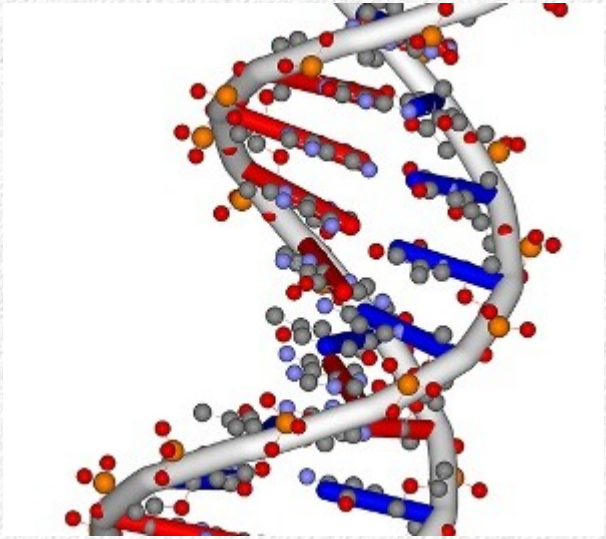
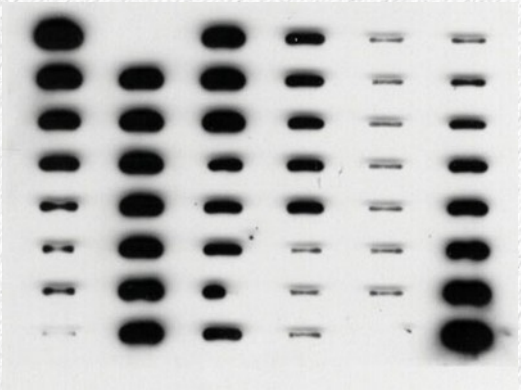
Likelihood and DNA





$$\mathrm{LR} = \frac{p(\mathrm{M}|\mathrm{P})}{p(\mathrm{M}|\mathrm{E})} = \frac{1}{p(\mathrm{M}|\mathrm{E})}$$

DNA evidence

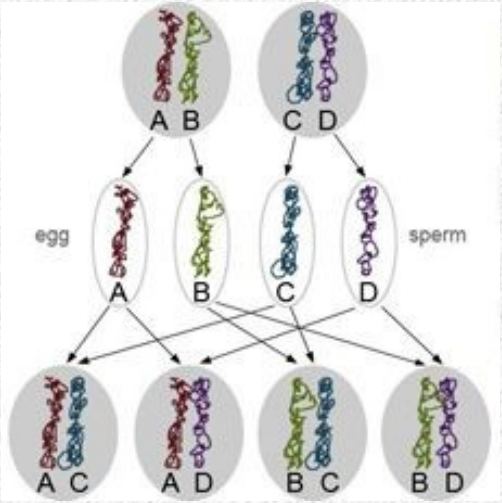
- Like fingerprints, but less subjective
  - Genes are discrete
  - Matches are either positive or negative
  - No judgment calls
- If matches are either positive or negative, why do they talk about probabilities of a match?

Statistics and DNA evidence

- Probability of matches in DNA cases involve some pretty small probabilities
  - Ex. “The chances that this sample belongs to somebody other than the defendant is 100 billion to one”
- What do statisticians mean when they say this?
- Where do these numbers come from?

Genes and alleles

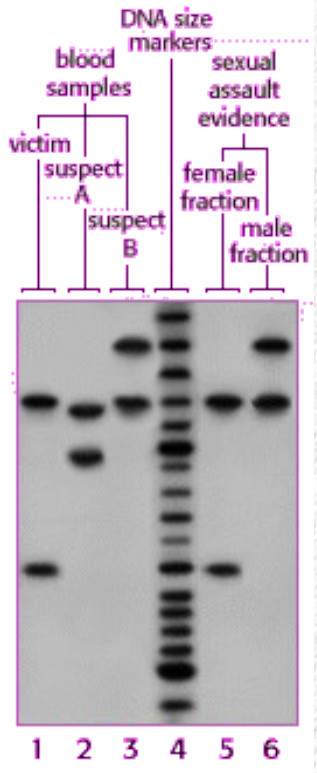
Mom Dad



Kids

- One copy of our 23 chromosomes from each parent
- Each chromosome in a pair has the same genes
- Genes can have different variants, called “alleles”
  - Can get the same alleles from mom and dad = homozygous
  - Can get different alleles from mom and dad = heterozygous

Example of a DNA fingerprint - one locus



- Evidence collected from a sexual assault victim
- Male fraction from recovered semen
- Two suspects, A and B, are apprehended
- Samples from the victim and the suspects are compared with the crime scene evidence
- Alleles at a locus appear on a gel as bands
- A suspect that has both of the bands present in the evidence is a match
- Which suspect matches?

