What is a race?)

Errors and Fallacies

Once the numerical strength of the evidence has been calculated, it is important that it be presented in a way that does not distort its meaning.

There are a series of common fallacies that can be avoided by careful application of the Principles of Evidence Interpretation.

Transposed conditional

Correct: *The evidence is 1000 times more likely if the suspect left the crime stain than if some unknown person left it.*

Incorrect: *It is 1000 times more likely that the suspect left the crime stain than some unknown person.*

The second statement is true only if the prior odds are 1.

Prosecution Fallacy

From the O.J. Simpson trials:

“You testify that there is . . . a 1 in 71 chance that a pair of contributors at random could have left the stain.” (Defense attorney at transcript p. 33,242.)

“The chances are at least 1-in-170 million that anybody else’s DNA besides Simpson’s could be contained in a blood drop found near the bodies of Nicole Brown Simpson and Ronald Goldman, testified Robin Cotton, director of Cellmark Diagnostics in Maryland.” (Associated Press, 11/14/96)

Defense Fallacy

The matching DNA profile has a probability of 1 in 100,000. The crime was committed in a city of 1,000,000 people.

Correct: Therefore 10 people in the city are expected to have that profile.

Incorrect: Therefore the suspect (who has the profile) has a probability of 1 in 10 of being guilty.

Defense Fallacy

The fallacy is to assign equal (prior) probabilities to all 10 people who are expected to have the profile. Also note that the actual number of people in the city with the profile could be any number from 0 to 1,000,000. Expected numbers are not actual numbers.

Uniqueness Fallacy

The matching DNA profile has a probability of 1 in 1,000,000. The crime was committed in a city of 1,000,000 people.

Correct: Therefore 1 person in the city is expected to have that profile.

Incorrect: Therefore the suspect (who has the profile) is the guilty person.

Uniqueness Fallacy

The fallacy is not to recognize that the actual number of people in the city with the profile could be any number from 0 to 1,000,000.

Expected numbers are not actual numbers.

Database Fallacy

Probabilities needed in LR are estimated on the basis of a sample from a population. Ideally this sample is drawn from the population defined by H_d and I. This is not practical.

If H_d is true, the racial background of the suspect does not define the population to be sampled. Do not need a database of (exactly) the same ethnicity as the suspect.

Aitken and Taroni, Science and Justice 1998; 38:165–177

The laboratory results and the expert's testimony were described in 12 cases.

1. Ross v. State of Indiana (Indiana Court of Appeal, May 13, 1996). The DNA expert knew that the frequency of the DNA profile found in the vaginal swab sample and in the suspect's blood sample was 1 in 80,000. He said that Ross was the source of the seminal fluid.

2. State of Washington v. Gentry (125 Wash. 2d 570, 888 P.2d 1105 (1995)]. The matching DNA profile had a frequency of 0.18%. The expert said that the percentage of the population from which the blood found on the defendant's shoelaces could have originated is 0.18%.

Aitken and Taroni, Science and Justice 1998; 38:165–177

3. R v. Deen (Court of Appeal, Criminal Division, December 21, 1993). The matching DNA profile had a frequency of 1 in 3 million. The Prosecutor and Expert had this exchange: Q (Prosecutor) “So the likelihood of this being any other man but Andrew Deen is one in 3 million?” A (Expert): ”In 3 million, yes.” Q: “On the figure which you have established according to your research, the probability of it being anybody else being one in 3 million what is your conclusion?” Expert: “My conclusion is that the semen originated from Andrew Deen.” Q. Are you sure of that?” A. “Yes.”

Aitken and Taroni, Science and Justice 1998; 38:165–177

4. U.S. v. Jakobetz [955 F. 2d 786 (2nd Cir. 1992)]. In that case the FBI expert knew that the frequency of the matching DNA profile in the population is 1 in 300 million. The expert testified that the DNA profiles from the two samples constituted a match and calculated there was one chance in 300 million that the DNA from the semen sample could have come from someone in the Caucasian sample other than Jakobetz.

Aitken and Taroni, Science and Justice 1998; 38:165–177

5. Gordon (Court of Appeal, November 22, 1993, April 22, May 26, 1994). The DNA profile had a frequency of 1 in 10,500,000. The expert agreed that there was a visual match between the critical samples and the appellant's sample which showed a likelihood that the appellant was the rapist in each case.

6. Lonsdale (Court of Appeal, March 9, 16 1995.) The DNA profile frequency was 1 in 1,000,000. The expert said that the chances of a sample from another Afro-Caribbean (the relevant population) matching the crime sample were one in a million.

Aitken and Taroni, Science and Justice 1998; 38:165–177

7. U.S. v. Bonds [12 F. 3d 540 (6th Cir. 1993)] The DNA profile frequency was 1 in 270,000. The FBI calculated a probability of 1 in 270,000 that an unrelated individual selected randomly from the Caucasian population (the relevant population) would have a DNA profile matching that of Bonds.

8. U.S. v. Martinez [13 F.3d 1191 (8th Cir. 1993)] The FBI expert knew that the population frequency of the matching DNA profile is 1 in 2,600. The expert testified that only 1 in 2,600 American Indians (relevant population) would be expected to produce the identical genetic characteristics as Martinez.

Aitken and Taroni, Science and Justice 1998; 38:165–177

9. Arizona v. Johnson [905 P. 2d 1002; 192 Ariz. Rep. 19 91995)] The expert knew that the matching DNA profile has a frequency of 1 in 312 million. The expert testified that the victim's shirt was examined and found to contain human blood and semen. Testing performed showed that DNA extracted from these stains matched Johnson's blood at five different chromosome locations or loci. The expert testified that the possibility of a random match - two unrelated individuals having the same DNA pattern across five loci - was 1 in 312 million.

Aitken and Taroni, Science and Justice 1998; 38:165–177

10. Ross v. State [B14-90-00659 (Tex. App. Feb. 13, 1992)] The DNA profile frequency was 1 in 209,100,000. The expert said that he has a database of blood samples from all over the country and he asks the question "How many people would we have to look at before we saw another person like this?" The answer is 209,100,000.

11. Harrison v. Indiana [Supreme Court of Indiana (Jan. 4, 1995)] The DNA profile frequency was 7.4 in 100. The expert said that although 92.6% of all white males could be excluded as the source of the specimen, the defendant had not been excluded. She acknowledged that for 13,000 white men (the size of the city where the crime was committed) the specimen could have come from any 962 [7.4% of 13,000] of them. Hence, the suspect is one of 962 men who might have committed the crime.

Aitken and Taroni, Science and Justice 1998; 38:165–177

12. R v. Montella [1 NZLR High Court (1992) 63-68]. The DNA profile frequency was 8.06 $\times 10^{-5}$. The expert said "A DNA profiling examination of the samples strongly supports a contention that the semen stain on the underpants of the complainants came from the accused. It is said that the likelihood of obtaining such DNA profiling results is at least 12,400 times greater if the semen stain originated from the accused than from another individual."

Nevada v Troy Brown

In the early morning hours of January 29, 1994, Jane Doe was sexually assaulted in the bedroom of her trailer home at 1637 Pruett Street in Carlin. Jane Doe and her four-year-old sister were home alone while their mother, Pam, was drinking at a bar, and their step-father, Wayne, was working the night shift at his job. Troy was arrested, tried, and convicted for the crime.

Nevada v Troy Brown

At trial, Renee Romero testified that she had conducted a DNA test on stains found on Jane Doe's underwear. Romero explained in detail what DNA is and how it is tested. Romero testified that the DNA sample tested from Jane Doe's underwear matched Troy's and that only 1 in 3,000,000 people had the same DNA code as the one tested. Troy's counsel cross-examined Romero regarding how she conducted the tests, the amount of DNA required to run the tests, and the databases against which the DNA tests were compared to determine the statistical probability that others would have the same DNA code. However, Troy's counsel did not call his own expert DNA witness even though the court provided funds for such a witness.

Nevada Supreme Court, February 26, 1997

Appellant Troy Brown was tried and convicted of sexually assaulting Jane Doe, a nine-year-old girl. Troy was convicted of two counts of sexual assault of a child under fourteen years of age, and one count of child abuse by sexual abuse. He was acquitted of one count of attempted murder. Troy claims on appeal that (1) he was improperly denied bail; (2) the DNA evidence was improperly admitted because no evidentiary hearing was held; (3) sufficient evidence did not exist to support his conviction; (4) double jeopardy barred his convictions for both sexual assault and child abuse by sexual abuse; and (5) the district judge abused his discretion during the sentencing phase of the trial.

Nevada Supreme Court, February 26, 1997

We conclude that the district judge properly denied bail for Troy, that the DNA evidence was properly admitted at trial, and that sufficient evidence existed to support Troy's conviction. However, we conclude that Troy's conviction for both sexual assault and child abuse by sexual abuse violated the double jeopardy provision of the Constitution and that the conviction for child abuse must be vacated. Finally, we conclude that the district judge abused his discretion during the sentencing phase of the trial and the case must be remanded to the district court for a new sentencing hearing on the remaining sexual assault conviction.

US District Court, February 6, 2004

On February 6, 2004, Troy filed his federal petition for writ of habeas corpus pursuant to 28 U.S.C. 2254, arguing, inter alia, violations of due process and ineffective assistance of counsel. Judge Pro permitted Troy to expand the record, admitting, among other things, an uncontested report discrediting Romero's testimony by Dr. Laurence Mueller (the "Mueller Report"), a professor of Ecology and Evolutionary Biology at the University of California, Irvine.

US District Court, February 6, 2004

The district court granted Troy's petition. First, the district court concluded that, in light of the Mueller Report, Romero's testimony was unreliable. Absent that testimony, no rational trier of fact could conclude beyond a reasonable doubt that Troy was guilty of each and every element of the offenses with which he was charged. The district court also concluded that Troy's attorney's failure to diligently defend against Respondents' DNA testimony, as well as his failure to investigate the alibi of Henle, a potential suspect, amounted to ineffective assistance of counsel. Respondents [the State of Nevada] timely appealed.

US Court of Appeals, May 8, 2008

At trial, Respondents presented the testimony of DNA expert Renee Romero of the Washoe County Sheriff's Office Crime Lab. Romero testified that, among other things, there was a 99.99967 percent chance that Troy was the assailant.

At Petitioner Troy Brown's trial for sexual assault, the Warden and State's ("Respondents") deoxyribonucleic acid ("DNA") expert provided critical testimony that was later proved to be inaccurate and misleading. Respondents have conceded at least twice that, absent this faulty DNA testimony, there was not sufficient evidence to sustain Troy's conviction. In light of these extraordinary circumstances, we agree with District Judge Philip M. Pro's conclusions that Troy was denied due process, and we affirm the district court's grant of Troy's petition for writ of habeas corpus.

US Court of Appeals, May 8, 2008

Troy asserts that there was insufficient evidence to convict him. His argument rests on the admission of Romero's later discredited testimony regarding the DNA evidence, which was introduced without rebuttal at trial. Respondents have conceded that absent introduction of Romero's DNA evidence, the remaining evidence is insufficient to sustain Troy's conviction. Having reviewed the record ourselves, we affirm the district court's conclusion that, had Romero's inaccurate and unreliable testimony on the DNA evidence been excluded, there would have been insufficient evidence to convict Troy on each essential element of the offenses beyond a reasonable doubt. We further agree with the district court's conclusion that the Nevada Supreme Court's decision was both "contrary to" and an "unreasonable application of" established United States Supreme Court precedent.

US Court of Appeals, May 8, 2008

Here, Romero initially testified that Troy's DNA matched the DNA found in Jane's underwear, and that 1 in 3,000,000 people randomly selected from the population would also match the DNA found in Jane's underwear (random match probability). After the prosecutor pressed her to put this another way, Romero testified that there was a 99.99967 percent chance that the DNA found in Jane's underwear was from Troy's blood (source probability). This testimony was misleading, as it improperly conflated random match probability with source probability. In fact, the former testimony (1 in 3,000,000) is the probability of a match between an innocent person selected randomly from the population; this is not the same as the probability that Troy's DNA was the same as the DNA found in Jane's underwear, which would prove his guilt. Statistically, the probability of guilt given a DNA match is based on a complicated formula known as Bayes's Theorem, see id. at 170-71 n. 2, and the 1 in 3,000,000 probability described by Romero is but one of the factors in this formula.

US Court of Appeals, May 8, 2008

Because we affirm the district court's grant of Troy Brown's habeas petition on due process grounds, we need not reach his arguments regarding ineffective assistance of counsel. The district court's grant of Troy's petition for writ of habeas corpus and reversal of his conviction is AFFIRMED. Respondents shall retry Troy within 180 days or shall release him from custody.

US Supreme Court, January 11, 2010

The prosecutor's fallacy is the assumption that the random match probability is the same as the probability that the defendant was not the source of the DNA sample. In other words, if a juror is told the probability a member of the general population would share the same DNA is 1 in 10,000 (random match probability), and he takes that to mean there is only a 1 in 10,000 chance that someone other than the defendant is the source of the DNA found at the crime scene (source probability), then he has succumbed to the prosecutors fallacy. It is further error to equate source probability with probability of guilt, unless there is no explanation other than guilt for a person to be the source of crime-scene DNA. This faulty reasoning may result in an erroneous statement that, based on a random match probability of 1 in 10,000, there is a .01% chance the defendant is innocent or a 99.99% chance the defendant is guilty.

US Supreme Court, January 11, 2010

In sum, the two inaccuracies upon which this case turns are testimony equating random match probability with source probability, and an underestimate of the likelihood that one of Troys brothers would also match the DNA left at the scene.

We have stated before that “DNA testing can provide powerful new evidence unlike anything known before.”

The State acknowledges that Romero committed the prosecutor’s fallacy. Regardless, ample DNA and non-DNA evidence in the record adduced at trial supported the jury’s guilty verdict. Accordingly, the judgment of the Court of Appeals is reversed, and the case is remanded for further proceedings consistent with this opinion.

Hierarchy of Propositions

(Evett et al., 2002. Journal of Forensic Sciences 47:520–523.)

Third level: Offense level propositions:

The defendant raped the victim.

Some unknown person raped the victim.

Second level: Activity level propositions:

The defendant smashed the window.

The defendant has never been at the scene.

First level: Source level propositions:

The glass on the defendant's clothing came from the broken window.

The glass on the defendant's clothing is from some other source.

Meaning of Frequencies

What is meant by “the frequency of the matching profile is 1 in 57 billion”?

It is an estimated probability, obtained by multiplying together the allele frequencies, and refers to an infinite random mating population. It has nothing to do with the size of the world’s population.

Meaning of Frequencies

With 13 STR loci having (at least) 10 alleles each, there are $55^{13} = 4.2 \times 10^{22}$ possible genotypes, even though there are only 6 billion people. The total world population is itself a sample from all possible genotypes. Almost all the possible genotypes are not in the present population, and have expected frequencies that are very small: e.g. if all 26 alleles were independent, and had frequency of 0.1, we could quote an estimated frequency of 8.2×10^{-23} for a completely heterozygous. We don't *expect* that anyone living will have that profile – but of course we know that someone does.

Meaning of Frequencies

The question is really whether we would see the profile in two people, given that we have already seen it in one person. This conditional probability may be very low, but has nothing to do with the size of the population.

ALLELIC INDEPENDENCE

Testing for Allelic Independence

What is the probability a person has a particular DNA profile? What is the probability a person has a particular profile if it has already been seen once?

The first question is a little easier to think about, but difficult to answer in practice: it is very unlikely that a profile will be seen in any sample of profiles. Even for one STR locus with 10 alleles, there are 55 different genotypes and most of those will not occur in a sample of a few hundred profiles.

For locus D3S1358 in the African American population, the FBI frequency database shows that 31 of the 55 genotype counts are zero. Estimating the population frequencies for these 31 types as zero doesn't seem sensible.

D3S1358 Genotype Counts

Observed	< 12	12	13	14	15	16	17	18	19	> 19
< 12	0									
12	0	0								
13	0	0	0							
14	0	0	0	2						
15	0	0	1	19	15					
16	1	1	1	15	39	19				
17	0	0	2	10	26	24	9			
18	1	0	1	2	6	10	3	0		
19	0	0	0	1	0	0	1	0	0	
> 19	0	0	0	0	1	0	0	0	0	0

Hardy-Weinberg Law

A solution to the problem is to assume that the Hardy-Weinberg Law holds. For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

D3S1358 Hardy-Weinberg Calculations

The allele counts for D3S1358 in the African-American sample are:

											Total
Allele	< 12	12	13	14	15	16	17	18	19	> 19	
Count	2	1	5	51	122	129	84	23	2	1	420

If the Hardy-Weinberg Law holds, then we would expect to see $n\tilde{p}_{13}^2 = 210 \times (5/420)^2 = 0.03$ individuals of type 13,13 in a sample of 210 individuals.

Also, we would expect to see $2n\tilde{p}_{13}\tilde{p}_{14} = 420 \times (5/420) \times (51/420) = 0.61$ individuals of type 13,14 in a sample of 210 individuals.

Other values are shown on the next slide.

D3S1358 Observed and Expected Counts

		< 12	12	13	14	15	16	17	18	19	> 19
< 12	Obs.	0									
	Exp.	0.0									
12	Obs.	0	0								
	Exp.	0.0	0.0								
13	Obs.	0	0	0							
	Exp.	0.0	0.0	0.0							
14	Obs.	0	0	0	2						
	Exp.	0.2	0.1	0.6	3.1						
15	Obs.	0	0	1	19	15					
	Exp.	0.6	0.3	1.5	14.8	17.7					
16	Obs.	1	1	1	15	39	19				
	Exp.	0.6	0.3	1.5	15.7	37.5	19.8				
17	Obs.	0	0	2	10	26	24	9			
	Exp.	0.4	0.2	1.0	10.2	24.4	25.8	8.4			
18	Obs.	1	0	1	2	6	10	3	0		
	Exp.	0.1	0.1	0.3	2.8	6.7	7.1	4.6	0.6		
19	Obs.	0	0	0	1	0	0	1	0	0	
	Exp.	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.1	0.0	
> 19	Obs.	0	0	0	0	1	0	0	0	0	0
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0	0.0

Testing for Hardy-Weinberg Equilibrium

A test of the Hardy-Weinberg Law will somehow decide if the observed and expected numbers are sufficiently similar that we can proceed as though the law can be used.

In one of the first applications of Hardy-Weinberg testing in a US forensic setting:

> “To justify applying the classical formulas of population genetics in the Castro case the Hispanic population must be in Hardy-Weinberg equilibrium. Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium.”

E.S. Lander. 1989. DNA fingerprinting on trial. Nature 339: 501-505.

VNTR “Coalescence”

Forensic DNA profiling initially used minisatellites, or VNTR loci, with large numbers of alleles. Heterozygotes would be scored as homozygotes if the two alleles were so similar in length that they coalesced into one band on an autoradiogram. Small alleles often not detected at all, and this is the cause of Lander’s finding.

Considerable debate in early 1990s on alternative “binning” strategies for reducing the number of alleles (Science 253:1037-1041, 1991).

Typing has moved to microsatellites with fewer and more easily distinguished alleles, but testing for Hardy-Weinberg equilibrium continues. There are still reasons why the law may not hold.

Population Structure can Cause Departure from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Structure

Effect of population structure taken into account with the "theta-correction." Matching probabilities allow for a variance in allele frequencies among subpopulations.

$$\text{Pr}(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

where p_A is the average allele frequency over all subpopulations. We will come back to this expression.

Population Admixture

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marrieages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Exact HWE Test

The preferred test for HWE is an "exact" one. The test rests on the assumption that individuals are sampled randomly from a population so that genotype counts have a multinomial distribution:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}}$$

This equation is always true, and when there is HWE ($P_{AA} = p_A^2$ etc.) there is the additional result that the allele counts have a binomial distribution:

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}$$

Exact HWE Test

Putting these together gives the conditional probability of the genotypic data given the allelic data and given HWE:

$$\begin{aligned}
\mathrm{Pr}(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \mathrm{HWE}) &= \frac{\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(p_A^2)^{n_{AA}}(2p_Ap_a)^{n_{Aa}}(p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}} \\
&= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\frac{2^{n_{Aa}}n_A!n_a!}{(2n)!}
\end{aligned}$$

Reject the Hardy-Weinberg hypothesis if this probability is unusually small.

Exact HWE Test Example

Reject the HWE hypothesis if the probability of the genotypic array, conditional on the allelic array, is among the smallest probabilities for all the possible sets of genotypic counts for those allele counts.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a, \text{HWE})$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data on the previous slide, conditional on the allele frequencies and on HWE, is $1/99 = 0.01$. This is less than the conventional 5% significance level.

In general, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true.

Exact HWE Test

Still in the two-allele case, for a sample of size $n = 100$ with minor allele frequency of 0.07, there are only 8 sets of genotype counts:

			Exact	
n_{AA}	n_{Aa}	n_{aa}	Prob.	p-value
93	0	7	0.0000	0.0000*
92	2	6	0.0000	0.0000*
91	4	5	0.0000	0.0000*
90	6	4	0.0002	0.0002*
89	8	3	**0.0051**	**0.0053***
88	10	2	0.0602	0.0654
87	12	1	0.3209	0.3863
86	14	0	0.6136	1.0000

So, for a nominal 5% significance level, the actual significance level is 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$.

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Mark a set of five index cards to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Check the following theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	$\frac{1}{21} = 0.048$
2	2	1	$\frac{12}{21} = 0.571$
1	4	0	$\frac{8}{21} = 0.381$

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

Permutation Test for D3S1358

For a STR locus, where $\{n_g\}$ are the genotype counts and $n = \sum_g n_g$ is the sample size, and $\{n_a\}$ are the alleles counts with $2n = \sum_a n_a$, the exact test statistic is

$$\Pr(\{n_g\}|\{n_a\}, \text{HWE}) = \frac{n!2^H \prod_a n_a!}{\prod_g n_g!(2n)!}$$

where H is the count of heterozygotes.

This probability for the African American genotypic counts at D3S1358 is 0.6163×10^{-13}, which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even small probability.

Linkage Disequilibrium

This term is generally reserved for association between pairs of alleles – one at each of two loci. In the present context, it may simply mean some lack of independence of profile or match probabilities at different loci.

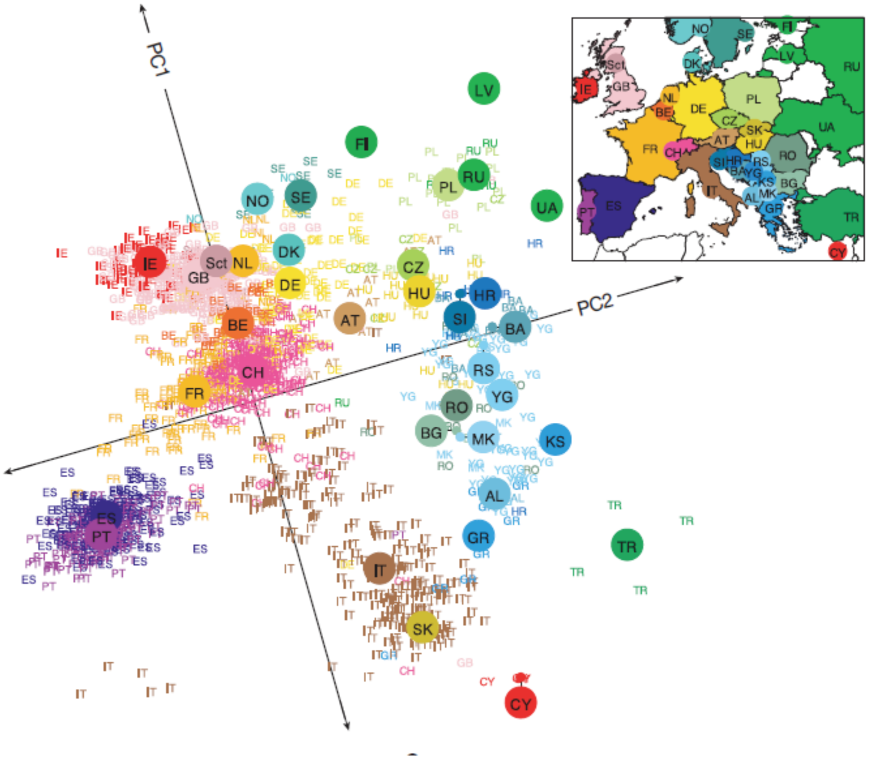
Unlinked loci are expected to be almost independent.

