



Vzdělávání v oblasti forenzní genetiky  
reg. č. CZ.1.07/2.3.00/09.0080

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

## *Evidentiary strength of a rare haplotype match: What is the right number?*

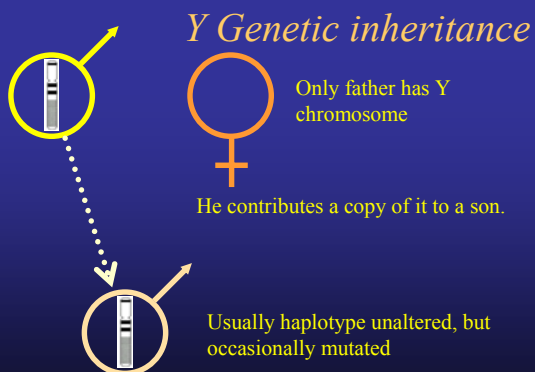
Charles Brenner, PhD  
DNA·VIEW and UC Berkeley Public Health  
www.dna-view.com c@dna-view.com

Brenner CH (2010) Fundamental problem of forensic mathematics –  
The evidential value of a rare haplotype  
Forensic Sci. Int. Genet. doi:10.1016/j.fsigen.2009.10.013  
<http://dna-view.com/downloads/documents/BerlinTalk>

The rules of genetics are simple. Their consequences are not always obvious.

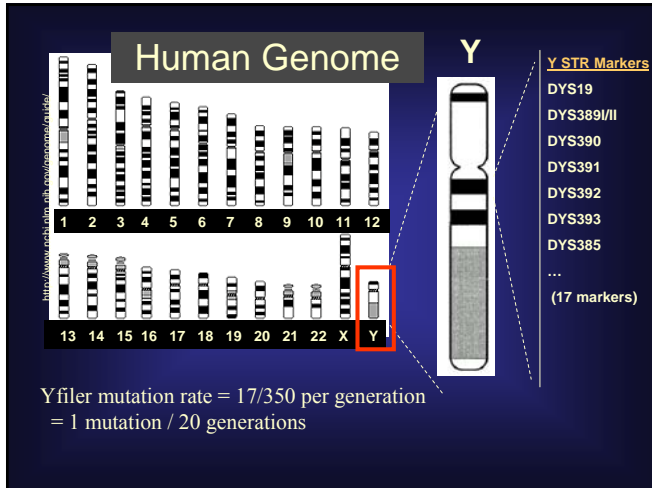
## Understanding Y haplotypes

1. Evolutionary history and population genetics
2. Evidential value



## *All men are related*

- All men alive today have a common Y-chromosome ancestor
- (probably 3,000 generations ago)

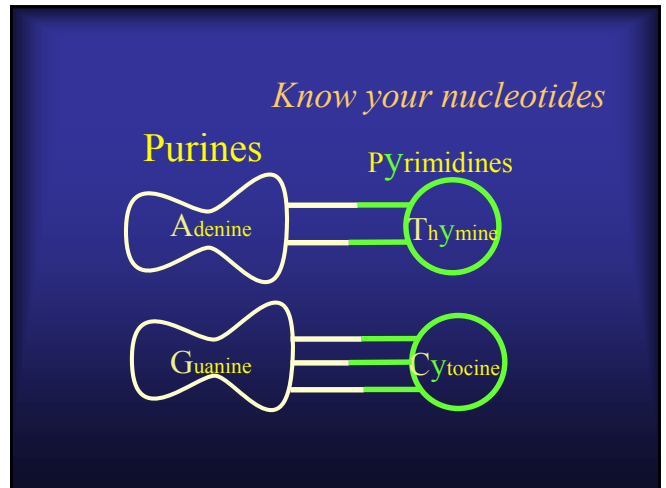
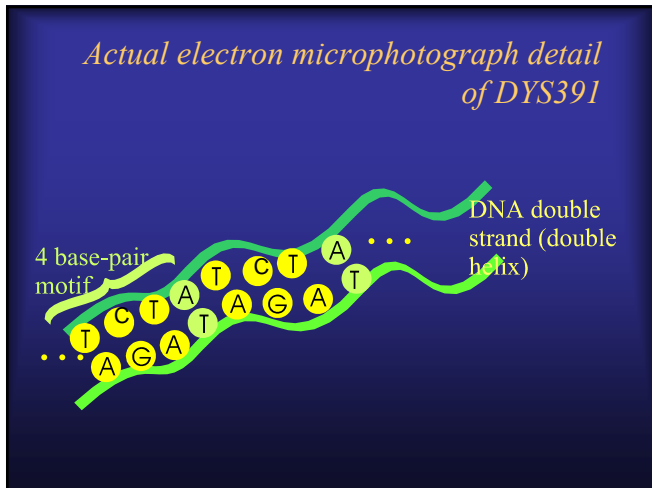