What's the probability of a genotype?

- What are the chances of a match to a randomly selected person?
  - If an (innocent) person matches the evidence → false accusation, conviction
  - If the probability of this is high, then the method isn't usable
- We need to know two things:
  - What is the frequency of each allele in the population?

    ...which can be used to tell us…

  - What is the frequency of each genotype in the population?

Hardy-Weinberg frequencies with two alleles

Frequency of A in population = p

Frequency of B in population = q

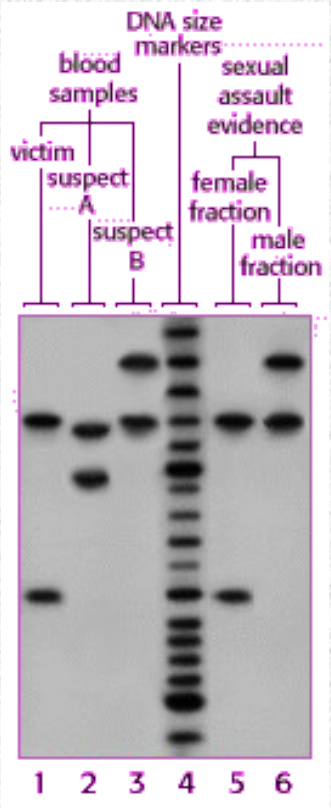
Frequency of AA homozygotes: p^2

Frequency of BB homozygotes: q^2

Frequency of heterozygotes: 2pq

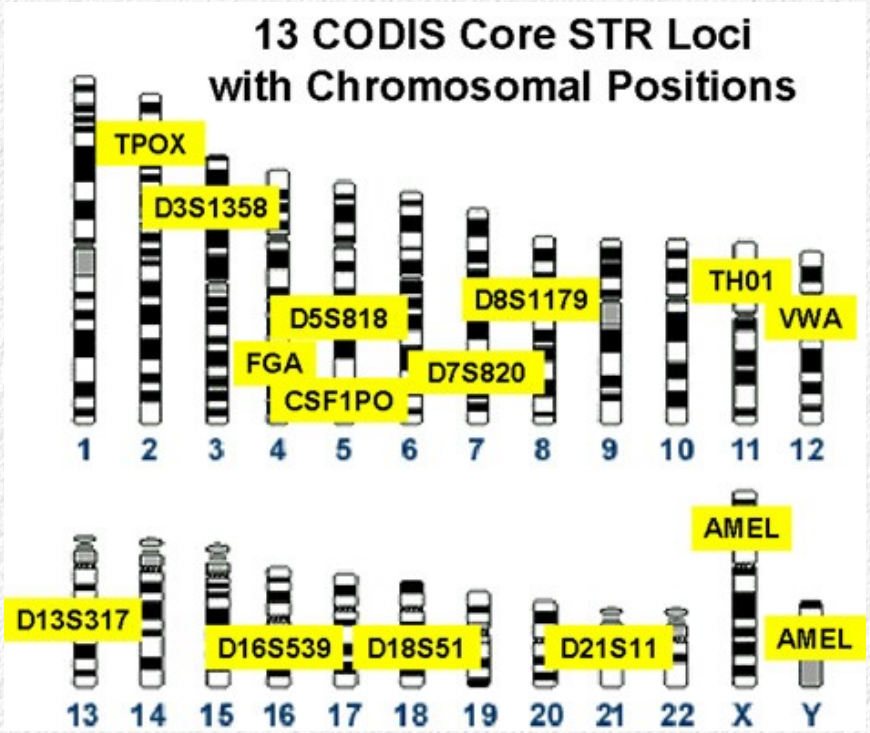
	A (p)	B (q)
A (p)	AA (p^2)	AB (pq)
B (q)	AB (pq)	BB (q^2)