Human Populations: History and Structure

In the paper
Novembre J, Johnson, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann A, Nelson MB, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456:98
there is quite dramatic evidence that our genetic profiles contain information about where we live, suggesting that these profiles reflect the history of our populations.

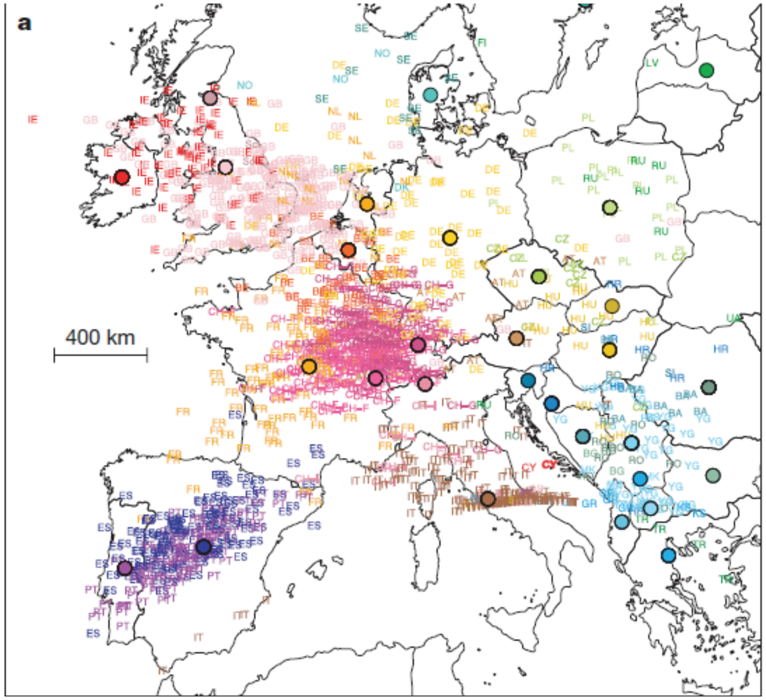
The authors collected "SNP" (single nucleotide polymorphism) data on over people living in Europe. Either the country of origin of the people's grandparents or their own country of birth was known. On the next slide, these geographic locations were used to color the location of each of 1,387 people in "genetic space." Instead of latitude and longitude on a geographic map, their first two principal components were used: these components summarize the 500,000 SNPs typed for each person.

Novembre et al., 2008



Novembre et al., 2008

As a follow-up, the authors took the genetic profile of each person and used it to predict their latitude and longitude, and plotted these on a geographic map. These predicted positions are colored by the country of origin of each person.



Y SNP Data Haplogroups

Another set of SNP data, this time from around the world, is available for the Y chromosome. These data were collected for the 1000 Genomes project (http://www.1000genomes.org/): there are 26 populations:

East Asia: CDX. Chinese Dai in Xishuangbanna; CHB. Han Chinese in Beijing; JPT. Japanese in Tokyo; KHV. Kinh in Ho Chi Minh City; CHS. Southern Han Chinese.

South Asian: BEB. Bengali in Bangladesh; GIH. Gujarati Indian in Houston; ITU. India Telugi in UK; PJL. Punjabi in Lahore; STU. Sri Lankan Tamil in UK.

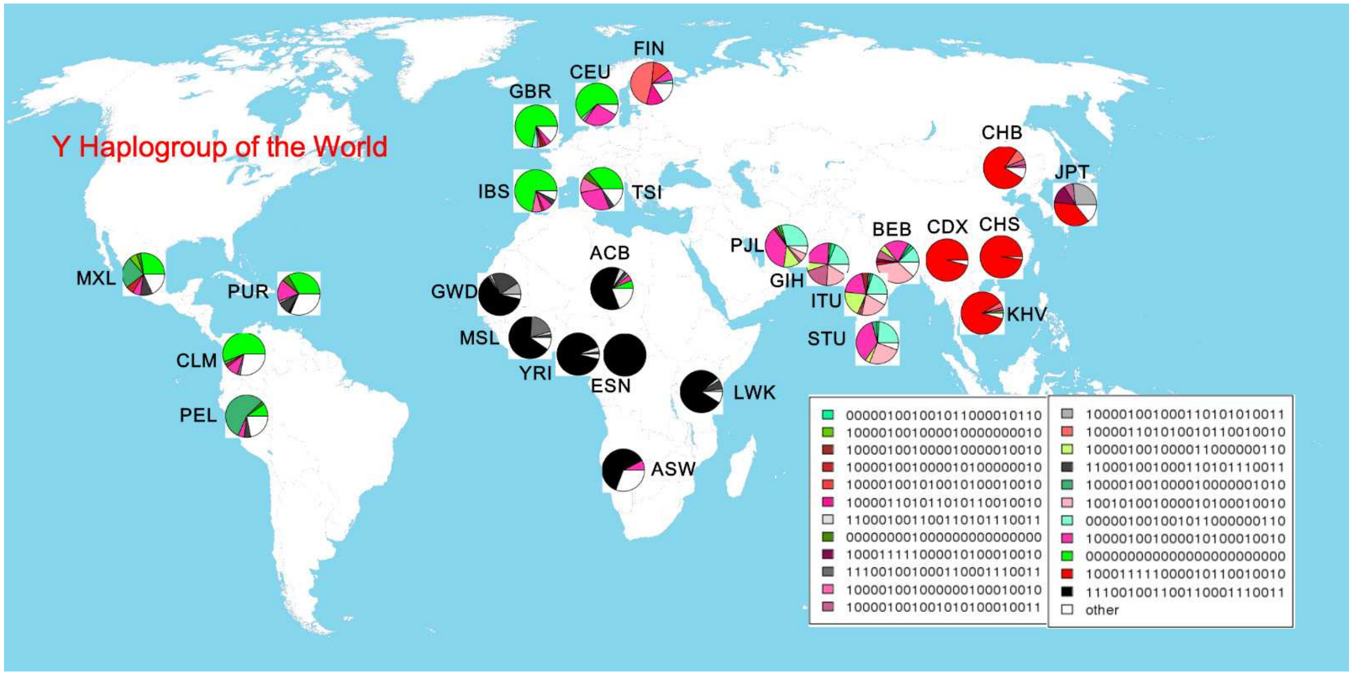
Y SNP Data Haplogroups

African: ASW. African Ancestry in Southwest US; ACB. African Caribbean in Barbados; ESN. Esan in Nigeria; GWD. Gambian in the Gambia; LWK. Luthya in Kenya; MSL. Mende in Sierra Leone; YRI. Yoruba in Nigeria.

European: GBR. British in UK; FIN. Finnish in Finland; IBS. Iberian in Spain; TSI. Toscani in Italy; CEU. Utah residents with European ancestry.

Americas: CLM. Columbian in Medellin; MXL. Mexican in Los Angeles; PEL. Peruvian in Lima, PUR. Puerto Rican in Puerto Rico.

Y SNP Data Haplogroups



Migration History of Early Humans

An interesting video of the migration of early humans is available at:

http://www.bradshawfoundation.com/journey/

Migration Map of Early Humans

https://genographic.nationalgeographic.com/human-journey/

This map summarizes the migration patterns of early humans.

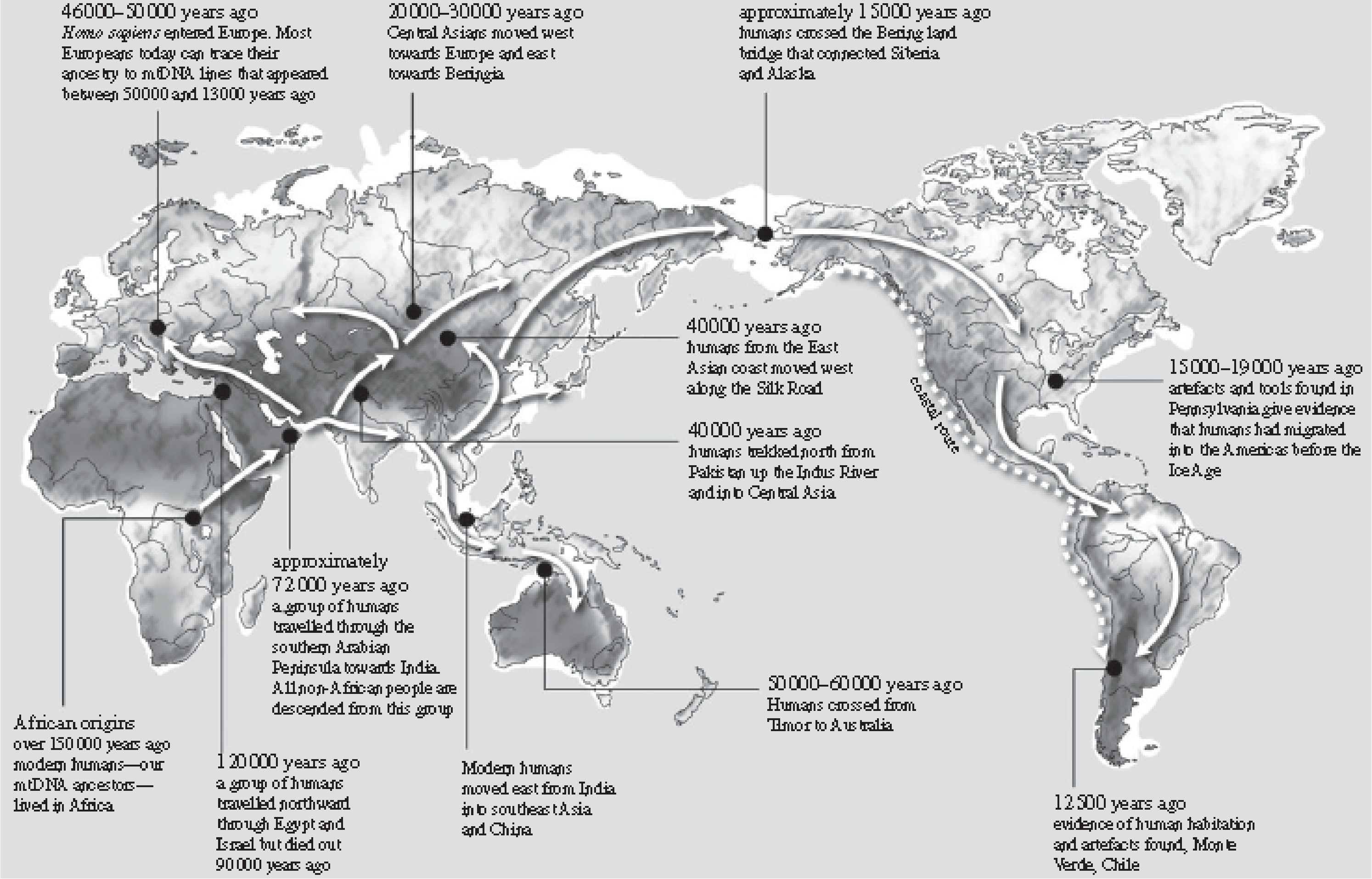
Migration Map of Early Humans

The map on the next slide, based on mitochondrial genetic profiles, is taken from:

Oppenheimer S. 2012. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Phil. Trans. R. Soc. B (2012) 367, 770-784 doi:10.1098/rstb.2011.0306.

The first two pages of this paper give a good overview, and they contain this quote: “The finding of a greater genetic diversity within Africa, when compared with outside, is now abundantly supported by many genetic markers; so Africa is the most likely geographic origin for a modern human dispersal.”

Migration Map of Early Humans



46000–50 000 years ago
Homo sapiens entered Europe. Most Europeans today can trace their ancestry to mtDNA lines that appeared between 50000 and 13000 years ago

20 000–30000 years ago
Central Asians moved west towards Europe and east towards Beringia

approximately 15000 years ago
humans crossed the Bering land bridge that connected Siberia and Alaska

40000 years ago
humans from the East Asian coast moved west along the Silk Road

40 000 years ago
humans trekked north from Pakistan up the Indus River and into Central Asia

15 000–19 000 years ago
artefacts and tools found in Pennsylvania give evidence that humans had migrated into the Americas before the Ice Age

approximately
72 000 years ago
a group of humans travelled through the southern Arabian Peninsula towards India. All non-African people are descended from this group

African origins
over 150 000 years ago modern humans—our mtDNA ancestors—lived in Africa

120 000 years ago
a group of humans travelled northward through Egypt and Israel but died out 90 000 years ago

Modern humans moved east from India into southeast Asia and China

50 000–60 000 years ago
Humans crossed from Timor to Australia

12 500 years ago
evidence of human habitation and artefacts found, Monte Verde, Chile

Forensic Implications

What does the theory about the spread of modern humans tell us about how to interpret matching profiles?

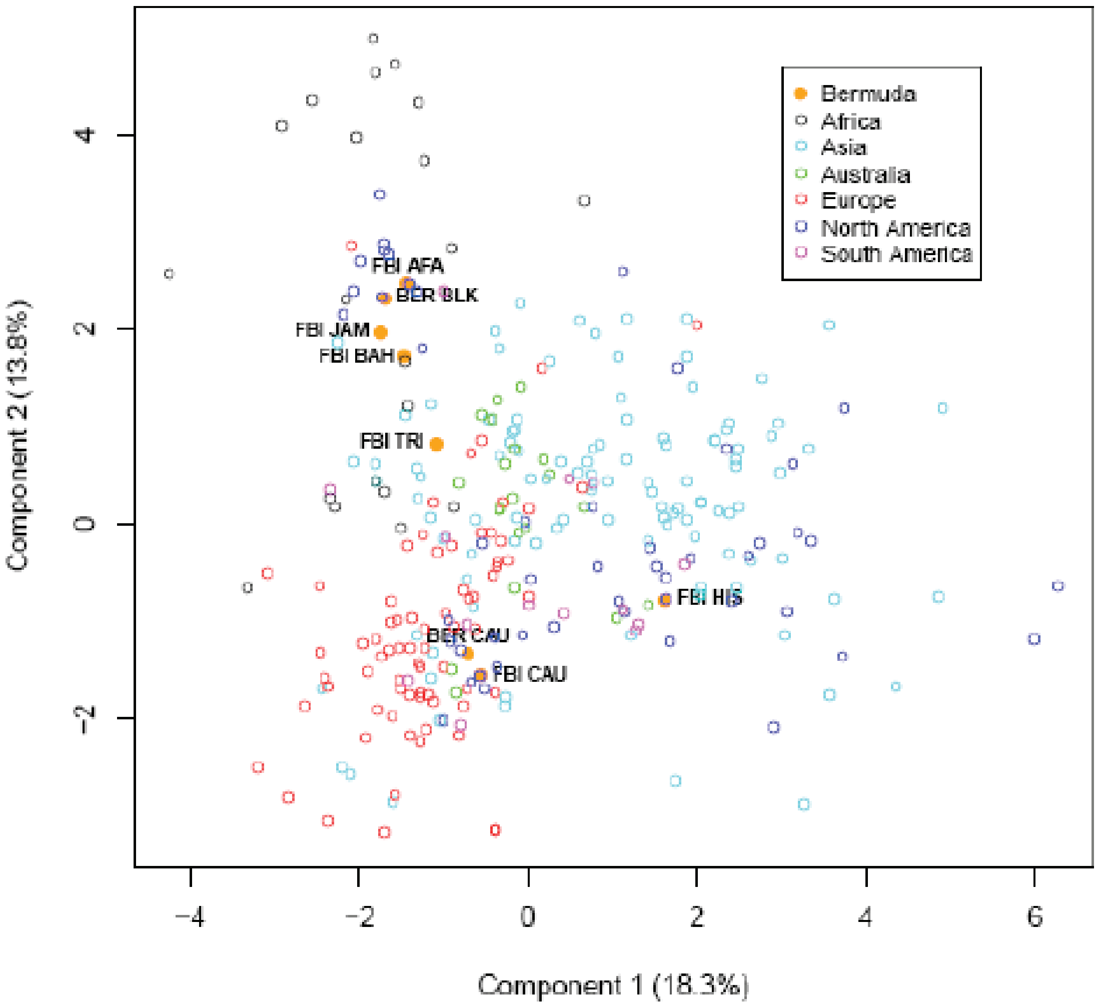
Matching probabilities should be bigger within populations, and more similar among populations that are closer together in time.

Forensic allele frequencies are consistent with the theory of human migration patterns.

Forensic STR PCA Map

A large collection of forensic STR allele frequencies was used to construct the principal component map on the next page. Also shown are some data collected by forensic agencies in the Caribbean, and by the FBI. The Bermuda police has been using FBI data - does this seem to be reasonable?

Forensic STR PCA Map

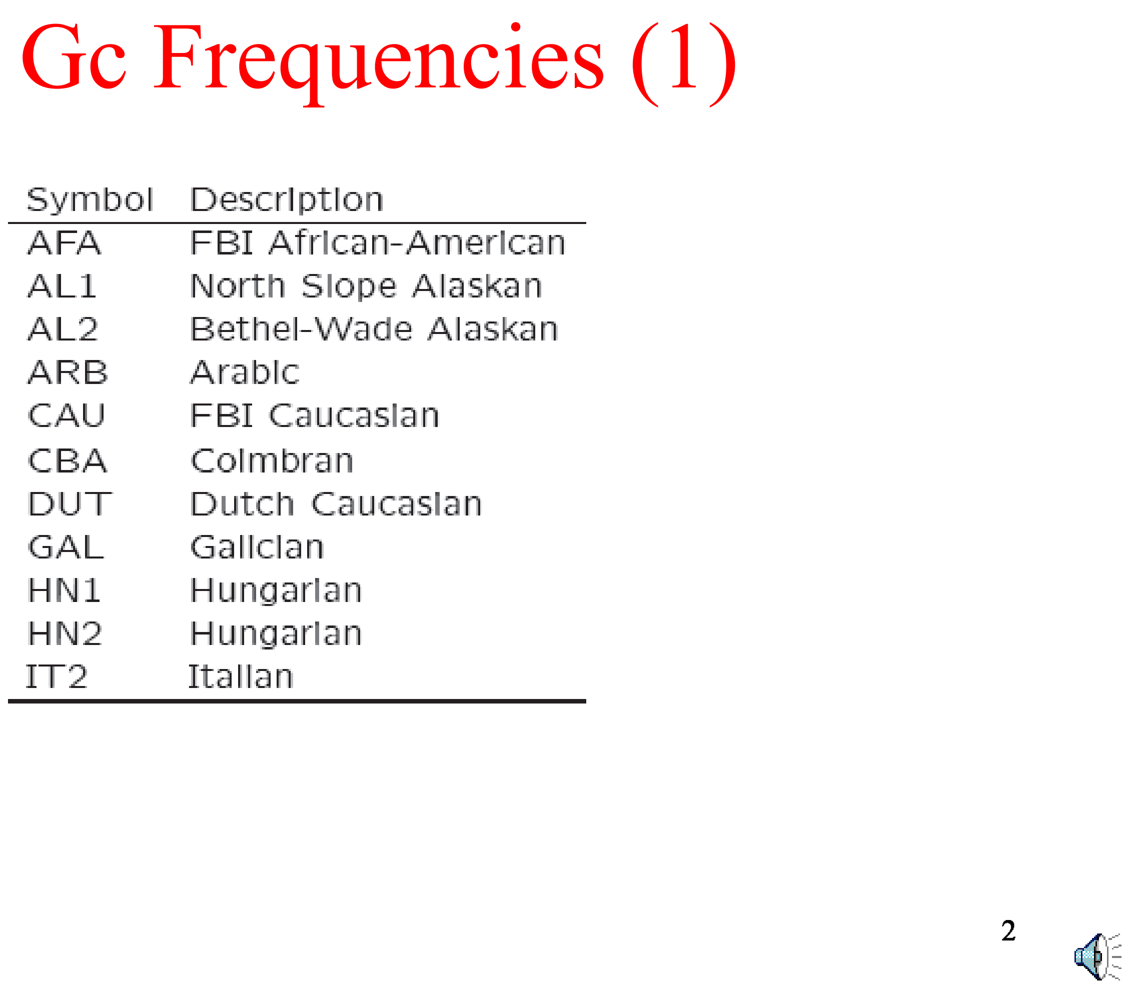


Genetic Distances

Forensic allele frequencies were collected from 21 populations. The next slides list the populations and show allele frequencies for the Gc marker. This has only three alleles, A, B, C.

The matching proportions within each population, and between each pair of populations, were calculated. These allow distances ("theta" or β_{ij}) to be calculated for each pair of populations i, j: $\widehat{\beta}_{ij} = (\tilde{M}_i + \tilde{M}_j - 2\tilde{M}_{Bij})/[2(1 - \tilde{M}_{Bij})]$.

Genetic Distances

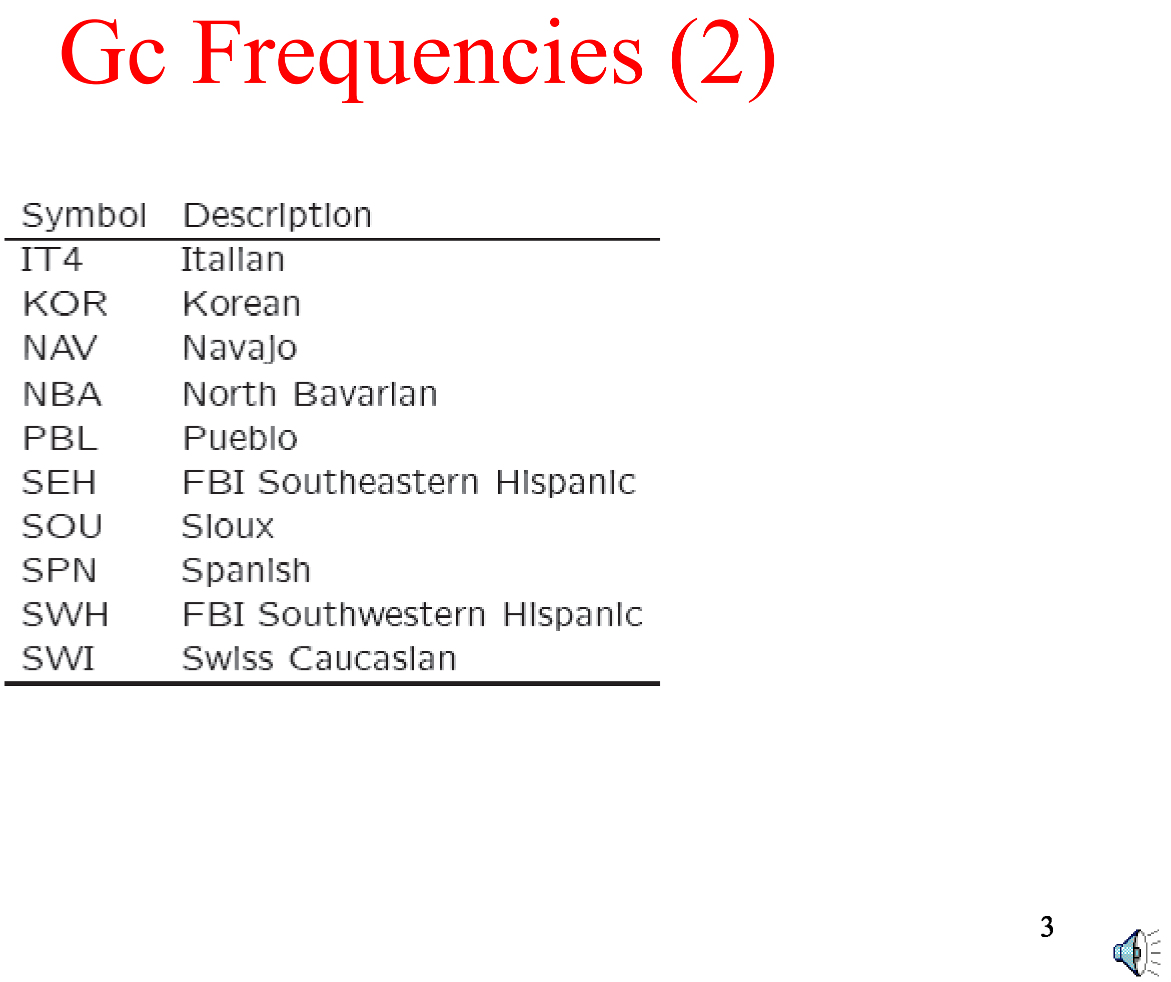


Gc Frequencies (1)