### Y chromosome & haplotype

**Y STR Markers**

- DYS19
- DYS385a/b
- DYS389I/II
- DYS390
- DYS391
- DYS392
- DYS393
- DYS385
- ... (17 markers)

- Chromosome – a physical object
- Haplotype – information on chromosome(s)
  - locus DYS391 (one locus, two loci)
  - Tetrameric repeat (TCTA)<sup>6-14</sup>
  - E.g. a person might be {10} at DYS390 – 10 tandem copies of the motif
  - A Yfiler™ haplotype is 17 loci, e.g. (14,13,14,29,22,10...)



### Likelihood ratio (LR) – the central concept of forensic mathematics

ratio of the probabilities of one event (E) under two different hypotheses (H1, H0).

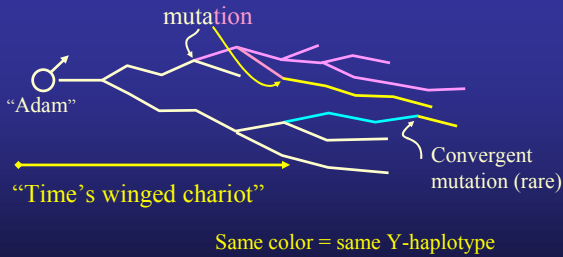
LR principle: Data (E) is evidence for one hypothesis over another to the extent that the data is more probable under the one hypothesis than under the other.

Think of the hypotheses as explanations or indicators for the events.

### Identity by descent or by state?

- Two men have the same Yfiler haplotype.
- Connected to a common ancestor without mutation (IBD), or not?
- (Terminology):
  - IBD = Identity by descent = related with no intervening mutations
  - IBS = Identity by state = same haplotype maybe coincidentally

## Y-haplotype lineage



## Convergent Y mutation

- Y haplotype = 17 numbers = position in 17-space
- Mutation is random walk in 17 dimensions
  - Each step is  $\pm 1$  in some dimension
  - $2 \times 17 = 34$
- Random walks rarely return to start.
  - 2 mutation separation: 1/34 chance that 2<sup>nd</sup> mutation reverses 1<sup>st</sup> one.
  - Probability to converge otherwise is negligible.
- Identical Y-filer haplotype => relationship to common ancestor without mutations (IBD)

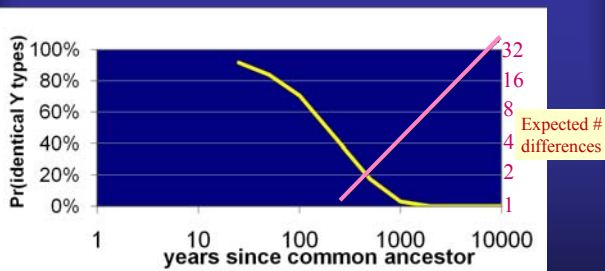
## Convergence experiment

- Simulated Y-filer population (N=90000)
- Small proportion of pair-wise matches
  - Pr(match)= 1/9000
- Given match (IBS), are all IBD?
  - Pr(IBD | IBS) = 33/34 (experimental, from simulation)
  - Close to computed estimate of non-convergence (previous slide).
    - (Why? They are not the same experiment.)

## Time to diverge

- $\mu \approx 1/350$  per locus per generation (1/150-1/3000)
- $\mu \approx 5\%$  per generation (17 loci)
- Suppose 4 generations / century
  - Common ancestor century ago = 3<sup>rd</sup> cousins
  - 8 meioses per century of separation between two contemporary men
- Pr( Y’s equal after 1 century) = 70%
- Expected # differences = 4/millennium.

## Y-haplotype divergence



> virtual non overlap of races

## Familial Y searching

- Suspect who matches Y type is exonerated (maybe it was a population tawl).
- So it’s his brother?