Probability of a match to our suspect



- Suspect B is a match
- If the frequency in the population of the top band is 0.15, and of the bottom band is 0.26...
  - The genotype frequency of this heterozygous genotype would be 2pq
  - 2(0.15)(0.26) = 0.078
- Therefore, we could sample from our population at random, and get a match to this genotype 7.8% of the time
- With San Diego County's population of 3 million, that would be 234,000 matches expected
- Not good enough! The solution? More genes

13 loci used by FBI



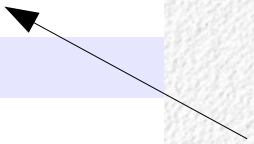
Loci used are highly polymorphic (i.e. many alleles), and neutral (i.e. not subject to natural selection)

Either on different chromosomes, or far apart on the same chromosome

Important?

An STR profile

STR Locus	Allele # (maternal)	Allele # (paternal)
D3S1358	16	16
VWA	15	16
FGA	19	20
D8S1179	12	12
D21S11	29	31.2
D18S51	12	17
D5S818	11	13
D13S317	10	11
D7S820	10	12
D16S539	8	11
THO1	6	7
TPOX	8	10
CSF1PO	8	12



Identifiers
(this is allele #16)

Convert the profile to genotype frequencies

STR Locus	Allele 1	Allele 2	p	q	Genotype freq.
D3S1358	16	16	0.34	0.34	0.11621
VWA	15	16	0.15	0.26	0.07932
FGA	19	20	0.06	0.07	0.00839
D8S1179	12	12	0.13	0.13	0.01819
D21S11	29	31.2	0.18	0.05	0.01787
D18S51	12	17	0.05	0.18	0.01766
D5S818	11	13	0.24	0.23	0.10986
D13S317	10	11	0.02	0.31	0.01510
D7S820	10	12	0.34	0.12	0.08309
D16S539	8	11	0.04	0.3	0.02285
THO1	6	7	0.15	0.38	0.11535
TPOX	8	10	0.32	0.09	0.05720
CSF1PO	8	12	0.06	0.29	0.03283

← Likelihood of genotype at each gene

Likelihood across all the genes is the product of the likelihoods for each = 3.03×10^{-19}

What can we do with this information?

- Use a likelihood ratio to compare two possibilities:
  - The suspect is the source of the crime scene evidence
  - The suspect is not the source of the crime scene evidence
- Need to know the likelihoods of each possibility
  - The likelihood of a match if the suspect is the source of the blood is 1
  - The likelihood of a match if the sample didn't come from the suspect is tiny – 3.3×10^{-19}
- Likelihood ratio = ratio of the likelihood if the suspect is the source of the blood to the likelihood if the suspect is not the source of the blood

Likelihood ratio

$$\mathrm{LR} = \frac{p(\mathrm{Match}|\mathrm{Perp})}{p(\mathrm{Match}|\mathrm{Whoops})} = \frac{1}{p(\mathrm{Match}|\mathrm{Whoops})}$$

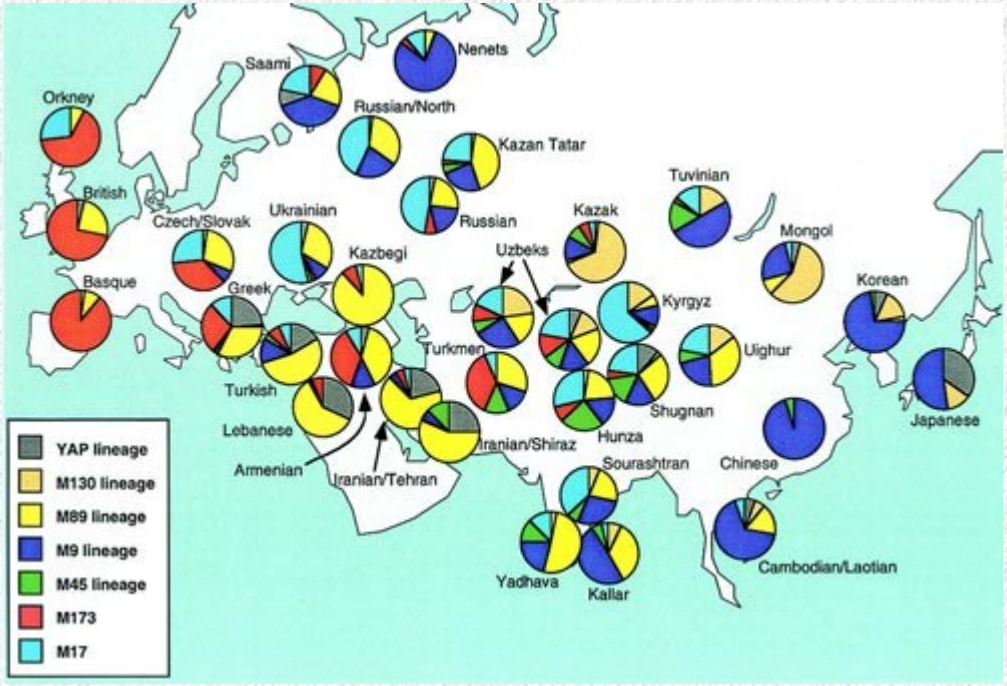
$$\frac{1}{3.3\times10^{-19}} = 1\times10^{19}$$