Symbol	DescrIptIon
AFA	FBI AfrIcan-AmerIcan
AL1	North Slope Alaskan
AL2	Bethel-Wade Alaskan
ARB	ArabIc
CAU	FBI CaucasIan
CBA	CoImbran
DUT	Dutch CaucasIan
GAL	GalIcIan
HN1	HungarIan
HN2	HungarIan
IT2	ItalIan

2

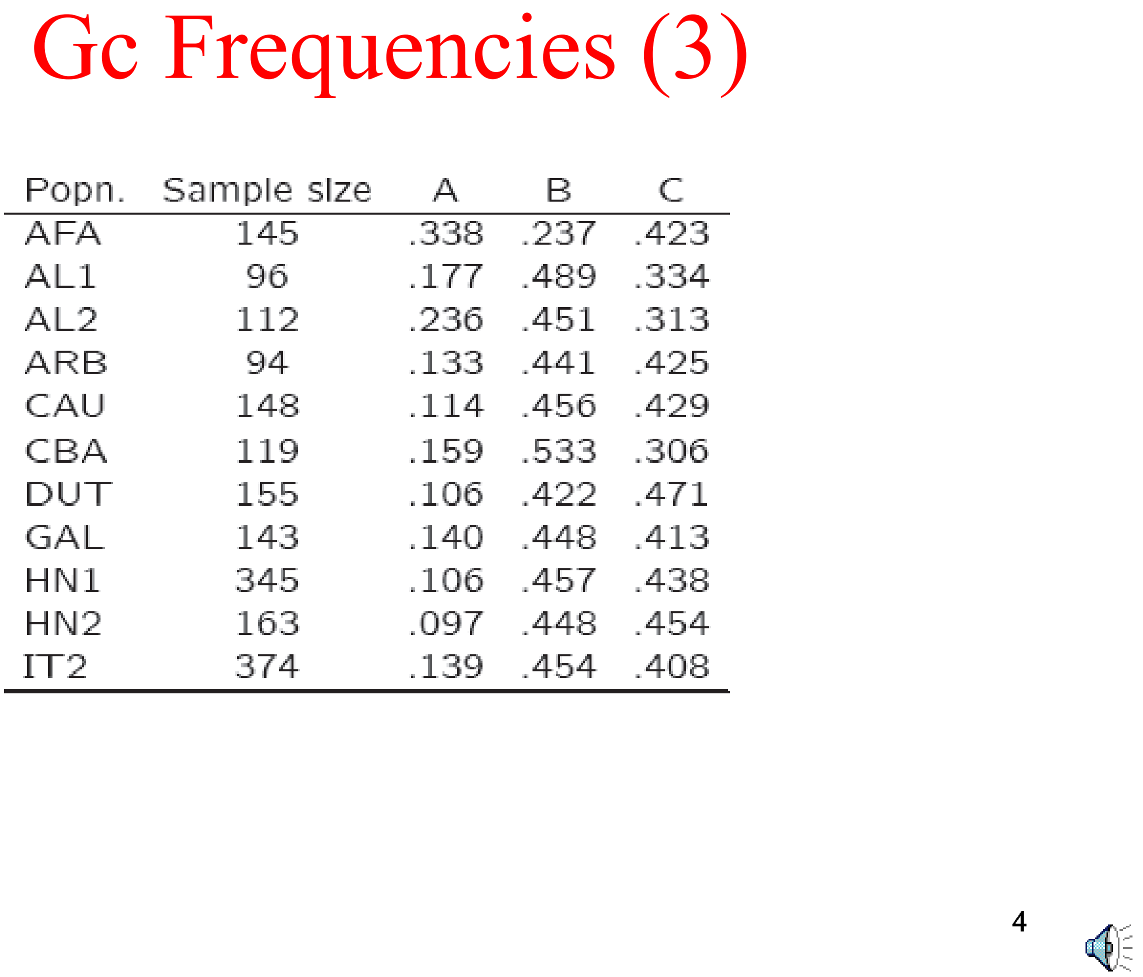
Genetic Distances



Gc Frequencies (2)

Symbol	Descrlptlon
IT4	Itallan
KOR	Korean
NAV	NavaJo
NBA	North Bavarlan
PBL	Pueblo
SEH	FBI Southeastern HIspanlc
SOU	Sloux
SPN	Spanlsh
SWH	FBI Southwestern HIspanlc
SWI	Swlss Caucaslan

Genetic Distances

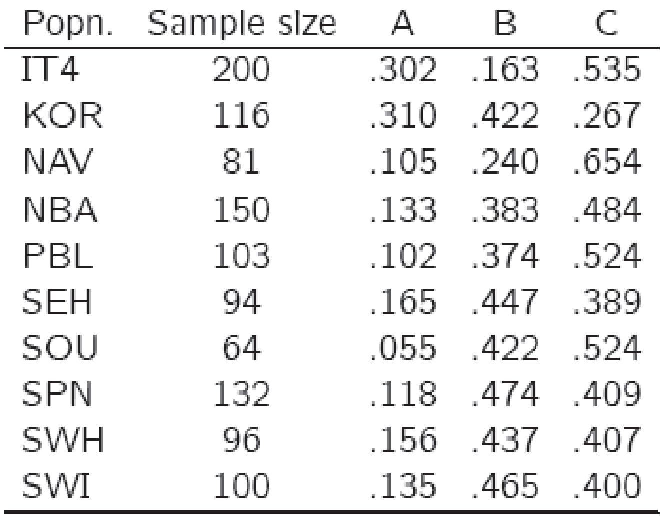


Gc Frequencies (3)

Popn.	Sample sIze	A	B	C
AFA	145	.338	.237	.423
AL1	96	.177	.489	.334
AL2	112	.236	.451	.313
ARB	94	.133	.441	.425
CAU	148	.114	.456	.429
CBA	119	.159	.533	.306
DUT	155	.106	.422	.471
GAL	143	.140	.448	.413
HN1	345	.106	.457	.438
HN2	163	.097	.448	.454
IT2	374	.139	.454	.408

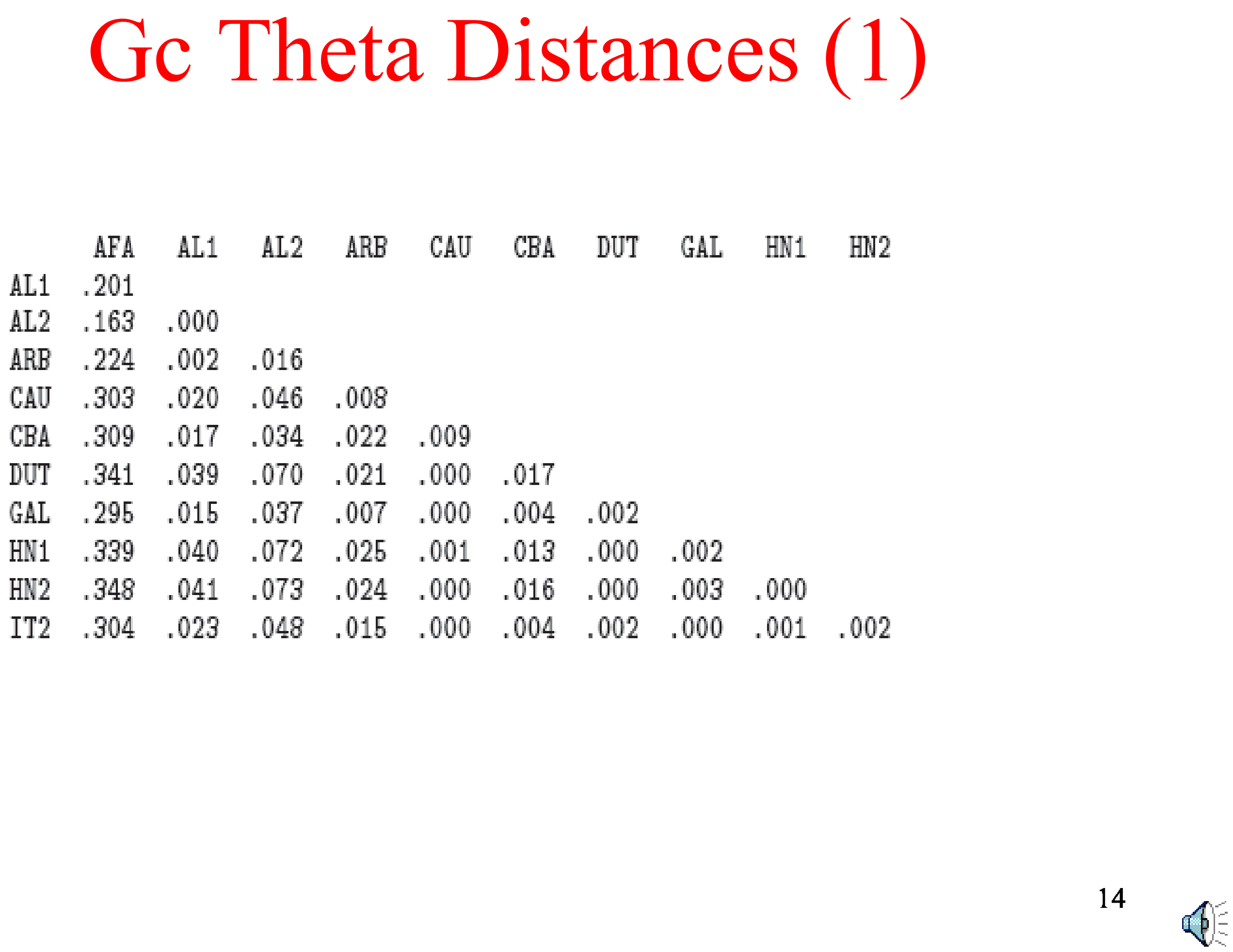
4

Gc Frequencies (4)



Popn.	Sample slze	A	B	C
IT4	200	.302	.163	.535
KOR	116	.310	.422	.267
NAV	81	.105	.240	.654
NBA	150	.133	.383	.484
PBL	103	.102	.374	.524
SEH	94	.165	.447	.389
SOU	64	.055	.422	.524
SPN	132	.118	.474	.409
SWH	96	.156	.437	.407
SWI	100	.135	.465	.400

Genetic Distances

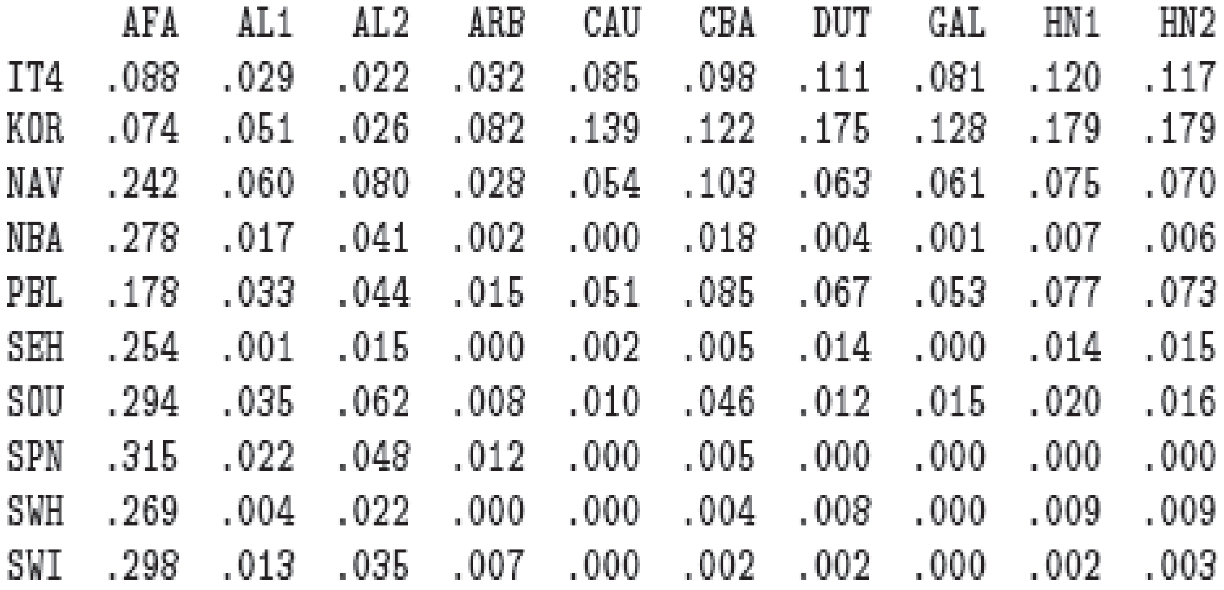


Gc Theta Distances (1)

	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
AL1	.201									
AL2	.163	.000								
ARB	.224	.002	.016							
CAU	.303	.020	.046	.008						
CBA	.309	.017	.034	.022	.009					
DUT	.341	.039	.070	.021	.000	.017				
GAL	.295	.015	.037	.007	.000	.004	.002			
HN1	.339	.040	.072	.025	.001	.013	.000	.002		
HN2	.348	.041	.073	.024	.000	.016	.000	.003	.000	
IT2	.304	.023	.048	.015	.000	.004	.002	.000	.001	.002

Genetic Distances

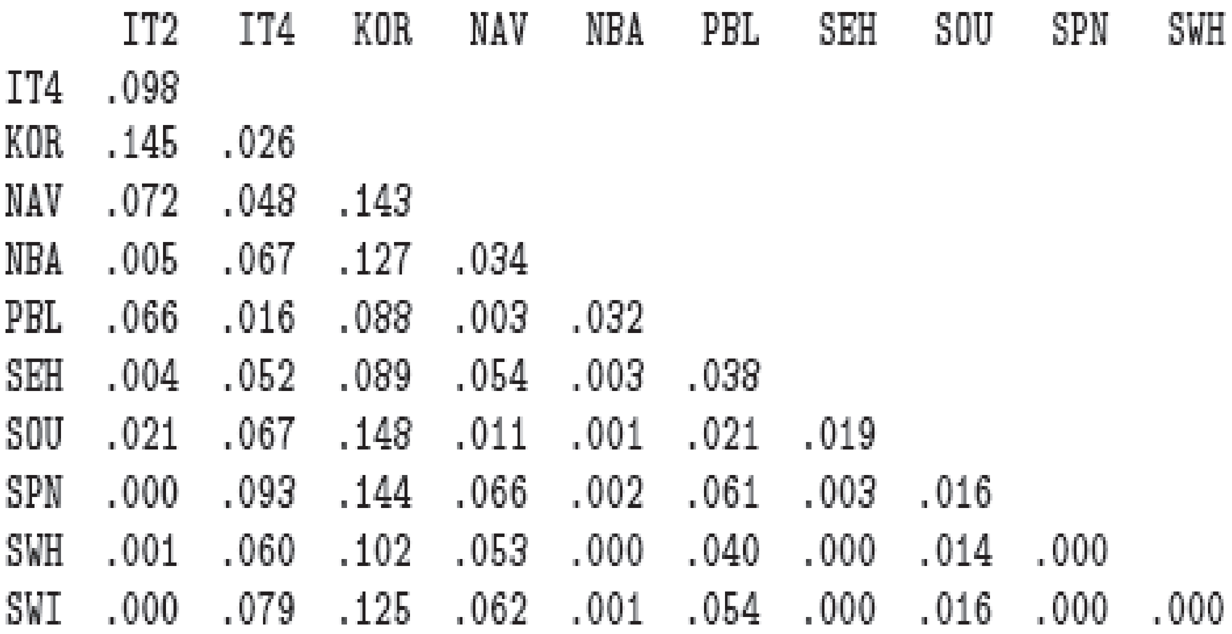
Gc Theta Distances (2)



	AFA	AL1	AL2	ARB	CAU	CBA	DUT	GAL	HN1	HN2
IT4	.088	.029	.022	.032	.085	.098	.111	.081	.120	.117
KOR	.074	.051	.026	.082	.139	.122	.175	.128	.179	.179
NAV	.242	.060	.080	.028	.054	.103	.063	.061	.075	.070
NBA	.278	.017	.041	.002	.000	.018	.004	.001	.007	.006
PBL	.178	.033	.044	.015	.051	.085	.067	.053	.077	.073
SEH	.254	.001	.015	.000	.002	.005	.014	.000	.014	.015
SOU	.294	.035	.062	.008	.010	.046	.012	.015	.020	.016
SPN	.315	.022	.048	.012	.000	.005	.000	.000	.000	.000
SWH	.269	.004	.022	.000	.000	.004	.008	.000	.009	.009
SWI	.298	.013	.035	.007	.000	.002	.002	.000	.002	.003

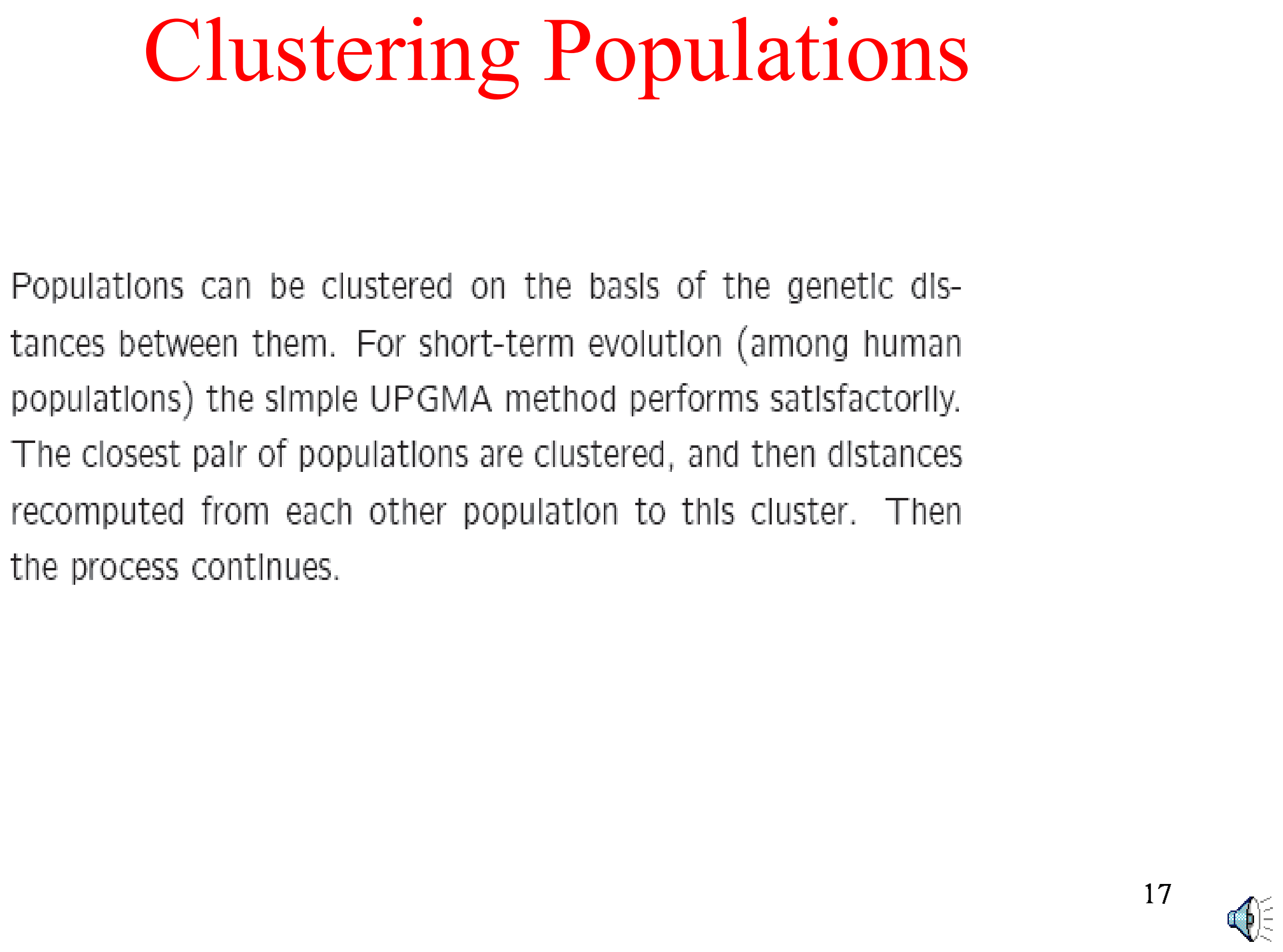
Genetic Distances

Gc Theta Distances (3)



	IT2	IT4	KOR	NAV	NBA	PBL	SEH	SOU	SPN	SWH
IT4	.098									
KOR	.145	.026								
NAV	.072	.048	.143							
NBA	.005	.067	.127	.034						
PBL	.066	.016	.088	.003	.032					
SEH	.004	.052	.089	.054	.003	.038				
SOU	.021	.067	.148	.011	.001	.021	.019			
SPN	.000	.093	.144	.066	.002	.061	.003	.016		
SWH	.001	.060	.102	.053	.000	.040	.000	.014	.000	
SWI	.000	.079	.125	.062	.001	.054	.000	.016	.000	.000

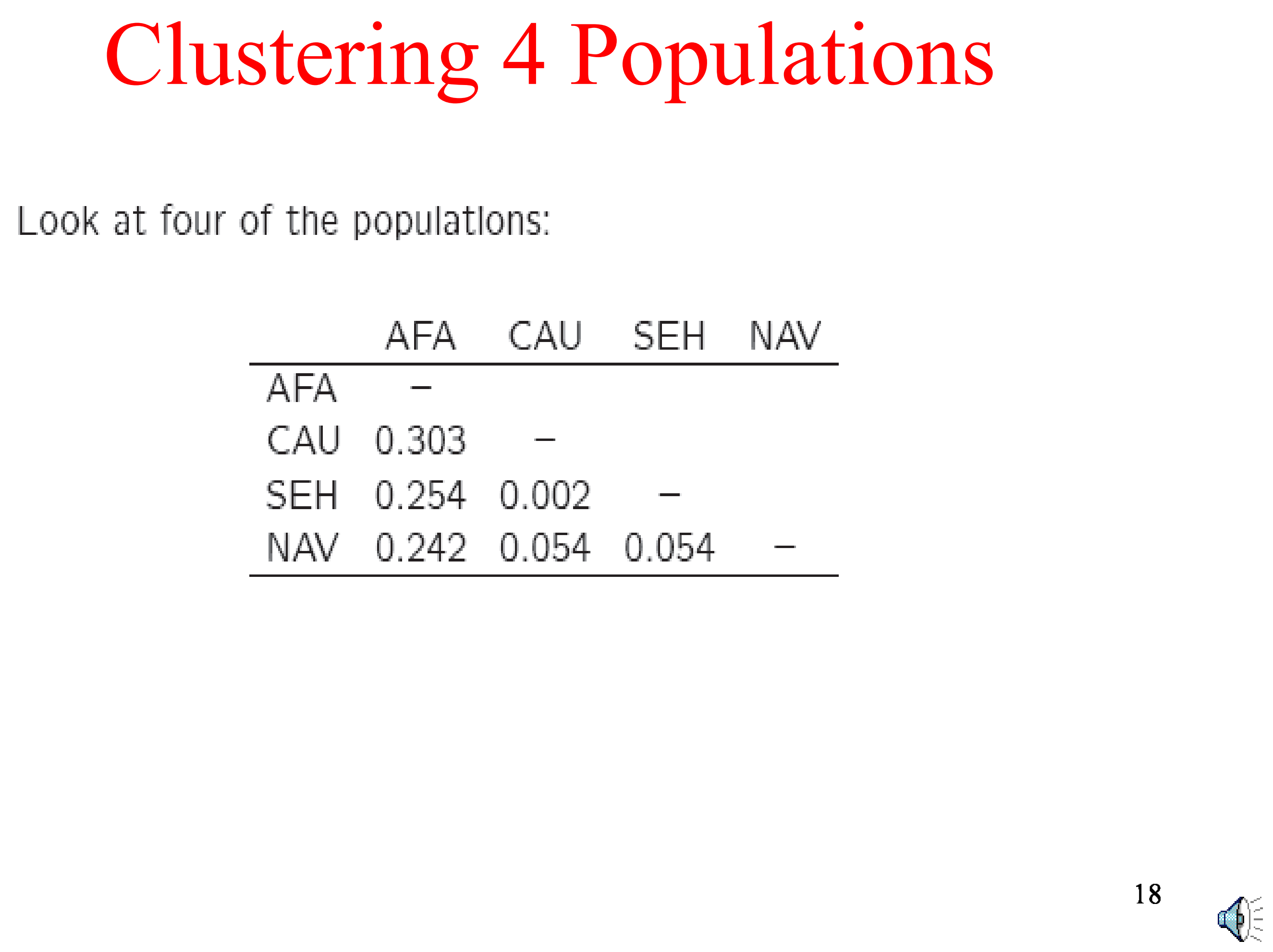
Genetic Distances



Clustering Populations

Populations can be clustered on the basis of the genetic distances between them. For short-term evolution (among human populations) the simple UPGMA method performs satisfactorily. The closest pair of populations are clustered, and then distances recomputed from each other population to this cluster. Then the process continues.

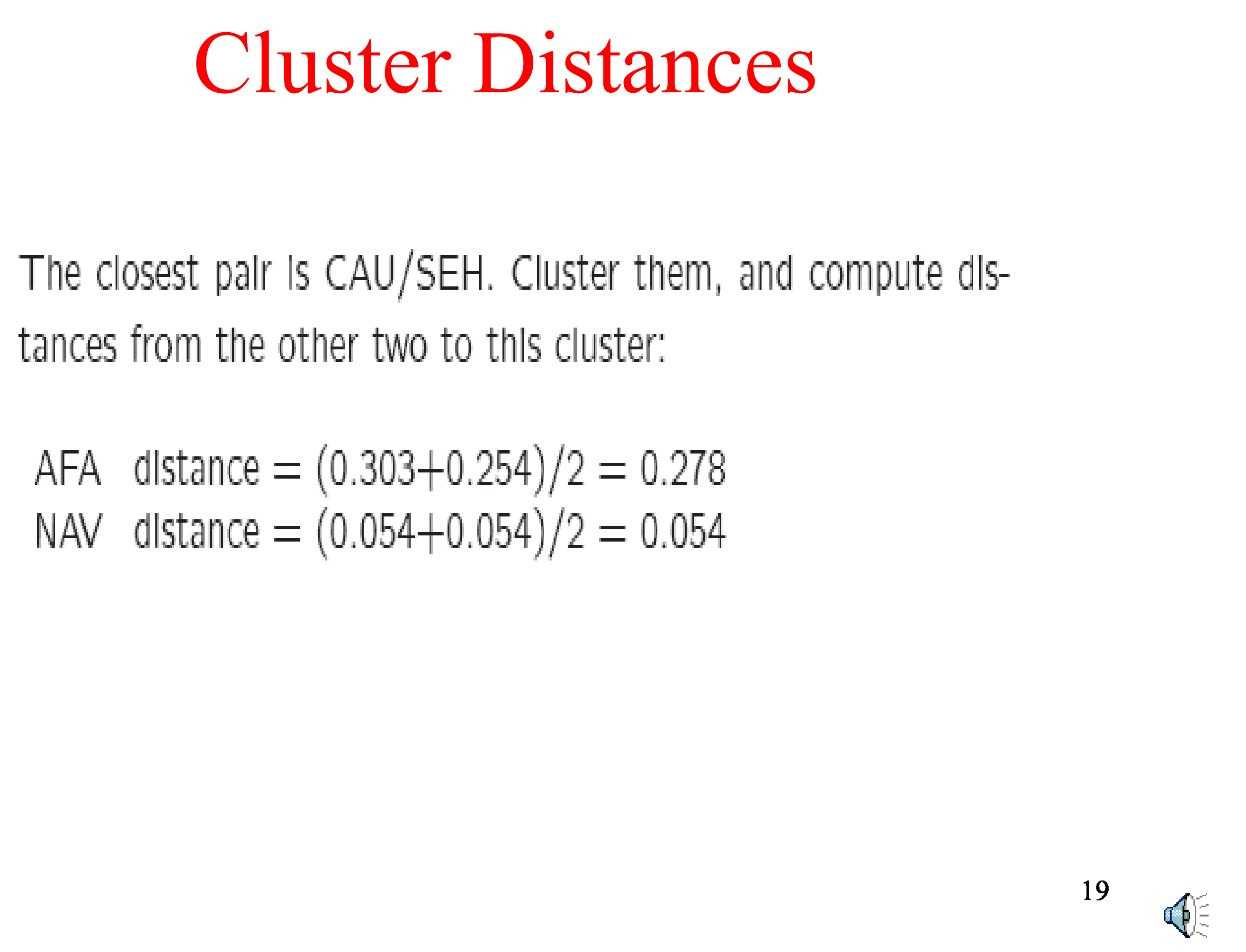
17



Clustering 4 Populations

Look at four of the populatlons:

	AFA	CAU	SEH	NAV
AFA	–			
CAU	0.303	–		
SEH	0.254	0.002	–	
NAV	0.242	0.054	0.054	–



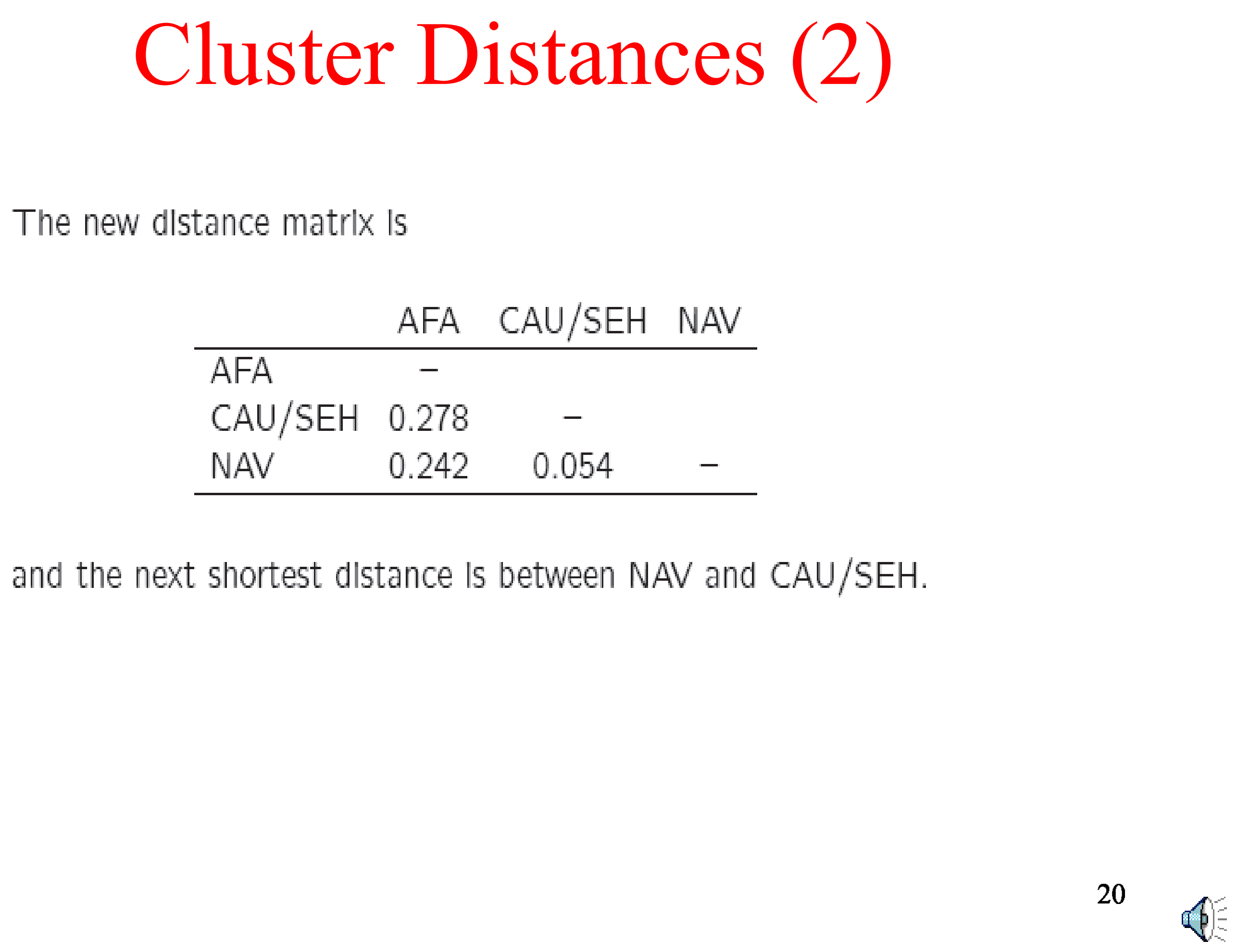
Cluster Distances

The closest palr ls CAU/SEH. Cluster them, and compute dlstances from the other two to thls cluster:

AFA dlstance $= (0.303+0.254)/2 = 0.278$
NAV dlstance $= (0.054+0.054)/2 = 0.054$

19

Genetic Distances



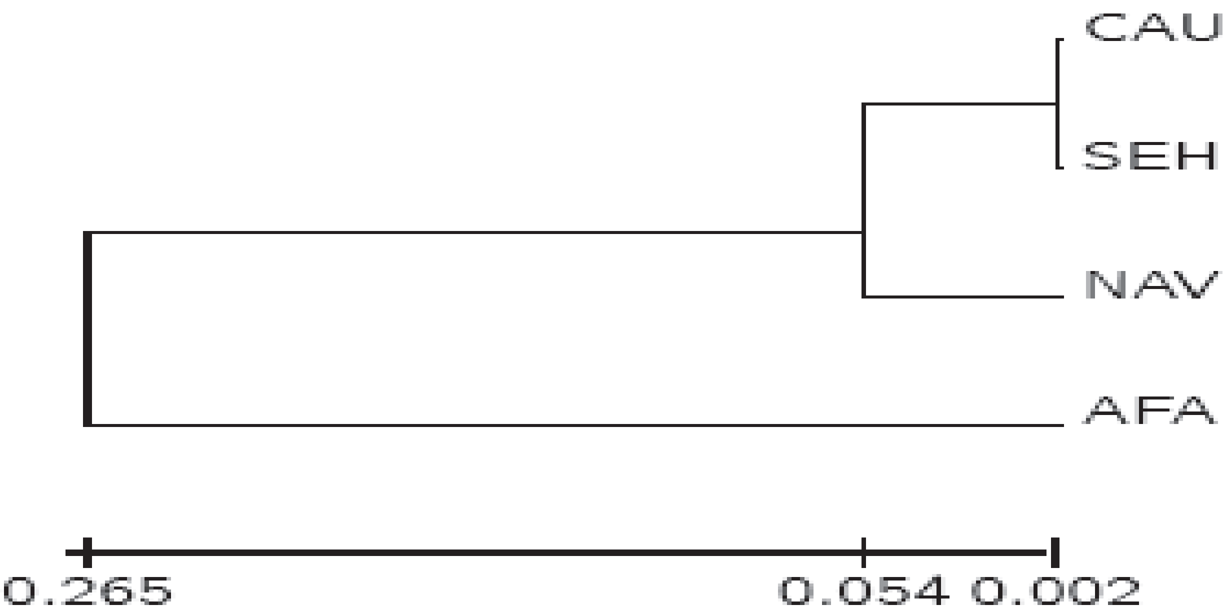
Cluster Distances (2)

The new dlstance matrlx ls

	AFA	CAU/SEH	NAV
AFA	–		
CAU/SEH	0.278	–	
NAV	0.242	0.054	–

and the next shortest dlstance ls between NAV and CAU/SEH.

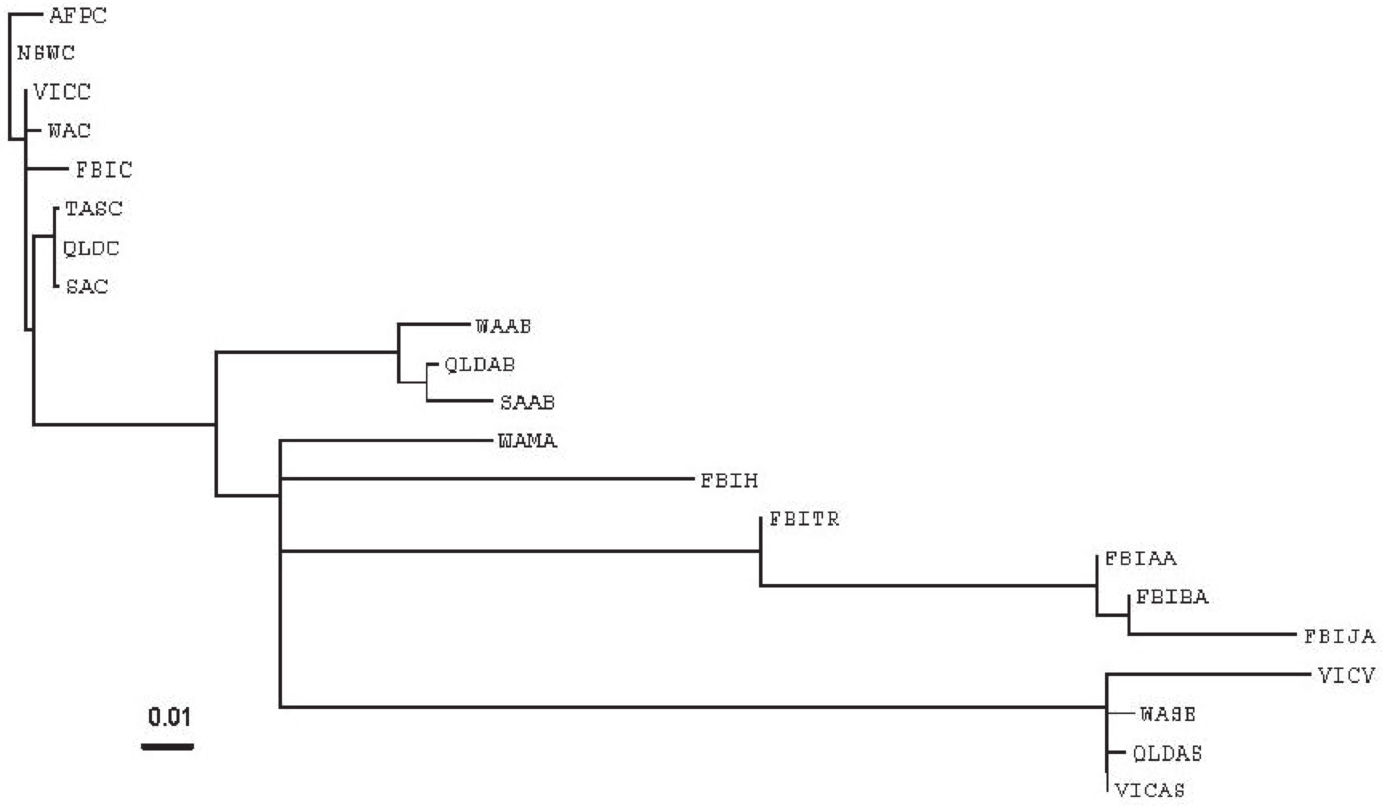
Gc UPGMA Dendrogram



21

Australian STR Data

Australian Values

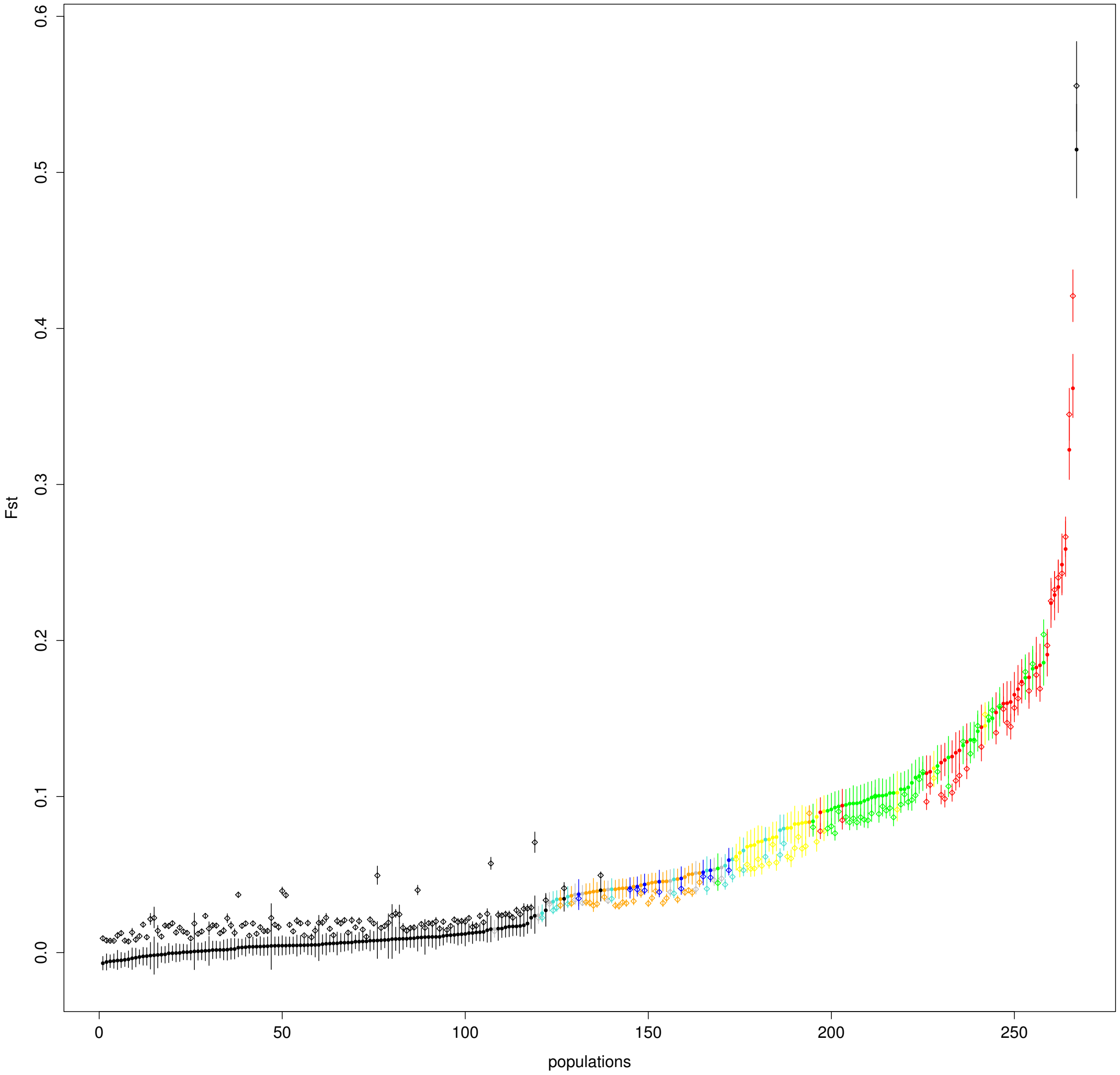


Worldwide Survey of STR Data

Published allele frequencies for 24 STR loci were obtained for 446 populations. For each population i, the within-population matching proportion $\tilde{M}_i$ was calculated. Also the average $\tilde{M}_B$ of all the between-population matching proportions. The "θ" for each population is calculated as $\widehat{\beta}_i = (\tilde{M}_i - \tilde{M}_B)/(1 - \tilde{M}_B)$. These are shown on the next slide, ranked from smallest to largest and colored by continent.

Africa: black; America: red; South Asia: orange; East Asia: yellow; Europe: blue; Latino: turquoise; Middle East: grey; Oceania: green.

Worldwide Survey of STR Data



Allelic Matching

Within-population Matching

The key forensic genetic issue is that of matching profiles. What is the probability that two people have the same STR profile?

We can get some empirical estimate of this when we have a set of profiles. For the African -American sample of 210 profiles for D3S1358, how many pairs of profiles match? Only those genotypes that occur more than once in the sample provide matches. To simplify this initial discussion, consider the following data for the Y-STR locus DYS390 from the NIST database:

Allele Counts in NIST Data for DYS390

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
20	4	1	1	0	6
21	176	4	17	1	198
22	43	45	14	17	119
23	36	116	50	17	219
24	56	145	129	21	351
25	23	46	21	36	126
26	3	2	2	4	11
27	0	0	2	0	2
Total	341	359	236	96	1032

Within- and Between-population Matching for DYS390

Within the African-American sample there are $341 \times 340 = 115,940$ pairs of profiles and the number of matches is

$$4\times3+176\times175+43\times42+36\times35+56\times55+23\times22+3\times2 = 37,470$$

so the within-population matching proportion is $37,470/115,940 = 0.323$.

Between the African-American and Caucasian samples, there are $341 \times 359 = 122,419$ pairs of profiles and the number of matches is

$$4\times1+176\times4+43\times45+36\times116+56\times145+23\times4+3\times2 = 12,403$$

so the between-population matching proportion is $12,403/122,419 = 0.101$.

Allele Counts in NIST Data for DYS391

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
7	0	0	1	0	1
8	0	1	0	1	2
9	2	12	16	3	33
10	238	162	128	79	607
11	93	175	89	13	370
12	7	9	2	0	18
13	1	0	0	0	1
Total	341	359	236	96	1032

The within-population matching proportion for the African-American sample is 65,006/115,940=0.561.

The between-population matching proportion for the African-American and Caucasian samples is 54,918/122,419=0.449.

Two-locus counts in NIST African-American Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	34	1122
22	11	9	72
24	10	15	210
24	11	39	1482
24	12	1	0
24	9	1	0
23	10	19	342
23	11	14	182
23	12	3	6
21	10	157	24492
21	11	15	210
21	12	2	2
21	9	1	0
21	13	1	0
25	10	11	110
25	11	12	132
26	10	1	0
26	11	2	2
20	10	1	0
20	11	2	2
20	12	1	0

Two-locus counts in NIST Caucasian Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	43	1806
22	11	1	0
22	9	1	0
24	10	48	2256
24	11	88	7656
24	12	4	12
24	9	5	20
23	10	50	2450
23	11	60	3540
23	12	2	2
23	9	3	6
23	8	1	0
21	10	3	6
21	11	1	0
25	10	18	306
25	11	22	462
25	12	3	6
25	9	3	6
26	11	2	2
20	11	1	0

Two-locus Matches

The within-population matching proportion for the African-American sample is 28,366/115,940=0.245.

The within-population matching proportion for the Caucasian sample is 18,536/128,522=0.144.

The between-population matching proportion for the African-American and Caucasian samples is 8,347/122,419=0.068.

There is a clear decrease in matching between populations from within populations. We can establish some theory that describes these proportions.

Partial Matching

For autosomal markers, two profiles may be:

Match:	AA, AA or AB, AB
Partially Match:	AA, AB or AB, AC
Mismatch:	AA, BB or AA, BC or AB, CD

How likely are each of these?

Database Matching

If every profile in a database is compared to every other profile, each pair can be characterized as matching, partially matching or mismatching without regard to the particular alleles. We find the probabilities of these events by adding over all allele types.

The probability P_2 that two profiles match (at two alleles) is

$$\begin{aligned} P_2 &= \sum_A \text{Pr}(AA, AA) + \sum_{A \neq B} \text{Pr}(AB, AB) \\ &= \frac{\sum_A p_A [\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\ &\quad + \frac{2\sum_{A \neq B} [\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)} \end{aligned}$$

Database Matching

This approach leads to probabilities P_2, P_1, P_0 of matching at 2,1,0 alleles:

$$
\begin{aligned}
P_2 &= \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) \\
&\quad + (1-\theta)^3(2S_2^2 - S_4)]
\end{aligned}
$$

$$
\begin{aligned}
P_1 &= \frac{1}{D}[8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) \\
&\quad + 4(1-\theta)^3(S_2 - S_3 - S_2^2 + S_4)]
\end{aligned}
$$

$$
\begin{aligned}
P_0 &= \frac{1}{D}[\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) \\
&\quad + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)]
\end{aligned}
$$

where $D = (1+\theta)(1+2\theta)$, $S_2 = \sum_A p_A^2$, $S_3 = \sum_A p_A^3$, $S_4 = \sum_A p_A^4$. For any value of θ we can predict the matching, partially matching and mismatching proportions in a database.

FBI Caucasian Matching Counts

One-locus matches in FBI Caucasian data (18,721 pairs of 13-locus profiles).