It is 10,000,000,000,000,000,000 times more likely that the suspect is the source of the crime scene evidence than not

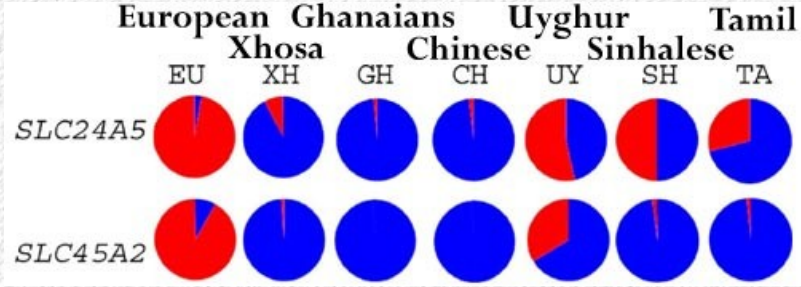
Finding your roots

- Allele frequencies vary among populations
- A DNA fingerprint can be used to assign a person to their most likely population of origin
- Need:
  - Gene frequency data from various populations
  - A DNA fingerprint

Why this works: allele frequencies vary among populations



We can make use of this variability to ask, which population is most likely to have produced the STR profile we have in hand?



Example: which population is this person most likely to have come from?

			Bahama African American			Navajo		
STR Locus	Allele 1	Allele 2	p	q	Genotype freq.	p	q	Genotype freq.
D3S1358	16	16	0.34	0.34	0.11621	0.14	0.14	0.02006
VWA	15	16	0.15	0.26	0.07932	0.02	0.43	0.01646
FGA	19	20	0.06	0.07	0.00839	0.19	0.09	0.03389
D8S1179	12	12	0.13	0.13	0.01819	0.11	0.11	0.01265
D21S11	29	31.2	0.18	0.05	0.01787	0.18	0.06	0.02219
D18S51	12	17	0.05	0.18	0.01766	0.09	0.12	0.02199
D5S818	11	13	0.24	0.23	0.10986	0.58	0.05	0.05766
D13S317	10	11	0.02	0.31	0.01510	0.15	0.22	0.06724
D7S820	10	12	0.34	0.12	0.08309	0.14	0.28	0.08088
D16S539	8	11	0.04	0.3	0.02285	0.01	0.15	0.00409
THO1	6	7	0.15	0.38	0.11535	0.17	0.61	0.20534
TPOX	8	10	0.32	0.09	0.05720	0.35	0.02	0.01151
CSF1PO	8	12	0.06	0.29	0.03283	0.02	0.29	0.00970

Allele frequencies from two different populations

Likelihood of being a Bahama African American = 3.03×10^{-19}

Likelihood of being a Navajo = 2.03×10^{-21}

Likelihood ratio

$$Likelihood\ ratio = \frac{L(profile|BAA)}{L(profile|Nav)} = \frac{3.03 \times 10^{-19}}{2.03 \times 10^{-21}} = 149.26$$

- 148.26 times more likely that this profile is from a person who is BAA rather than Navajo
- Conclusion depends on comparison group used
  - Think of these as competing hypotheses
  - We are only considering two possible alternatives here
  - We could do these calculations for many different possible source populations, and find the one with the highest likelihood of having produced the fingerprint