		θ				
Locus	Observed	.000	.001	.005	.010	.030
D3S1358	.077	.075	.075	.077	.079	.089
vWA	.063	.062	.063	.065	.067	.077
FGA	.036	.036	.036	.038	.040	.048
D8S1179	.063	.067	.068	.070	.072	.083
D21S11	.036	.038	.038	.040	.042	.051
D18S51	.027	.028	.029	.030	.032	.040
D5S818	.163	.158	.159	.161	.164	.175
D13S317	.076	.085	.085	.088	.090	.101
D7S820	.062	.065	.066	.068	.070	.080
CSF1PO	.122	.118	.119	.121	.123	.134
TPOX	.206	.195	.195	.198	.202	.216
THO1	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

FBI Database Matching Counts

Matching		Number of Partially Matching Loci												
loci	θ	0	1	2	3	4	5	6	7	8	9	10	11	12
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30
1	Obs.	0	12	48	203	574	1133	1516	1596	1206	602	193	43	3
	.000	0	7	50	212	600	1192	1704	1768	1320	692	242	51	5
	.010	0	5	40	178	527	1094	1637	1779	1393	767	282	62	6
2	Obs.	0	7	61	203	539	836	942	807	471	187	35	2	
	.000	1	9	56	210	514	871	1040	877	511	196	45	5	
	.010	1	8	50	193	494	875	1096	969	593	239	57	6	
3	Obs.	0	6	33	124	215	320	259	196	92	16	1		
	.000	1	7	36	116	243	344	334	220	94	23	3		
	.010	0	6	35	117	256	380	387	268	120	32	4		
4	Obs.	1	5	17	29	54	82	67	16	6	0			
	.000	0	3	15	40	70	81	61	29	8	1			
	.010	0	3	15	44	81	98	78	40	12	1			
5	Obs.	0	1	2	6	12	14	6	5	0				
	.000	0	1	4	9	13	11	6	2	0				
	.010	0	1	4	11	16	15	9	3	0				
6	Obs.	0	1	0	2	2	0	0	0					
	.000	0	0	1	1	1	1	0	0					
	.010	0	0	1	2	2	1	1	0					

Predicted Matches when $n = 65,493$

Matching	Number of partially matching loci							
loci	0	1	2	3	4	5	6	7
6	4,059	37,707	148,751	322,963	416,733	319,532	134,784	24,125
7	980	7,659	24,714	42,129	40,005	20,061	4,150	
8	171	1,091	2,764	3,467	2,153	530		
9	21	106	198	163	50			
10	2	7	8	3				
11	0	0	0					
12	0	0						
13	0							

Y-chromosome Profiles

The Y-chromosome also has several STR markers that are useful in forensic science. In one respect, the profiles are easier to interpret as each man has only one allele at an STR locus. Otherwise interpretation is made more complicated by the lack of recombination on the Y chromosome, meaning that alleles at different loci are not independent. Or are they?

We expect that mutations act independently at different loci and this may counter the lack of recombination to some extent.

Y-STR Databases

There are three public databases of Y-STR profiles:

- Y-Chromosome Haplotype Reference Database (YHRD)

- Human Genome Diversity Project (HGDP)

- Data published by Xu et al. (XU)

Two-locus LD for Y-STR Loci

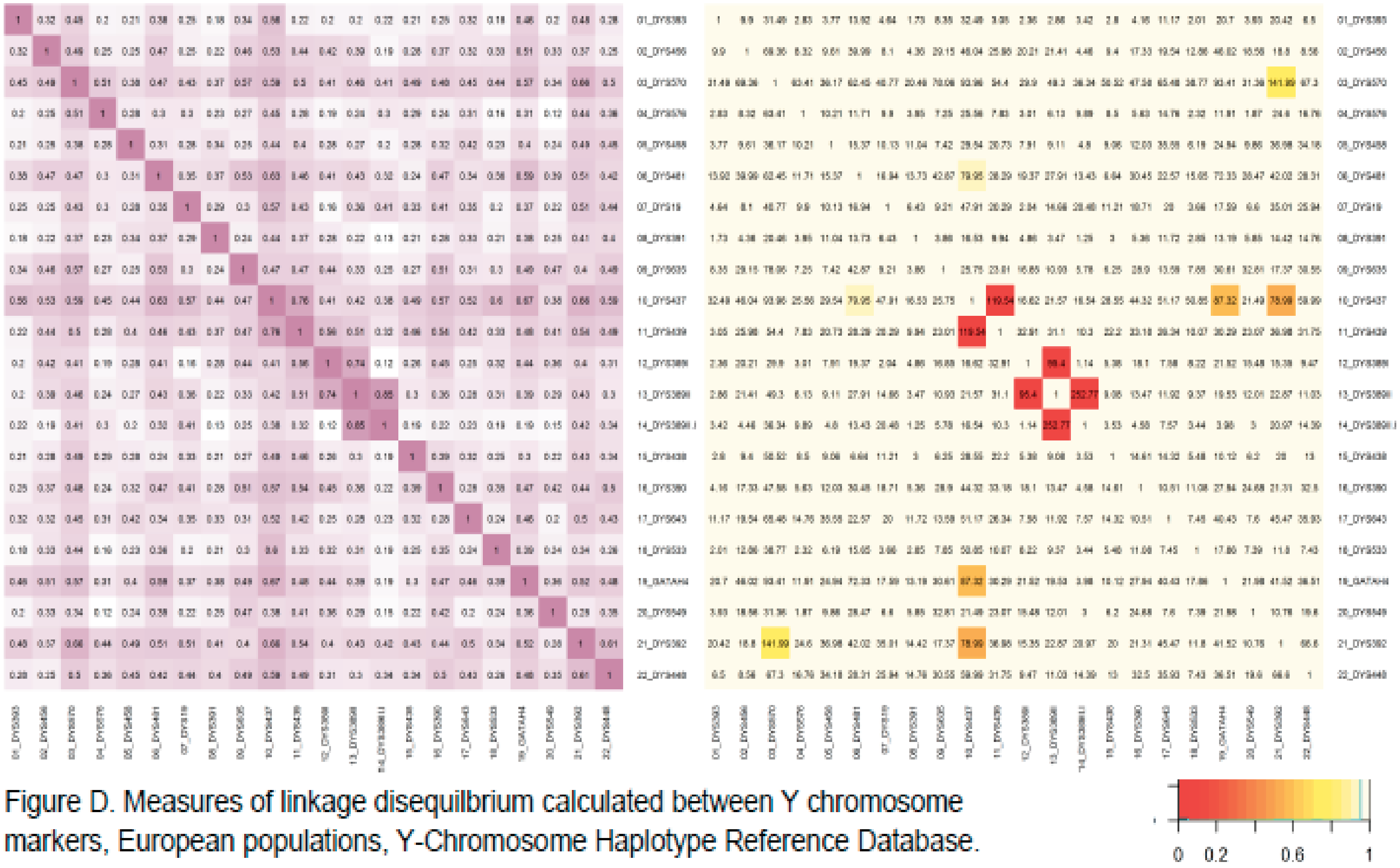


Figure D. Measures of linkage disequilibrium calculated between Y chromosome markers, European populations, Y-Chromosome Haplotype Reference Database.

Entropy

How do we measure independence among loci? The traditional measures of linkage disequilibrium don't work very well for multiple loci.

Instead we turn to entropy. For a single locus l, with alleles of type u having sample frequencies $\tilde{p}_{u_l}$, the entropy is

$$H_l \ = \ -\sum_u \tilde{p}_{u_l} \ln(\tilde{p}_{u_l})$$

How does this quantity behave? If there is only one allele, $u = 1, \tilde{p}_{1_l} = 1$ then $H_l = 0$. If there are m equally frequent alleles $\tilde{p}_{u_l} = 1/m, u = 1, 2, \ldots m$ then $H_l = -\ln(1/m) = \ln(m)$ and this gets larger as m gets larger. Entropy therefore indicates the amount of variation at the locus.

Two-locus Entropy

For haplotypes uv for alleles u, v at loci l, l' with sample frequencies $\tilde{p}_{u_l v_{l'}}$, the entropy is

$$H_{ll'} \; = \; -\sum_u \sum_v \tilde{p}_{u_l v_{l'}} \ln(\tilde{p}_{u_l v_{l'}})$$

If the two loci are independent: $\tilde{p}_{u_l v_{l'}} = \tilde{p}_{u_l} \tilde{p}_{v_{l'}}$ so, in this case,

$$H_{ll'} \; = \; H_l + H_{l'}$$

If the two loci are completely dependent: $\tilde{p}_{u_l v_{l'}} = \tilde{p}_{u_l} = \tilde{p}_{v_{l'}}$ so, in this case,

$$H_{ll'} \; = \; H_l = H_{l'}$$

Three-locus Entropy

For haplotypes uvw for alleles u, v, w at loci l, l', l'' with sample frequencies $\tilde{p}_{u_l v_{l'} w_{l''}}$, the entropy is

$$H_{ll'l''} \;=\; -\sum_u \sum_v \sum_w \tilde{p}_{u_l v_{l'} w_{l''}} \ln(\tilde{p}_{u_l v_{l'} w_{l''}})$$

If the third locus is independent of the first two loci: $\tilde{p}_{u_l v_{l'} w_{l''}} = \tilde{p}_{u_l v_{l'}} \tilde{p}_{w_{l''}}$ so, in this case,

$$H_{ll'l'} \;=\; H_{ll'} + H_{l''}$$

If the third locus is completely dependent on the first two loci: $\tilde{p}_{u_l v_{l'} w_{l''}} = \tilde{p}_{u_l v_{l'}} = \tilde{p}_{w_{l''}}$ so, in this case,

$$H_{ll'l'} \;=\; H_{ll'} = H_{l''}$$

Conditional Entropy

The conditional entropy for locus l' given loci l is

$$H_{l'|l} \;=\; H_{ll'} - H_l = \begin{cases} H_{l'} & l, l' \text{ independent} \\ 0 & l, l' \text{ dependent} \end{cases}$$

The conditional entropy for locus l'' given loci l, l' is

$$H_{l''|ll'} \;=\; H_{ll'l''} - H_{ll'} = \begin{cases} H_{l''} & ll', l'' \text{ independent} \\ 0 & ll', l'' \text{ dependent} \end{cases}$$

A locus independent of the haplotype of the previous loci adds its own entropy to the entropy of the haplotype of previous loci.

A locus completely dependent of the previous loci adds nothing to the haplotype of the previous loci.

Constructing Y-STR Haplotypes

Entropy values let us build up haplotypes by adding the most informative loci first: i.e. add loci to maximize the entropy at each stage.

If we have a haplotype with L loci, add the next locus with the maximum entropy conditional on the haplotype. Need to consider all possible haplotypes for a specific set of loci.

Constructing Haplotypes with Maximum Entropy

For each Y-STR marker l with allele u_l sample frequencies $\tilde{p}_{u_l}$ form the entropies $H_l = -\sum_u \tilde{p}_{u_l} \ln(\tilde{p}_{u_l})$. Find the largest of these, and choose that marker ("1") to be the first one. Its entropy is H_1.

For each of the other $L-1$ markers form the two-locus entropies H_{1l} with the two-locus haplotype u_1v_l frequencies $\tilde{p}_{u_1v_l}$:

$$H_{1l} \;=\; -\sum_u \sum_v \tilde{p}_{u_1v_l} \ln(\tilde{p}_{u_1v_l})$$

and then form the $L-1$ conditional entropies $H_{l|1} = H_{1l} - H_1$.

Choose the marker "2" with the largest conditional entropy to be the second one selected. The combined entropy is H_{12}

$$H_{12} \;=\; -\sum_u \sum_v \tilde{p}_{u_1v_2} \ln(\tilde{p}_{u_1v_2})$$

Constructing Haplotypes with Maximum Entropy

For each of the $L-2$ remaining markers form the three-locus entropies H_{12l} with the three-locus haplotype $u_1v_2w_l$ frequencies $\tilde{p}_{u_1v_2w_l}$:

$$H_{12l} \;=\; -\sum_u\sum_v\sum_w \tilde{p}_{u_1v_2w_l}\ln(\tilde{p}_{u_1v_2w_l})$$

and then form the $L-2$ conditional entropies

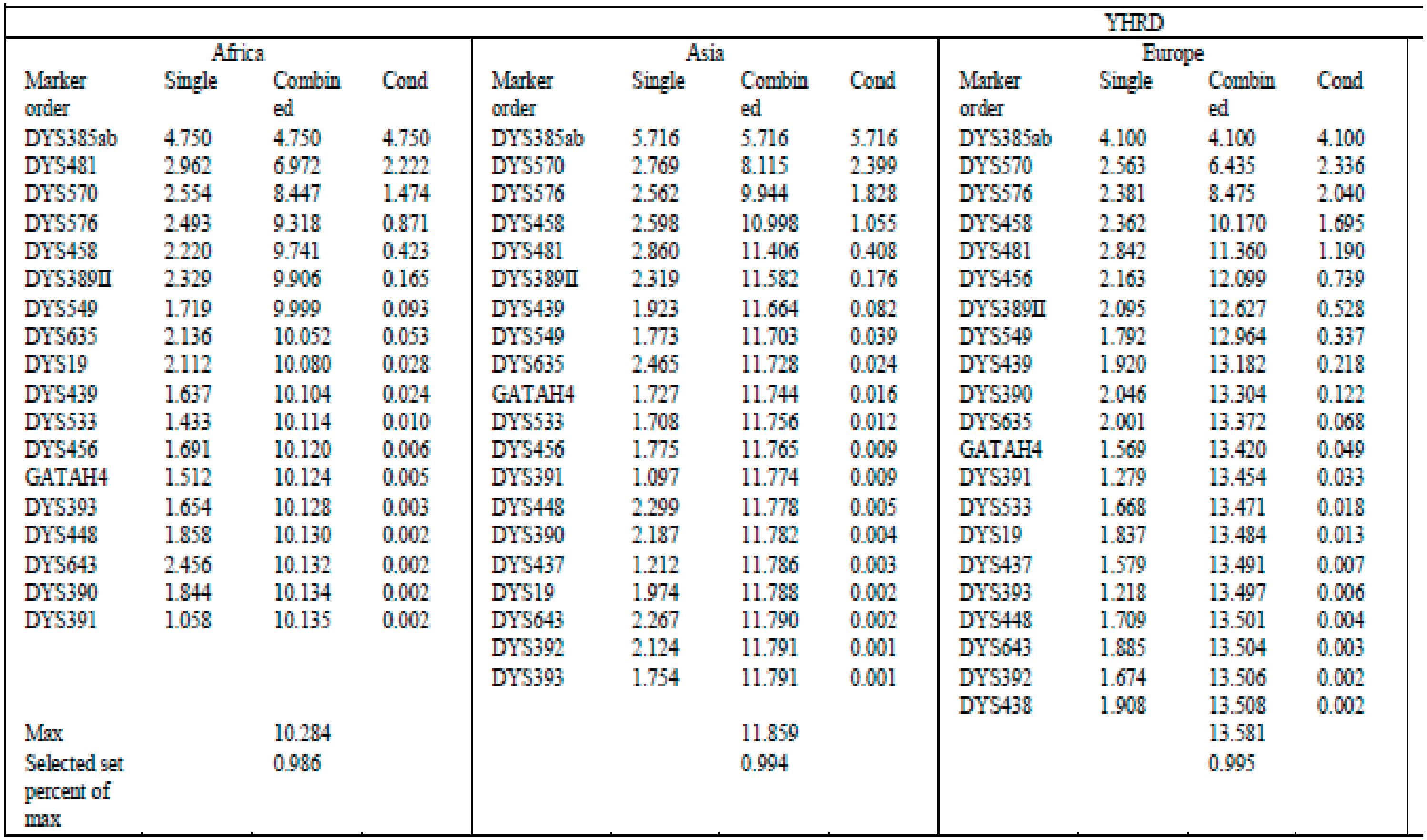
$$H_{l|12} \;=\; H_{12l} - H_{12}$$

Choose the marker "3" with the largest conditional entropy to be the third one selected. The combined entropy is H_{123}

$$H_{123} \;=\; -\sum_u\sum_v\sum_w \tilde{p}_{u_1v_2w_3}\ln(\tilde{p}_{u_1v_2w_3})$$

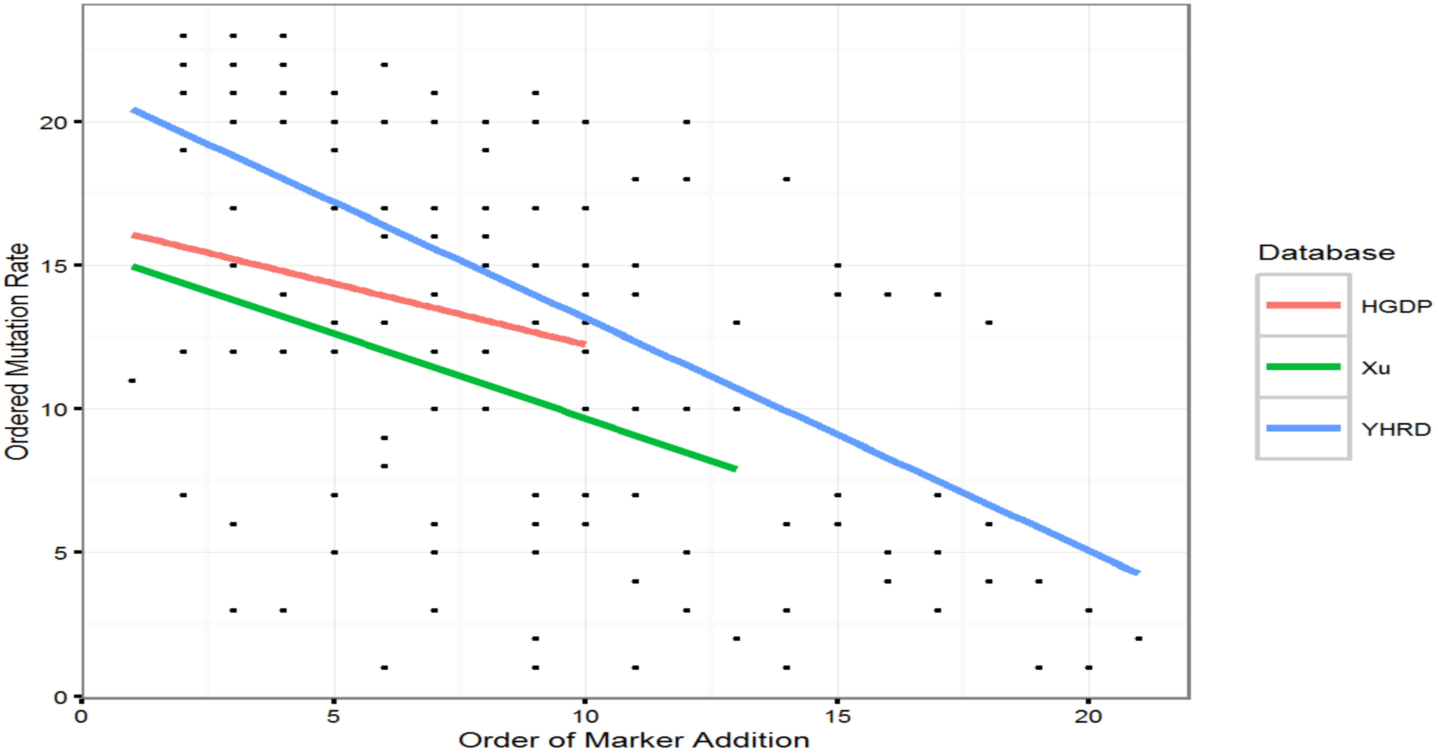
etc.

Examples



									YHRD		
Africa				Asia				Europe			
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.750	4.750	4.750	DYS385ab	5.716	5.716	5.716	DYS385ab	4.100	4.100	4.100
DYS481	2.962	6.972	2.222	DYS570	2.769	8.115	2.399	DYS570	2.563	6.435	2.336
DYS570	2.554	8.447	1.474	DYS576	2.562	9.944	1.828	DYS576	2.381	8.475	2.040
DYS576	2.493	9.318	0.871	DYS458	2.598	10.998	1.055	DYS458	2.362	10.170	1.695
DYS458	2.220	9.741	0.423	DYS481	2.860	11.406	0.408	DYS481	2.842	11.360	1.190
DYS389II	2.329	9.906	0.165	DYS389II	2.319	11.582	0.176	DYS456	2.163	12.099	0.739
DYS549	1.719	9.999	0.093	DYS439	1.923	11.664	0.082	DYS389II	2.095	12.627	0.528
DYS635	2.136	10.052	0.053	DYS549	1.773	11.703	0.039	DYS549	1.792	12.964	0.337
DYS19	2.112	10.080	0.028	DYS635	2.465	11.728	0.024	DYS439	1.920	13.182	0.218
DYS439	1.637	10.104	0.024	GATAH4	1.727	11.744	0.016	DYS390	2.046	13.304	0.122
DYS533	1.433	10.114	0.010	DYS533	1.708	11.756	0.012	DYS635	2.001	13.372	0.068
DYS456	1.691	10.120	0.006	DYS456	1.775	11.765	0.009	GATAH4	1.569	13.420	0.049
GATAH4	1.512	10.124	0.005	DYS391	1.097	11.774	0.009	DYS391	1.279	13.454	0.033
DYS393	1.654	10.128	0.003	DYS448	2.299	11.778	0.005	DYS533	1.668	13.471	0.018
DYS448	1.858	10.130	0.002	DYS390	2.187	11.782	0.004	DYS19	1.837	13.484	0.013
DYS643	2.456	10.132	0.002	DYS437	1.212	11.786	0.003	DYS437	1.579	13.491	0.007
DYS390	1.844	10.134	0.002	DYS19	1.974	11.788	0.002	DYS393	1.218	13.497	0.006
DYS391	1.058	10.135	0.002	DYS643	2.267	11.790	0.002	DYS448	1.709	13.501	0.004
				DYS392	2.124	11.791	0.001	DYS643	1.885	13.504	0.003
				DYS393	1.754	11.791	0.001	DYS392	1.674	13.506	0.002
								DYS438	1.908	13.508	0.002
Max		10.284				11.859				13.581	
Selected set percent of max		0.986				0.994				0.995	

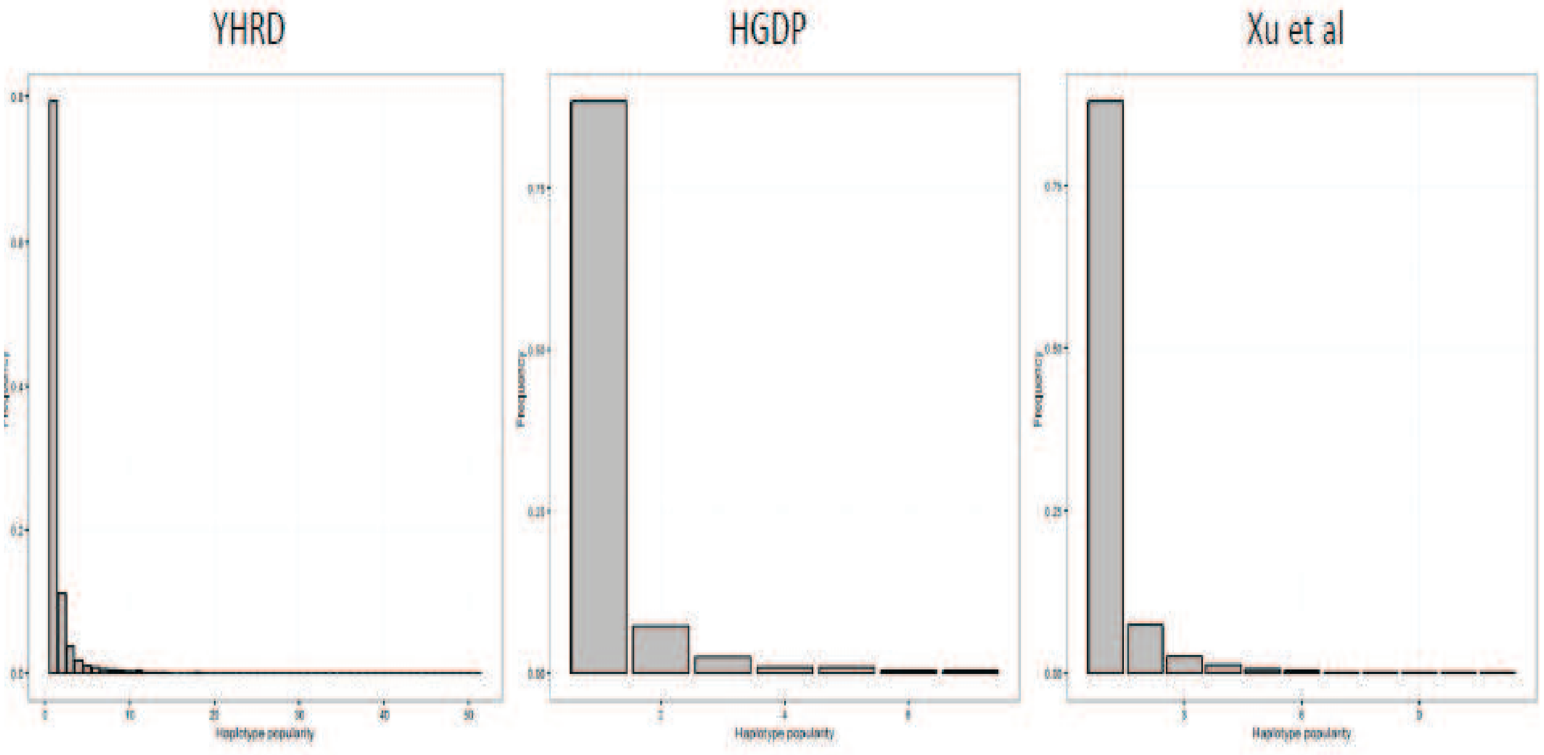
Examples


Ordered Mutation Rate
0
5
10
15
20
Order of Marker Addition
0
5
10
15
20
Database
HGDP
Xu
YHRD

Counting Method

The problem is that, with 10 or more alleles at a locus, the number of possible profiles quickly exceeds the database size as the number of loci increases. It is not uncommon for a profile of interest not to appear in a database, although some profiles do reach appreciable numbers as is shown for 12-locus profiles on the next slide. Eventually, however, the number of loci makes it likely that a profile occurs once or not all in a database unless profiles from close relatives such as father-son or brothers are present.

Counting Method: 12-locus Profiles



Exact Confidence Intervals

Exact confidence limits follow from the binomial distribution. For events with low probabilities p, how large could p be for there to be at least a 5% chance of seeing no more than x (i.e. $0, 1, 2, \ldots x$) occurrences of that event among n events. If this upper bound is p_U,

$$\begin{aligned} \sum_{k=0}^{x} \Pr(k) &\geq 0.05 \\ \sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} &\geq 0.05 \end{aligned}$$

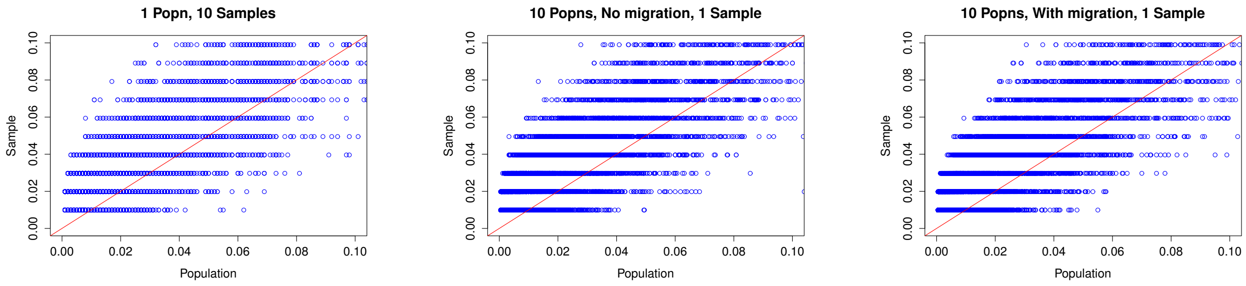
If $x = 0$, then $(1-p_U)^n \geq 0.05$ or $p_U \leq 1 - 0.05^{1/n}$ and this is 0.0295 if $n = 100$. More generally $p_U \approx 3/n$ when $x = 0$, but the same result holds if the profile is based on 7 or 17 or 27 loci.

Sampling Effects

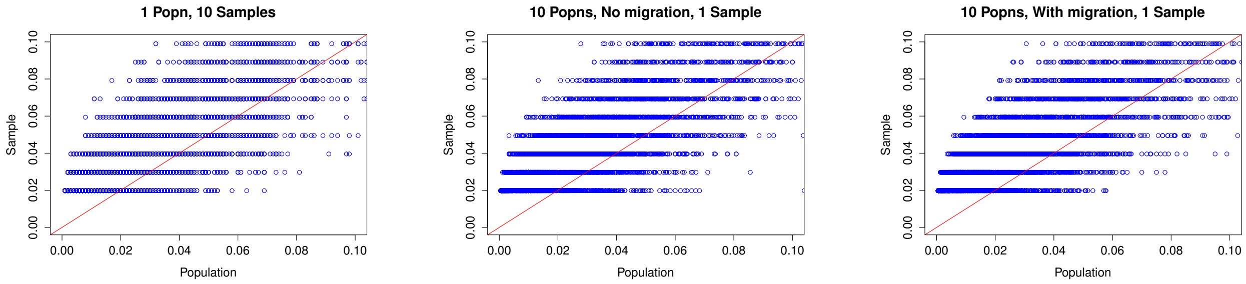
To illustrate the behavior of the sample proportion of a haplotype in samples of n haplotypes from populations of N haplotypes, we simulated a finite random-mating population of N haplotypes. The founding population consisted of N unique haplotypes. The subsequent loss of variation was countered by stepwise mutation, one repeat unit in each direction. Loci were completely linked, but underwent mutation independently. Loci were completely linked, but underwent mutation independently.

On the next slides we show results for $N = 1000$ and $n = 100$. There were 10 loci, each with mutation rate $\mu = 5 \times 10^{-4}$. The migration rate was either $m = 0$ or $m = 0.1$.

Sampling Effects



Sample and population frequencies for each profile in the population.



Sample and population frequencies for each profile in the sample.

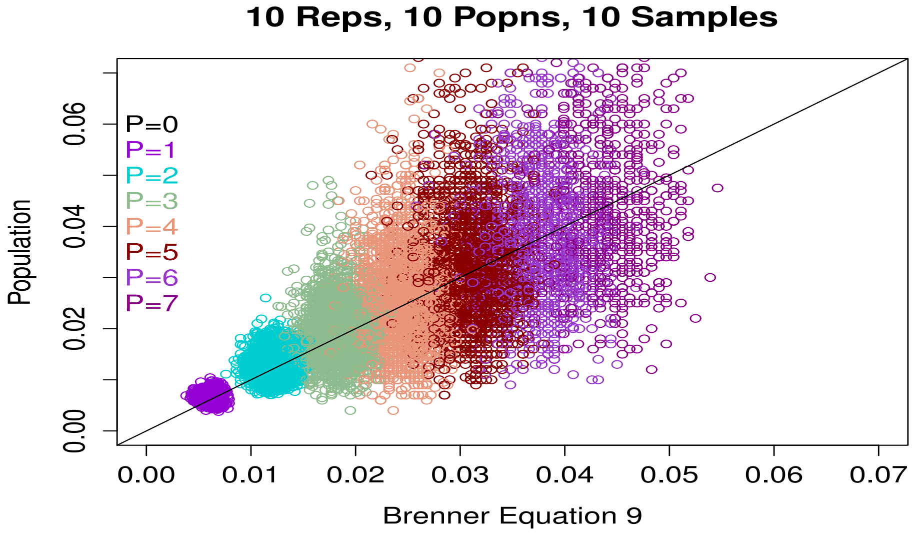
Brenner's Method

Brenner (2010) proposed the use of the proportion κ of profiles that occurred only once in a database that had been augmented by the evidentiary profile. His approach did not require a genetic model, although κ values can be predicted for some genetic models. The probability of a person taken randomly from a population would have the same profile as the evidentiary type when that type was not present in a sample of size $(n-1)$ (i.e. occurred once in the sample augmented by the evidentiary profile) was given by $(1-\kappa)/n$.

For profiles that occur p times in the augmented sample (those with "popularity" p), Brenner suggested a modification to $p(1-\kappa)/n$ that approaches the sample proportion $\tilde{p}$ when the proportion of singletons in the database becomes small.

Brenner's Method

Here we compare Brenner's estimates for every profile in the augmented database with the proportion of profiles of that type in the population from which the sample was drawn. Brenner's values appear better than the sample proportions for profiles not seen in the sample before it was augmented, as desired by Brenner. The quality decreases as the sample proportion of the evidentiary profile increases.



Genetic Model

A genetic approach can be built on the notion of identity by descent. For large numbers of loci, profiles of the same type are likely to match because they have a common ancestral haplotype. If θ_i is the probability of identity by descent of two random haplotypes in population i, the probability a random profile in population i is of type A given the evidentiary profile, also from population i, is that type is $\Pr(A|A)_i = \theta_i + (1-\theta_i)p_{Ai}$.

As profile proportions p_{Ai} become small the matching probabilities approach θ_i. These quantities, in turn, decrease as the number of loci increases. Kimura and Ohta (1968) showed that, for single-step mutations, STR loci have predicted θ values of $1/\sqrt{1+4N\mu}$. For L loci undergoing independent mutation we could replace μ by $1-(1-\mu)^L \approx L\mu$.

Y-STR Matches

The chance of a random man having Y-STR haplotype A is written as p_A, the profile probability.

The chance that two men have haplotype A is written as P_{AA}.

The chance that a man has haplotype A given that another man has been seen to have that profile is $P_{A|A}$, the match probability. The three quantities are related by $P_{A|A} = P_{AA}/p_A$.

A major difficulty is that we generally do not have samples from the relevant (sub)population to give us estimates of p_A or P_{AA}. Instead we have a database of profiles that may represent a larger population.

Interpreting Evidence

Two hypotheses for observed match between suspect and evidence:

H_P: Suspect is source of evidence.
H_D: Suspect is not source of evidence.

Then

$$\frac{\Pr(H_P|\text{Match})}{\Pr(H_D|\text{Match})} = \frac{\Pr(\text{Match}|H_P)}{\Pr(\text{Match}|H_D)} \times \frac{\Pr(H_P)}{\Pr(H_D)}$$

Interpreting Evidence

Suppose matching Y-STR profile is type A. The likelihood ratio reduces to

$$\begin{aligned}\frac{\text{Pr}(\text{Match}|H_P)}{\text{Pr}(\text{Match}|H_D)} &= \frac{\text{Pr}(A|A,H_P)}{\text{Pr}(A|A,H_D)} \\ &= \frac{1}{\text{Pr}(A|A)}\end{aligned}$$

A population genetic model introduces the quantity θ:

$$\begin{aligned}\text{Pr}(AA) &= \theta p_A + (1-\theta)p_A^2 \\ \text{Pr}(A|A) &= \theta + (1-\theta)p_A\end{aligned}$$

where θ is the probability that two profiles are identical by descent.

Sample Within-population Matching

If the sample from population i has n_{Ai} copies of allele (or haplotype) A, and these sum to n_i, the sample within-population matching proportion for this population is

$$\begin{aligned}
\tilde{M}_i &= \frac{1}{n_i(n_i-1)}\sum_A n_{Ai}(n_{Ai}-1) \\
&= \frac{n_i}{n_i-1}\sum_A \frac{n_{Ai}}{n_i}\left(\frac{n_{Ai}}{n_i}-\frac{1}{n_i}\right) \\
&= \frac{n_i}{n_i-1}\left(\sum_A \tilde{p}_{Ai}^2-\frac{1}{n_i}\right)
\end{aligned}$$

Averaging over populations:

$$\tilde{M}_W = \frac{1}{r}\sum_{i=1}^{r}\tilde{M}_i$$

Sample Between-population Matching

The sample between-population matching proportion for populations i and j is

$$\begin{aligned}\tilde{M}_{ij} &= \frac{1}{n_i n_j}\sum_A n_{Ai} n_{Aj} \\ &= \sum_A \frac{n_{Ai}}{n_i}\frac{n_{Aj}}{n_j} \\ &= \sum_A \tilde{p}_{Ai}\tilde{p}_{Aj}\end{aligned}$$

Averaging over pairs of populations:

$$\tilde{M}_B = \frac{1}{r(r-1)}\sum_{i\neq j}^{r} \tilde{M}_{ij}$$

One-locus NIST Y-STR Estimates

Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\widehat{\beta}_W$
DYS19	0.32571062	0.24309148	0.10915340
DYS385a/b	0.07982377	0.04427420	0.03719640
DYS389I	0.41279418	0.38319082	0.04799436
DYS389II	0.26072434	0.23741323	0.03056847
DYS390	0.28981997	0.18813203	0.12525182
DYS391	0.52191425	0.48517426	0.07136392
DYS392	0.39961865	0.35168087	0.07394164
DYS393	0.50285122	0.48769253	0.02958906
DYS437	0.46400112	0.38595032	0.12710828
DYS438	0.36817530	0.23212655	0.17717601
DYS439	0.35507469	0.34990863	0.00794667
DYS448	0.30091326	0.22640195	0.09631787
DYS456	0.33444029	0.32578009	0.01284478
DYS458	0.21642167	0.19701369	0.02416976
DYS481	0.18867019	0.14121936	0.05525373
DYS533	0.39365769	0.37177174	0.03483757
DYS549	0.33976578	0.30691346	0.04740003
DYS570	0.21298105	0.20775666	0.00659442
DYS576	0.20955290	0.18125443	0.03456321
DYS635	0.27720127	0.20653182	0.08906400
DYS643	0.28394262	0.20058158	0.10427710
Y-GATA-H4	0.40667782	0.39899963	0.01277568

Multiple-locus US-YSTR Estimates

No. Loci	Added Locus	$\widetilde{M}_W$	$\widetilde{M}_B$	$\widehat{\beta}_W$
1	DYS_438	0.37903281	0.27283973	0.14603806
2	DYS_392	0.22353526	0.10233258	0.13501958
3	DYS_19	0.11294942	0.05471374	0.06160639
4	DYS_390	0.05923470	0.02393636	0.03616398
5	DYS_643	0.04798422	0.02456341	0.02401059
6	YGATA_C4	0.03119210	0.01541060	0.01602851
7	DYS_533	0.01979150	0.00777794	0.01210774
8	DYS_393	0.01482393	0.00650531	0.00837309
9	DYS_456	0.01073170	0.00396487	0.00679377
10	DYS_438	0.00889934	0.00287761	0.00603912
11	DYS_549	0.00524369	0.00123093	0.00401770
12	DYS_481	0.00317518	0.00055413	0.00262250
13	DYS_389I	0.00240161	0.00031517	0.00208710
14	DYS_391	0.00200127	0.00017039	0.00183119
15	DYS_576	0.00106995	0.00005877	0.00101124
16	DYS_ 389II	0.00089896	0.00004205	0.00085695
17	DYS_385	0.00065020	0.00002729	0.00062293
18	YGATA_H4	0.00063652	0.00002427	0.00061227
19	DYS_448	0.00055062	0.00000713	0.00054349
20	DYS_458	0.00051100	0.00000423	0.00050677
21	DYS_570	0.00043010	0.00000423	0.00042587
22	DYS_439	0.00038612	0.00000423	0.00038189

Y-STR Match Probabilities

Within subpopulation i, if we knew the haplotype frequencies and if we assumed random mating, the chance that two unrelated men have haplotype A is p_{Ai}^2 and the match probability is just the profile probability p_{Ai}.

If we allow for the evolutionary variation that led to the current subpopulation, then the total population allele frequencies p_A can be used when the specific population frequencies p_{Ai} are not known:

$$P_{AAi} \;=\; \theta_i p_A + (1-\theta_i)p_A^2$$

Estimating Match Proportions

To take account of what $\sum_A \tilde{p}_A^2$ is actually estimating, we form an estimate of the within subpopulation matching as

$$\widehat{M}_W = \beta_W + (1-\beta_W)\sum_A \tilde{p}_A^2$$

where

$$\beta_W = \frac{\theta_W - \theta_B}{1-\theta_B}$$

This is the "theta" for the "theta-correction" expression for match probabilities.

When there are data from the subpopulations, it has a simple estimate:

$$\widehat{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1-\tilde{M}_B}$$

Inbreeding and Relatedness

Matching probabilities for DNA evidence can be affected quite substantially when the people whose profile are (or may be) related.

To begin to consider how to approach this question, we start by thinking about inbreeding.

Inbreeding

Inbreeding occurs when a person receives (copies of) the same allele from the same ancestral allele. The most likely example in people is for children of marriages between first cousins. First cousins have parents who are siblings, so they have two grandparents in common, and they might pass on the same one of the four alleles these two grandparents have to their child.

Cousin Marriages

“The United States has the only bans on cousin marriage in the Western world. As of February 2010, 30 U.S. states prohibit most or all marriages between first cousins, and a bill is pending in Maryland which would prohibit most first cousins from marrying there. Six states prohibit first-cousin-once-removed marriages. Some states prohibiting cousin marriage recognize cousin marriages performed in other states, but despite occasional claims that this holds true in general, laws also exist that explicitly void all foreign cousin marriages or marriages conducted by state residents out of state.”

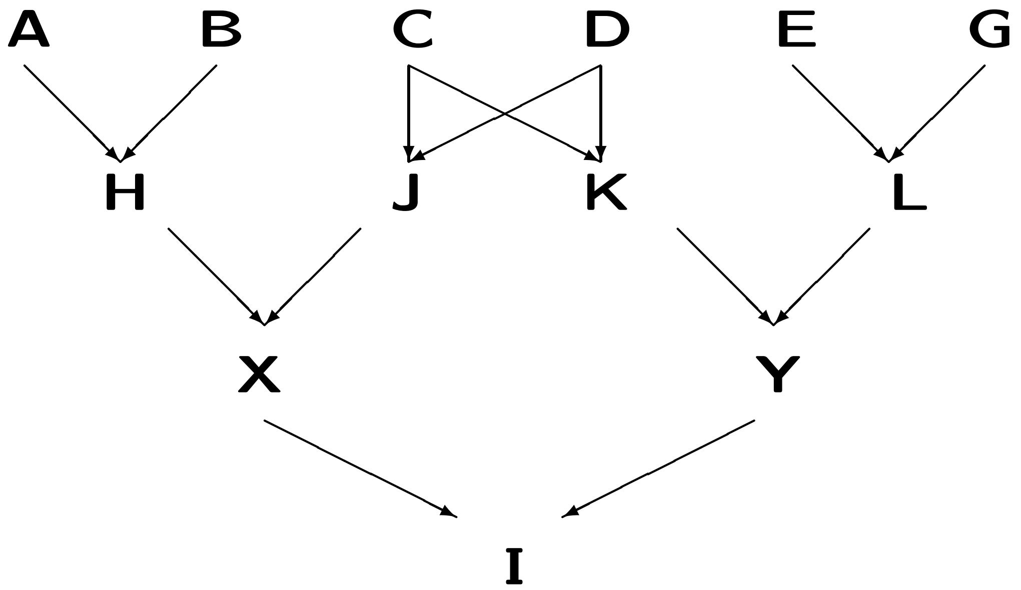
Wikipedia, “Cousin Marriages”

Cousin Marriages

“Twenty-five states prohibit marriages between first cousins. Six states allow first cousin marriage under certain circumstances, and North Carolina allows first cousin marriage but prohibits double-cousin marriage. States generally recognize marriages of first cousins married in a state where such marriages are legal.”

http://www.ncsl.org/research/human-services/state-laws-regarding-marriages-between-first-cousi.aspx

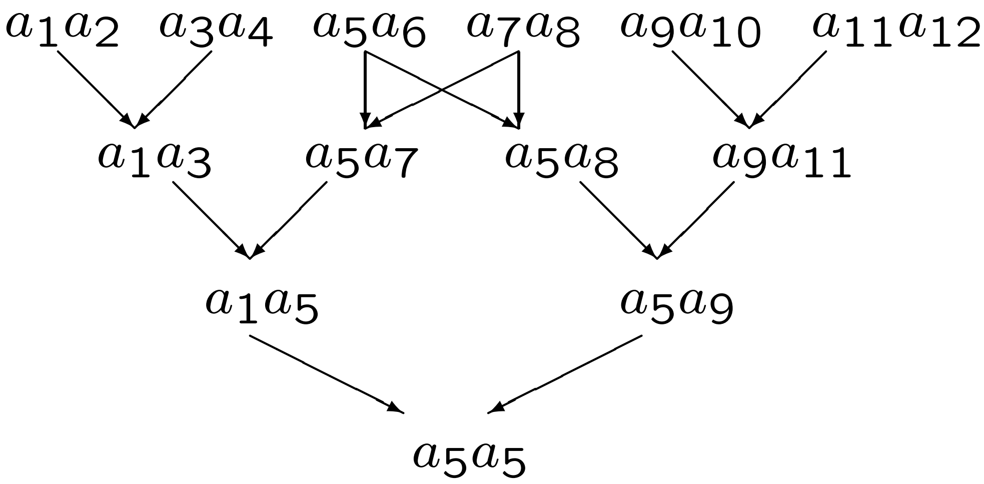
Offspring of First cousins



X, Y are first cousins. J, K are full sibs. C, D are the grandparents in common to X, Y. Because X, Y are related, individual I is inbred.

The next slide shows a possible set of genotypes for this pedigree.

Offspring of First cousins



Individual I in this case has received two (identical) copies of the same allele a_5.

In general, each of the child's two alleles has descended from one of 8 grandparental alleles: 64 possible combinations. Of these combinations, 4 involve two copies of the same allele. The probability the child is inbred is 4/64=1/16.

Cousin Marriage Inbreeding Coefficient

Using the grandparental alleles as numbered on the previous slide (but ignoring all the other genotypes shown there) the 64 allelic combinations have 4 that involve an identical pair. Such a pair are identical by descent (ibd). Inbreeding coefficient F is probability of ibd.

		First allele							
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
Second allele	a_5	—	—	—	—	ibd	—	—	—
	a_6	—	—	—	—	—	ibd	—	—
	a_7	—	—	—	—	—	—	ibd	—
	a_8	—	—	—	—	—	—	—	ibd
	a_9	—	—	—	—	—	—	—	—
	a_{10}	—	—	—	—	—	—	—	—
	a_{11}	—	—	—	—	—	—	—	—
	a_{12}	—	—	—	—	—	—	—	—

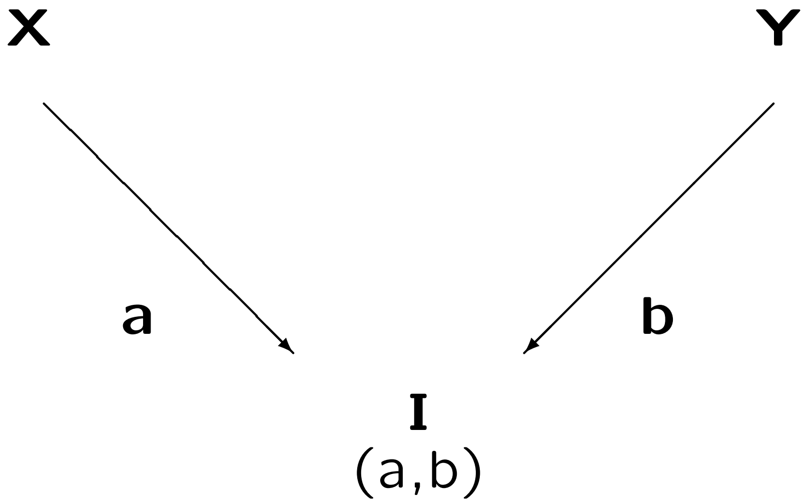
Coancestry

The inbreeding coefficient of an individual is the probability that the two alleles *going to* that individual are ibd. The inbreeding coefficient of individual I is written as F_I

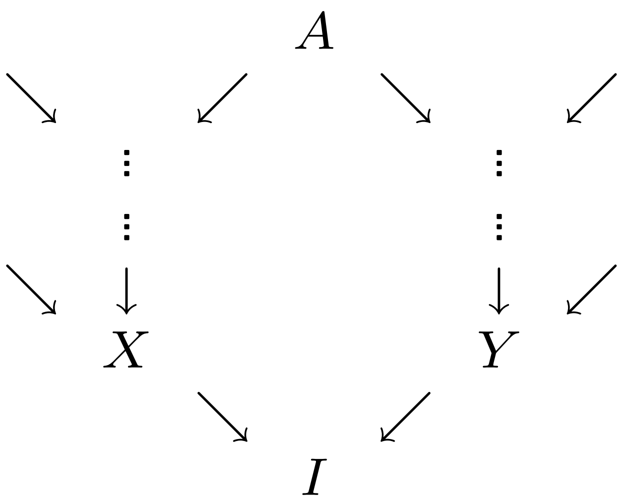
In other words, the two alleles *coming from* the two parents are ibd. The probability that two alleles, one chosen randomly from each of two individuals, are ibd is the coancestry of those two individuals. The coancestry of individuals X and Y is written as θ_{XY}.

Relatedness

The inbreeding coefficient of an individual is the coancestry of its parents; $F_I = \theta_{XY}$.



Path Counting



Identify the path linking the parents of I to their common ancestor(s).

Path Counting

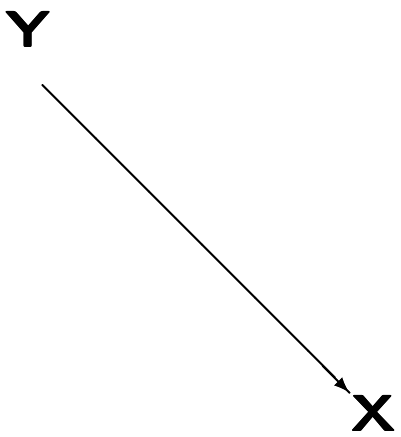
If the parents X, Y of an individual I have ancestor A in common, and if there are n individuals (including X, Y) in the path linking the parents through A, then the inbreeding coefficient of I, or the coancestry of X and Y, is

$$F_I = \theta_{XY} \quad = \quad \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several paths to an ancestor, sum over all paths.

If there are several ancestors, this expression is summed over all the ancestors.

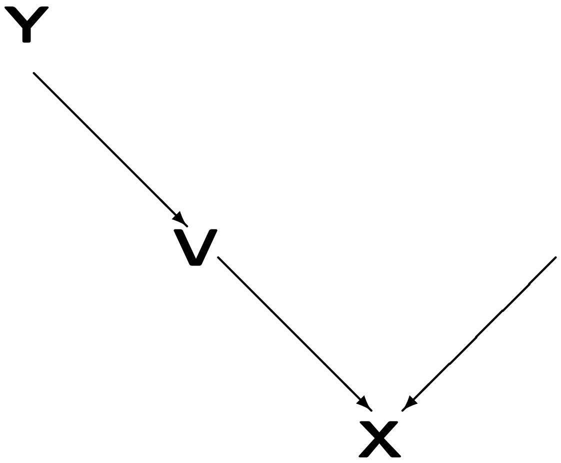
Parent-Child



The common ancestor of parent X and child Y is X. The path linking X, Y to their common ancestor is YX and this has $n = 2$ individuals. Therefore

$$\theta_{XY} \quad = \quad \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

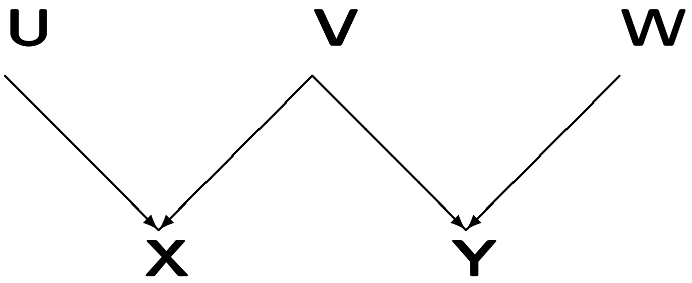
Grandparent-grandchild



The common ancestor of grandparent X and grandchild Y is X. The path linking X, Y to their common ancestor is YVX and this has $n = 3$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

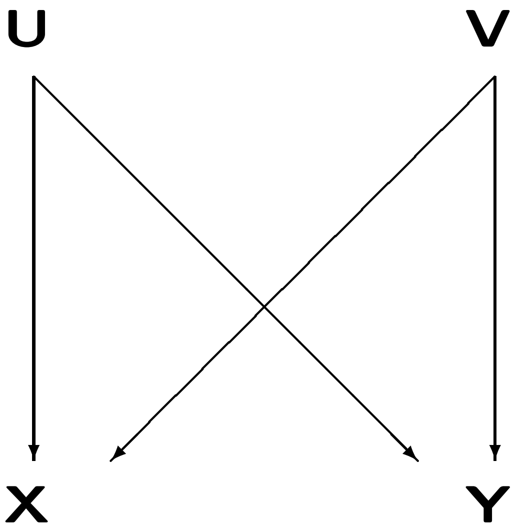
Half sibs



The common ancestor of half sibs X and Y is V. The path linking X, Y to their common ancestor is XVY and this has $n = 3$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

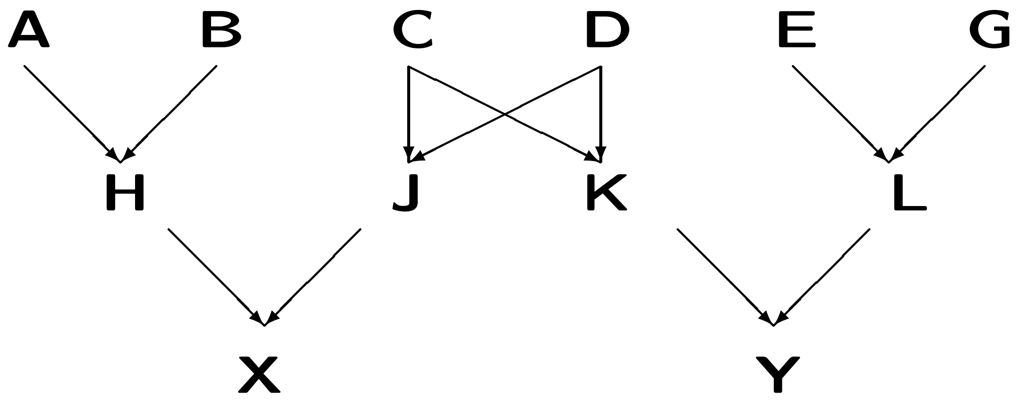
Full sibs



The common ancestors of full sibs X and Y are U and V. The paths linking X, Y to their common ancestors are XUY and XVY and these each have $n = 3$ individuals. Therefore

$$\theta_{XY} \; = \; \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

First cousins

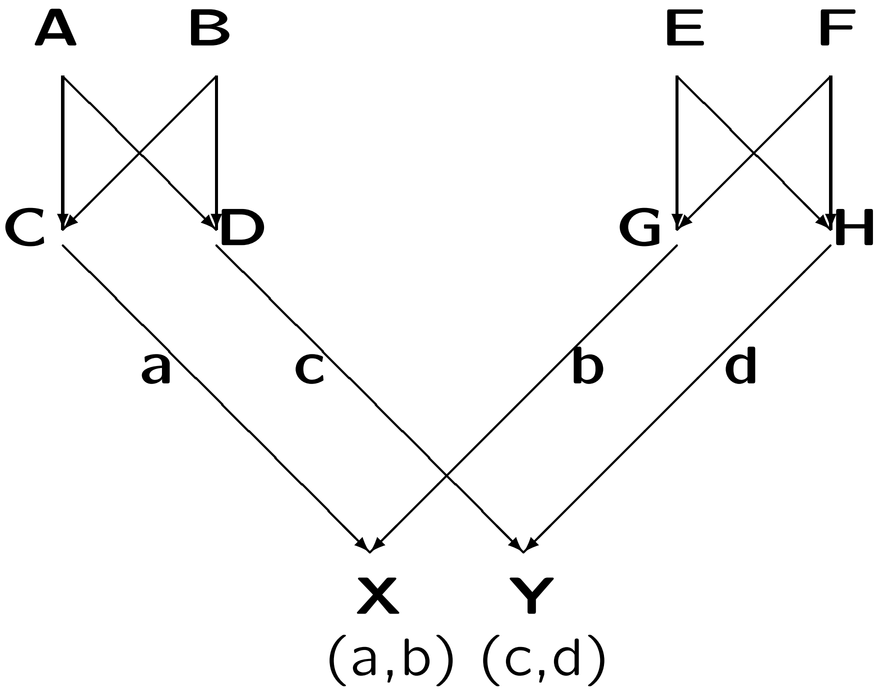


The common ancestors of cousins X and Y are C and D. The paths linking X, Y to their common ancestors are $XJCKY$ and $XJDKY$ and these each have $n = 5$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{16}$$

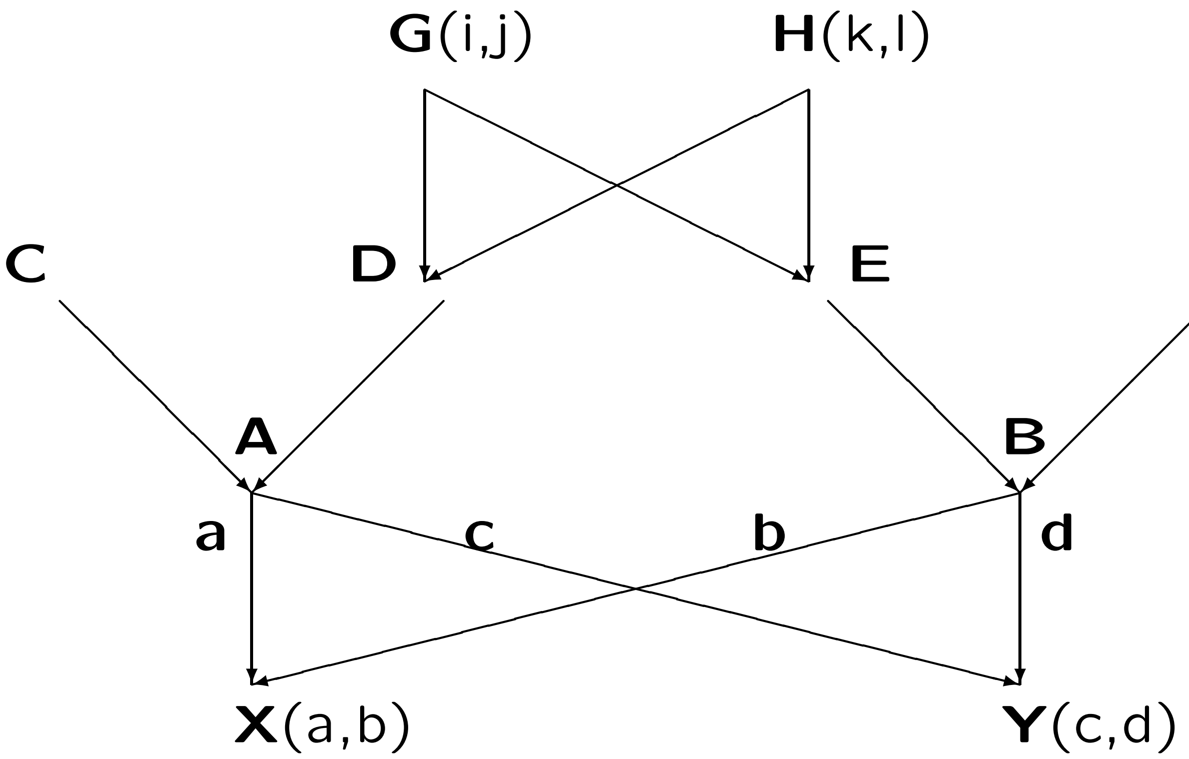
Double First Cousins

If two brothers C, D marry two sisters G, H, their children X, Y are both maternal and paternal first cousins: i.e. they are double first cousins. What is the coancestry coefficient of double first cousins?



Siblings whose Parents are Cousins

If two first cousins, A, B, marry and have two children X, Y, what is the coancestry coefficient of those children?



Genotype frequencies

Suppose individuals in a population all have inbreeding coefficients F. The probability an individual has two ibd alleles is F, and the probability of two non-ibd alleles is $(1 - F)$. The probability that any allele is type A is p_A, the population allele frequency. So, the probability an individual is homozygous is

$$P_{AA} \;=\; F \times p_A + (1 - F) \times p_A^2$$

Genotype frequencies

To emphasize the difference from Hardy-Weinberg:

$$P_{AA} \;=\; p_A^2 + Fp_A(1-p_A)$$

Because heterozygous individuals must have non-ibd alleles:

$$\begin{aligned} P_{Aa} &= 2(1-F)p_Ap_a \\ &= 2p_Ap_a - 2Fp_Ap_a \end{aligned}$$

These are profile probabilities, not match probabilities. The quantity F is not the θ in the match probability equations.

First Cousin Example

If every person in the population had parents who were first cousins, $F = 1/16 = 0.0625$. For a locus with allele frequencies $\{p_i\}$ that were all 0.10:

$$\begin{aligned} P_{ii} &= (0.1)^2 + 0.0625(0.1)(0.10) = 0.015625 \\ P_{ij} &= 2(0.1)(0.1) - 2(0.0625)(0.1)(0.1) = 0.018750 \end{aligned}$$

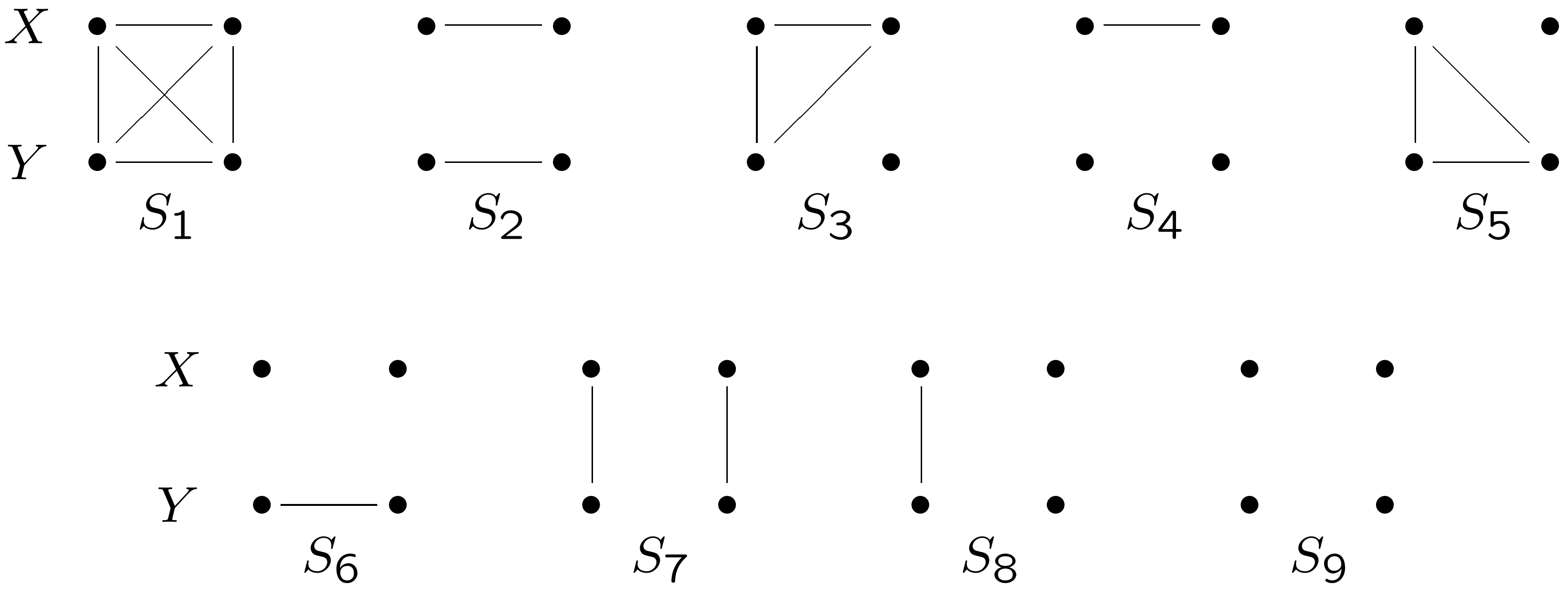
Probabilities of Pairs of Relatives

The inbreeding coefficient F is a statement about a pair of alleles, and it provides genotypic frequencies – the frequencies of pairs of alleles.

What about pairs of individuals? Their joint genotypic frequencies must involve four-allele analogs of F.

Nine-parameter IBD Set

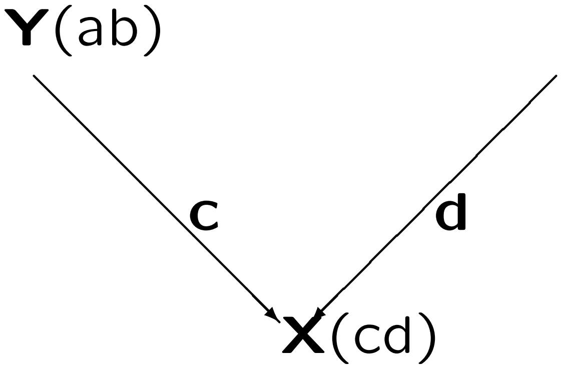
Solid lines join pairs of ibd alleles: top row is the pair of alleles for X, bottom row the pair of alleles for Y.



Non-inbred Relatives

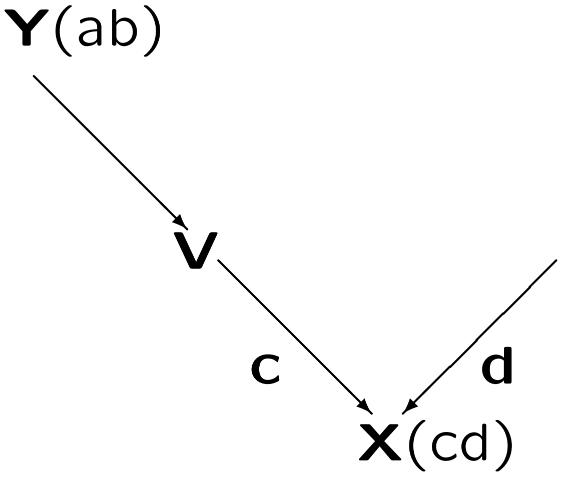
There is a reduction when neither individual is inbred, as then neither a, b nor c, d are ibd. There are then only three states and the three probabilities are often written as $k_2 = \Delta_7, k_1 = \Delta_8$ or $k_0 = \Delta_9$ to indicate the number of pairs of ibd alleles carried by the two individuals. Examples follow:

Parent-Child



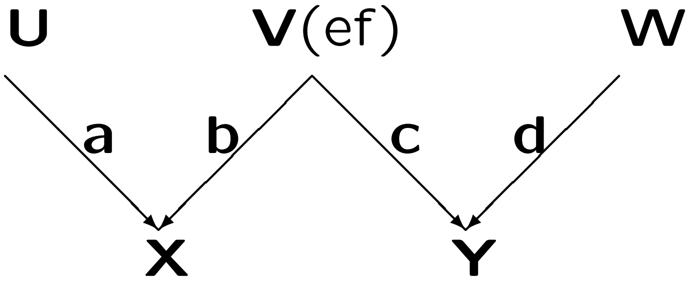
$$\Pr(c \equiv a) = 0.5, \quad \Pr(c \equiv b) = 0.5, \quad k_1 = 1$$

Grandparent-grandchild



$$\Pr(c \equiv a) = 0.25, \quad \Pr(c \equiv b) = 0.25, \quad k_1 = 0.5 \& k_0 = 0.5$$

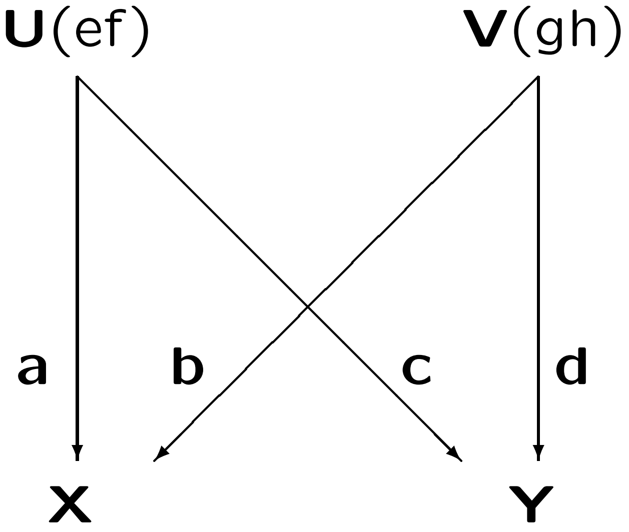
Half sibs



		0.5	0.5
		$c \equiv e$	$c \equiv f$
0.5	$b \equiv e$	0.25	0.25
0.5	$b \equiv f$	0.25	0.25

Therefore $k_1 = 0.5$ so $k_0 = 0.5$.

Full sibs



U(ef)

V(gh)

a

b

c

d

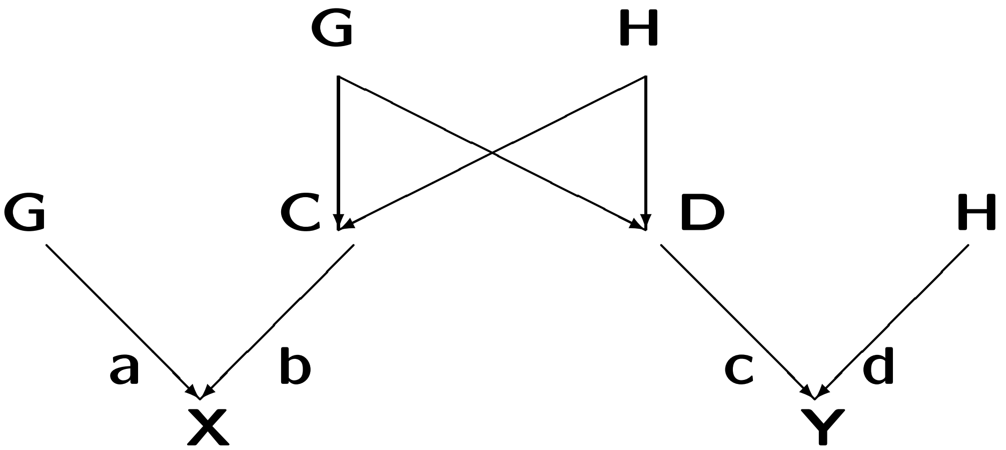
X

Y

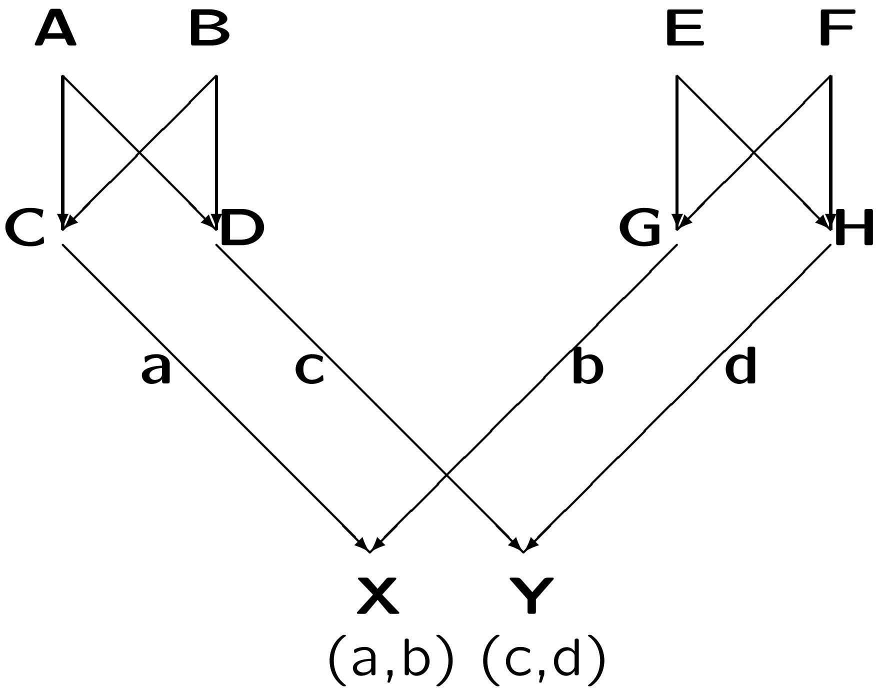
		0.5	0.5
		$b \equiv d$	$b \not\equiv d$
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

$$k_0 = 0.25, k_1 = 0.50, k_2 = 0.25$$

First cousins



Double First Cousins



Non-inbred Relatives

Values of the three probabilities for some common relationships between non-inbred relatives are:

Relationship	k_2	k_1	k_0	$\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$
Identical twins	1	0	0	$\frac{1}{2}$
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Parent-child	0	1	0	$\frac{1}{4}$
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{1}{8}$
Half sibs*	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
Unrelated	0	0	1	0

* Also grandparent-grandchild and avuncular (e.g. uncle-niece).

Joint genotypic probabilities

For any specific pair of genotypes, the Δ's or k's describe the identity-by-descent classes. For two A_iA_i:

- with probability k_2 there are two pairs of ibd alleles, sotwo independent A_i alleles. These A_i with probability p_i^2.

- with probability k_1 there is one pair of ibd alleles, so three independent A_i alleles. These are A_i with probability p_i^3.

- with probability k_0 the no ibd alleles, so four independent A_i alleles. These are all A_i with probability p_i^4.

$$\Pr(A_iA_i, A_iA_i) \;=\; k_2p_i^2 + k_1p_i^3 + k_2p_i^4$$

Joint genotypic probabilities

Genotypes	Probability
ii, ii	$k_2p_i^2 + k_1p_i^3 + k_0p_i^4$
ii, jj	$k_0p_i^2p_j^2$
ii, ij	$k_1p_i^2p_j + 2k_0p_i^3p_j$
ii, jk	$2k_0p_i^2p_jp_k$
ij, ij	$2k_2p_ip_j + k_1p_ip_j(p_i + p_j) + 4k_0p_i^2p_j^2$
ij, ik	$k_1p_ip_jp_k + 4k_0p_i^2p_jp_k$
ij, kl	$4k_0p_ip_jp_kp_l$

Example: Non-inbred full sibs

Genotypes	Probability
ii, ii	$p_i^2(1+p_i)^2/4$
ii, jj	$p_i^2 p_j^2/4$
ii, ij	$p_i p_j(p_i+p_j)/2$
ii, jk	$p_i^2 p_j p_k/2$
ij, ij	$p_i p_j(1+p_i+p_j+2p_i p_j)/2$
ij, ik	$p_i p_j p_k(1+2p_i)/2$
ij, kl	$p_i p_j p_k p_l$

“It was my brother.”

The defense hypothesis may be that the source of an evidentiary stain was a relative of the defendant. For example

> H_p: the defendant is the source of the crime stain.
> H_d: (an untyped) brother of the defendant is the source of the crime stain.

“It was my brother.”

If the evidence profile is $E : AB$ and the defendant has genotype $G_S : AB$, then the likelihood ratio is

$$
\begin{aligned}
\text{LR} &= \frac{\Pr E|H_p)}{\Pr(E|H_d)} \\
&= \frac{1}{\Pr(AB|\text{brother of } AB \text{ person})} \\
&= \frac{1}{\Pr(AB, AB|\text{brothers})/\Pr(AB)} \\
&= \frac{1}{[p_A p_B(1 + p_A + p_B + 2p_A p_B)/2]/(2p_A p_B)} \\
&= \frac{4}{1 + p_A + p_B + 2p_A p_B}
\end{aligned}
$$

Are These People Related?

Remains identification often involves the comparison of two profiles and comparing the hypotheses:

H_1: These profiles are from two people with a specific relationship.
H_2: These profiles are from two unrelated people.

If the profiles have genotypes ab and cd at a locus, then the likelihood ratio is

$$\text{LR} = \frac{\Pr(ab, cd|H_1)}{\Pr(ab, cd|H_2)}$$

Example: Full sibs vs Unrelated

Suppose two samples X, Y have genotypes AA and AB at a locus. For

H_p: X, Y are from full-sibs
H_d: X, Y are from unrelated individuals

The likelihood ratio is

$$\begin{aligned}
\text{LR} &= \frac{\text{Pr}(AA, AB|\text{Full sibs})}{\text{Pr}(AA, AB|\text{Unrelated})} \\
&= \frac{k_1 p_A^2 p_B + k_0 2 p_A^3 p_B | k_1 = 0.5, k_0 = 0.25)}{k_1 p_A^2 p_B + k_0 2 p_A^3 p_B | k_1 = 0.0, k_0 = 1.00)} \\
&= \frac{p_A^2 p_B + p_A^3 p_B}{4 p_A^3 p_B} = \frac{1 + p_A}{4 p_A}
\end{aligned}$$

Example: Full sib vs Half sibs

Suppose two samples X, Y have genotypes AB and AB at a locus.

For

H_p: X, Y are from full-sibs
H_d: X, Y are from half-sibs

The likelihood ratio is

$$\begin{aligned}
\text{LR} &= \frac{\Pr(AA, AB|\text{Full sibs})}{\Pr(AA, AB|\text{Half sibs})} \\
&= \frac{2\frac{1}{4}p_Ap_B + \frac{1}{2}p_Ap_B(p_A + p_B) + 4\frac{1}{4}p_A^2p_B^2}{\frac{1}{2}p_Ap_B(p_A + p_B) + 4\frac{1}{2}p_A^2p_B^2} \\
&= \frac{p_Ap_B + p_Ap_B(p_A + p_A) + 2p_A^2p_B^2}{p_Ap_B(p_A + p_B) + 2p_A^2p_B^2} = \frac{1 + p_A + p_B + 2p_Ap_B}{p_A + p_B + 2p_Ap_B}
\end{aligned}$$

Match Probabilities for Relatives

For relatives, described by k_0, k_1, k_2 in a structured population described by θ:

$$\begin{aligned}
\Pr(A_uA_u, A_uA_u) &= k_0 \Pr(A_uA_uA_uA_u) + k_1 \Pr(A_uA_uA_u) + k_2 \Pr(A_uA_u) \\
\Pr(A_uA_v, A_uA_v) &= 4k_0 \Pr(A_uA_uA_vA_v) + k_1[\Pr(A_uA_uA_v) + \Pr(A_uA_vA_v)] \\
&\quad + 2k_2 \Pr(A_uA_v), \quad u \neq v.
\end{aligned}$$

The allelic-set probabilities in these equations refer to the generation to which the relatives' most recent common ancestors belong. The match probabilities become

$$\Pr(A_uA_v|A_uA_v) = \begin{cases} k_0\dfrac{[2\theta+(1-\theta)p_u][3\theta+(1-\theta)p_u]}{(1+\theta)(1+2\theta)} \\ \quad + k_1\dfrac{2\theta+(1-\theta)p_u}{1+\theta} + k_2, & u = v, \\ \\ k_0\dfrac{2[\theta+(1-\theta)p_u][\theta+(1-\theta)p_v]}{(1+\theta)(1+2\theta)} \\ \quad + k_1\dfrac{2\theta+(1-\theta)(p_u+p_v)}{2(1+\theta)} + k_2, & u \neq v \end{cases}$$

Parameters p_u and θ are assumed to have the same value in successive generations.

Relatedness and Matching

What is the chance that two relatives, with relationship described by k_0, k_1, k_2, match?

$$\begin{aligned}
\text{Pr(Match)} &= k_2 + k_1[\sum_i \text{Pr}(A_iA_iA_i) + \sum_i \sum_{j\neq i} \text{Pr}(A_iA_jA_j)] + k_0P_2 \\
&= k_2 + k_1[\theta + (1-\theta)S_2] + k_0P_2 \\
\text{Pr(Partial Match)} &= k_1[2\sum_i \sum_{j\neq i} \text{Pr}(A_iA_iA_j) + \sum_i \sum_{j\neq i} \sum_{k\neq i,j} \text{Pr}(A_iA_jA_k)] \\
&\quad + k_0P_1 \\
&= k_1(1-\theta)(1-S_2) + k_0P_1 \\
\text{Pr(Mismatch)} &= k_0P_0
\end{aligned}$$

where P_2, P_1, P_0 are the match, partial match and mismatch probabilities for unrelated people. Setting $\theta = 0$ gives the results for unstructured populations.

Relatedness and Matching Data

If $\theta = 0.03$, using FBI allele frequencies for Caucasians.

	Not	First-	Parent	Full-
Locus	related	cousins	-child	sibs
D3S1358	.089	.124	.229	.387
vWA	.077	.111	.213	.376
FGA	.048	.078	.166	.345
D8S1179	.083	.119	.227	.384
D21S11	.051	.081	.172	.349
D18S51	.040	.068	.150	.335
D5S818	.175	.216	.339	.463
D13S317	.101	.139	.252	.401
D7S820	.080	.115	.219	.379
CSF1PO	.134	.173	.288	.428
TPOX	.216	.261	.397	.503
THO1	.096	.133	.241	.395
D16S539	.105	.143	.256	.404
Total	2×10^{-14}	2×10^{-12}	6×10^{-9}	5×10^{-6}

Mixtures

There are many situations where the DNA person from more than one person is present in an evidentiary sample:

- Rape: DNA from victim, assailant and possible consensual partners.

- Murder: DNA from victim and assailant.

- Touch DNA: several people who touched a surface.

Binary Model

The simplest approach is to examine the evidence profile and determine which alleles are present.

For example, suppose locus D8S1179 is typed and alleles 12,13,14 are seen. What are the genotypes of possible contributors?

Random Man Not Excluded/Inclusion Probability

The RMNE or CPI approach lists the genotypes that are not excluded, i.e. are included in the evidence profile. For the 12,13,14 example these are:
12,12; 12,13; 12,14; 13,13, 13,14, 14,14.
If all allele frequencies were 0.1, then these six genotypes have (Hardy-Weinberg) probabilities
0.01; 0.02; 0.02; 0.01; 0.02; 0.01
so the probability that a "random man" would not be excluded is 0.09.

What is wrong with this?

Problem with RMNE

The problem with the RMNE approach is that it does not take into account that there must be (at least) two contributors to the profile 12,13,14. Some pairs of people could not be the contributors:

12,12 and 12,12; 12,12 and 13,13; 12,12 and 14,14;
12,12 and 12,13; 12,12 and 12,14 etc.

Problem with RMNE

The possible pairs of random people who would produce a profile of type 12,13,14 are:

	12,12	12,13	12,14	13,13	13,14	14,14
12,12	no	no	no	no	yes	no
12,13	no	no	yes	no	yes	yes
12,14	no	yes	no	yes	yes	no
13,13	no	no	yes	no	no	no
13,14	yes	yes	yes	no	no	no
14,14	no	yes	no	no	no	no

Problem with RMNE

Putting in the genotype probabilities for allele frequencies of 0.10:

		12,12	12,13	12,14	13,13	13,14	14,14
		0.01	0.02	0.02	0.01	0.02	0.01
12,12	0.01					0.0002	
12,13	0.02			0.0004		0.0004	0.0002
12,14	0.02		0.0004		0.0002	0.0004	
13,13	0.01			0.0002			
13,14	0.02	0.0002	0.0004	0.0004			
14,14	0.01		0.0002				

The probability two random men have the profile 12,13,14 is 0.0036, which is less than the RMNE.

Problem with RMNE

Another problem with RMNE is that it does not consider that there may be a known contributor, e.g. the victim.

Suppose the evidence is the profile 12,13,14 and the victim is of type 12,13. The other contributor must have an allele type 14.

Now suppose a suspect has type 13,14. He is not excluded as a contributor. The two hypotheses may be:

> H_p: the evidence is from the victim and the suspect.
> H_d: the evidence is from the victim and an unknown man.

The probability of the evidence if H_p is true is 1.

Problem with RMNE

If H_d is true, the unknown man must be of type 12,14 or 13,14 or 14,14 and these have probabilities 0.02, 0.02, 0.01 if the allele frequencies are 0.1.

The likelihood ratio is

$$\text{LR} \; = \; \frac{\Pr(E|H_p)}{\Pr(E|H_d)} = \frac{1.0}{0.05} = 20$$

The RMNE approach would give a LR of 1/0.09= 11.1.

Victim not included

If the evidence is not an intimate sample, the hypotheses may be different:

H_p: the evidence is from the victim and the suspect.
H_d: the evidence is from two unknown men.

The probability of the evidence if H_p is true is still 1, but under H_d it is 0.0036 and the LR is 1/0.0036=277.8

The RMNE is still 0.09 for this case. Clearly miss-states the strength of the evidence.

Semi-continuous Model

Allow for allelic drop-out and allelic drop-in.

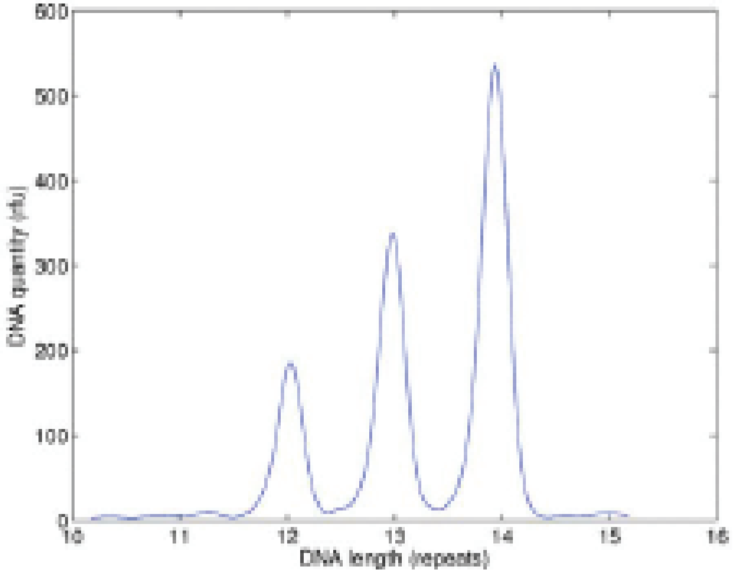
C or $\bar{C}$ are the probabilities an allele drops out, or does not drop out.

D or $\bar{D}$ are the probabilities an allele drops in, or does not drop in. Dp_A is the probability an allele of type A drops in.

Continuous Model

The “binary model” analyses ignore features of the electropherograms - in particular, how much DNA is present for each allele.

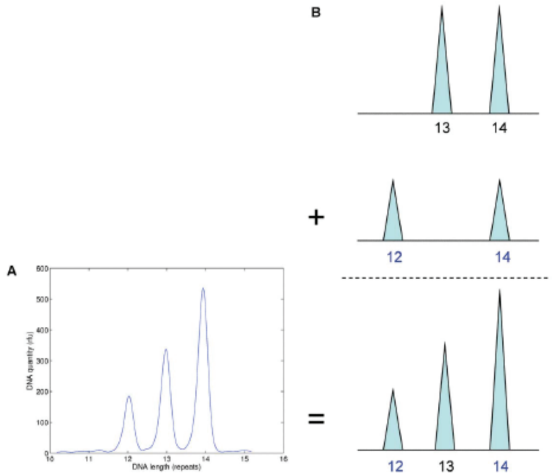
A mixture profile may actually look like



(Source: Perlin and Sinelnikov, PLoS One 4:e837, 2009)

Continuous Model

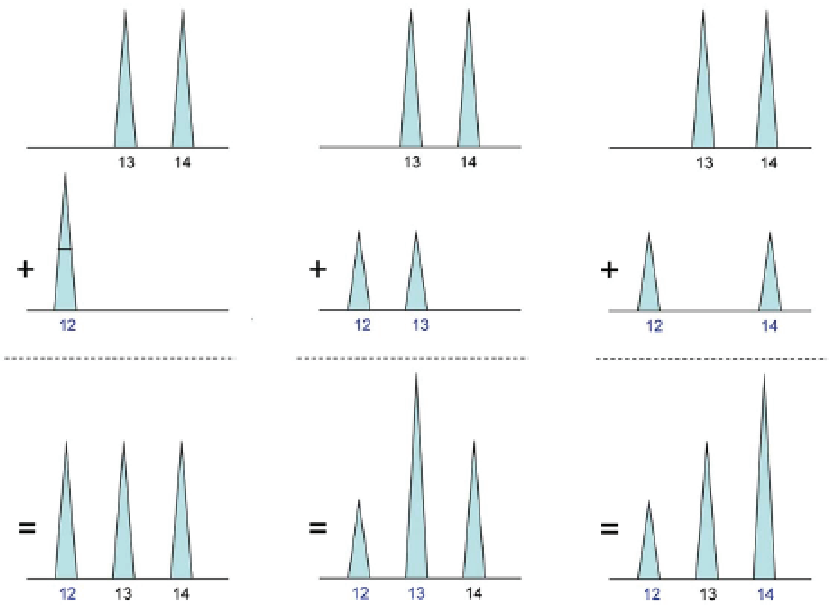
The mixture of alleles 12,13,14 from the previous slide; along with the “peak heights” (amount of DNA) could be explained as:



(Source: Perlin and Sinelnikov, PLoS One 4:e837, 2009) This approach allows for different amounts of DNA from each contributor – there may be more from the victim than the assailant.

Continuous Model

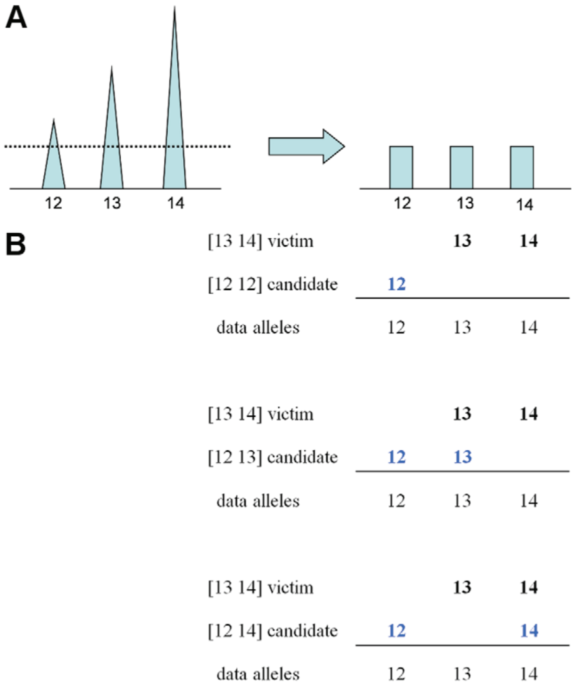
Taking peak heights into account can also rule out some contributors to the evidence profile. If the victim of type 13,14 is present, then only one of the possible second contributors is consistent with the electropherogram:



(Source: Perlin and Sinelnikov, PLoS One 4:e837, 2009)

Binary vs Continuous Model

To emphasize the difference between the binary and the continuous models, note that the binary model uses a threshold: all peaks above that threshold are retained and are given equal weight. Other peaks are discarded.



[13 14] victim		**13**	**14**
[12 12] candidate	**12**		
data alleles	12	13	14

[13 14] victim		**13**	**14**
[12 13] candidate	**12**	**13**	
data alleles	12	13	14

[13 14] victim		**13**	**14**
[12 14] candidate	**12**		**14**
data alleles	12	13	14

Binary vs Continuous Model

Suppose the alleles 12,13,14 have frequencies of 0.10.

For the hypotheses H_p: Victim (13,14) plus Suspect (12,13) versus H_d: Victim (13,14) plus Unknown:

the binary model allows the unknown contributor for H_d to be 12,12 or 12,13 or 12,14 and the LR is

$$\mathrm{LR} = \frac{1}{\Pr(1,12) + \Pr(12,13) + \Pr(12,14)} = \frac{1}{0.01 + 0.02 + 0.02} = 20$$

but the continuous model restricts the unknown contributor for H_d to be 12,13 and the LR is

$$\mathrm{LR} = \frac{1}{\Pr(12,13)} = \frac{1}{0.02} = 50$$

Peaks below threshold

When allelic peaks fall below a threshold, under the binary model that locus is ignored. This has been thought to be conservative but may actually be prejudicial if the alleles in question would weaken the strength of the evidence against the suspect.

The semi-continuous model takes all peaks into account by including for "drop-out" and "drop-in" probabilities in the calculations.

On the next slide are the profiles of the evidence, the victim and a suspect in the case against Charles Richard Smitj, Superior Court of California, County of Sacramento, Number 06FO0122. The prosection presented a likelihood ratio of 96,000 and the defense gave a value of 2. The defense expert, Professor David Balding said "The treatment of the DNA evidence in this case was the worst I've encountered."

California vs Smith

				Likelihood ratio	
Locus	Evidence	Victim	Suspect	Prosecution	Defense
D8	12,13,16	13,16	12,13	4.0	4.5
D21	28	28,30	29,29	1	0.4
D7	–	8,10	9,10	1	1.1
CSF	–	8,10	10,11	1	1.1
D3	16	14,16	16,17	4.3	1.4
TH01	7	7,7	9.3,9.3	1	0.4
D13	–	11,13	8,12	1	1.1
D16	12,13	12,13	11,12	4.0	0.9
D2	24	19,24	17,25	1	0.8
D19	12,13	12,13	13,15	6.5	1.1
vWA	18,20	18,20	19,20	18	1.4
TPO	9,11	9,9	11,12	7.0	1.3
D18	–	13,15	12,17	1	1.1
D5	8,11,12	8,12	11,13	1.7	0.9
FGA	–	21,22	22,24	1	1.1
Product				96,000	2.0

Balding and Buckleton, Forensic Science International: Genetics 4:1-10, 2009.

Case in Lohmueller and Rudin

(Lohmueller and Rudin, Journal of Forensic Sciences 58:S243-S249, 2012)

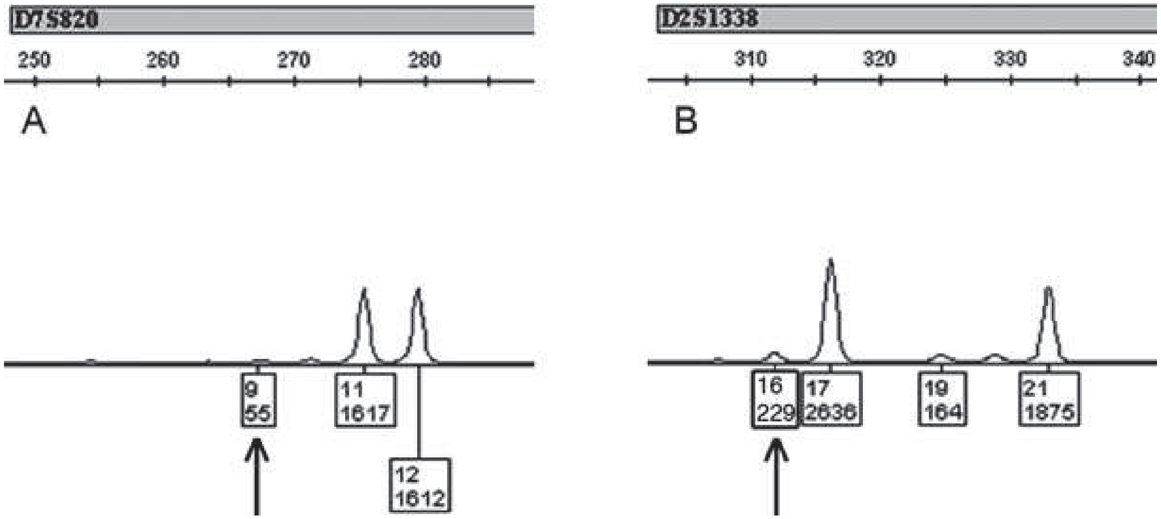
Female victim stabbed to death. Husband was suspect. Husband's girlfriend was an alternate suspect.

Agreement that victim was the major contributor to bloodstains.

For some loci, a complete analysis requires the probability of allelic drop-out to be included in the calculations.

Leaving out loci with peaks below the threshold was prejudicial to the suspect.

Case in Lohmueller and Rudin



Panel A: allele 9 was below detection threshold (so not called) and was used to exclude alternate suspect.

Panel B: allele 16 was below the stutter threshold (so regarded as stutter) and was used to exclude the alternate suspect.

Case in Lohmueller and Rudin

	Evidence					
Locus	Major	Minor	Stutter	Victim	Suspect	Alt. Suspect
D8	13,14	15	12	13,14	13,15	13,15
D21	30,31.2	32.2	29,30.2	30,31,2	30,31.2	30,32.2
D7	11,12	9	10	11,12	8,11	9,11
CSF	10,13	11,12	9	10,13	12,13	11,12
D3	15,17	16	14	15,17	15	16
TH01	7,8	9	6	7,8	8,9.3	8,9
D13	11	8,12	10	11	10,11	8,12
D16	11,13	9	10,12	11,13	11,14	9,11
D2	17,21	19	16,20	17,21	17,26	16,19
D19	13	14	12	13	13	13,14
vWA	17	15,18	16	17	17	15,18
TPO	8,9	10		8,9	9,11	9,10
D18	13,17		12,16	13,17	13,17	17
D5	11,12		10	11,12	11	11,12
FGA	20,21	24	19	20,21	21,24	19,24

Greater LR for H_p: Wife plus Girlfriend vs H_d: Wife plus Unknown than for H_p: Wife plus Husband vs H_d: Wife plus Unknown.