### Comments on crime-suspect match

- If suspect *not* donor, then
  - 97% that suspect and donor are IBD which is very unlikely if they are separated by more than a few centuries.
- A 17-locus haplotype is typically represented by a small number\* of men descended without mutation from a common ancestor 100-200 years ago.
- Probably way beyond immediate family.
  - \* 1/10000 of the population

### probability two random men match?

- Example: 1272 Caucasian men (ABI)
  - 808000 pairwise comparisons (big sample!)
    - 90% of 1272 men are singletons (no pairwise matches)
    - 49 pairs of matching haplotypes (49 matches)
    - 5 triples (5×3=15 pairwise matches)
  - ... in total 91 pairwise matches / 808000
  - Pairwise matching rate 1/8900
- Can evidential strength (new type) be less than that? (no matter what the “upper confidence” limit may be)

### Y-STR efficacy

- random match probability  $\approx 1/10000$
- eliminates all false leads (e.g. familial searching)

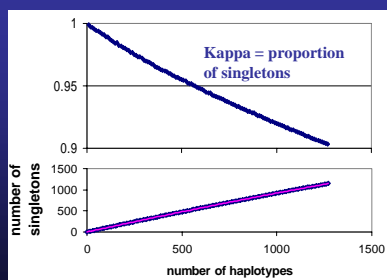
Black	1/14000
Asian	1/4100
Caucasian	1/8900

Y-haplotype matching odds for US populations (17 Yfiler STR loci)

### Probability of a new Y haplotype

- Assume Y-filer (17 STR loci)
- Probability in an actual database?
  - Example: 1272 Caucasian men (ABI sample)
    - 90% are “singletons”
- Smaller database
  - If  $n=1$ , 100% singletons
- Suppose we collect the entire world male population. What % of singletons?

### Growth of a (Y-)haplotype “database” (population sample)



### Y-filer population sample data

- size=# of chromosomes
- $\alpha$ =# of singletons (types not repeated)
- $\kappa = \alpha/\text{size}$ , proportion of sample that is singleton

	Size	$\alpha$	$\kappa = \alpha/n$	$1/(1-\kappa)$ (“inflation factor”)
US Black	985	925	0.94	16.4
Asian	330	312	0.95	18.3
Caucasian	1276	1152	0.90	10.3
Example D	$n-1$	$\alpha$	0.9	10

## Quiz: Probability of new type?

- Assume the Example Y-haplotype database.
  - $\kappa=90\%$  of the chromosomes are singletons.
    - Assume  $\kappa$  changes only slowly as  $D$  grows.
  - What is the probability that the next person sampled has a NEW type?
  - Answer:  $\kappa$  (90%), the same as the probability the last one added was new.
- H. Robbins, Ann Math Stat 1968
- Corollary:  $\kappa$  of the population is not represented in the database.
  - Corollary:  $1 - \kappa$  (e.g. 10%) = probability new observation (i.e. crime scene type) IS represented in the database.

## Crime occurs!



- Y-haplotype obtained
- Interesting case:
  - donor=criminal
  - Crime scene type  $S$  not found in database  $D$
- Assume database  $D$  representative of “suspect population”

## Suspect matches crime scene haplotype. Relevant number?

Relevant number is the matching probability, the probability that an innocent suspect would match the crime scene type given available data of crime scene type & population database and general scientific knowledge.

Is there another kind?



Innocent suspect is the test.

## Suspect matches crime scene haplotype. Relevant number?

Relevant number is the matching probability, the probability that an innocent suspect would match the crime scene type given available data of crime scene type & population database and general scientific knowledge



## SWGAM “Statistical Interpretation”

- Assumes that issue is to “estimate frequency”
  - Unlike probability, refers to unknown information
- “Confidence interval corrects for sampling variation.” (For “unobserved” haplotype, amounts to  $3/N$ .)
  - Purely statistical idea, ignores scientific knowledge, ignores crime scene occurrence.
- Summary: Confuses frequency for probability, and doesn’t even get frequency right.

## Relevant question: $Pr(\text{match})$

- What is the matching probability
  - that a random innocent suspect will match the crime scene DNA type  $S$ ?
  - given that the type was observed at the crime scene,
  - given the available population database  $D$ , which doesn’t have  $S$ . Let the size of  $D$  be  $n-1$ .



### Relevant question: $Pr(\text{match})$

- What is the matching probability
  - that a random innocent suspect will match the crime scene DNA type S?
  - given that the type was observed at the crime scene,
  - given the available population database D, which doesn't have S. Let the size of D be  $n-1$ .



### Relevant question: $Pr(\text{match})$

- What is the matching probability
  - that a random innocent suspect will match the crime scene DNA type S?
  - given that the type was observed at the crime scene,
  - given the available population database D, which doesn't have S. Let the size of D be  $n-1$ .



### Relevant question: $Pr(\text{match})$

- What is the matching probability
  - that a random innocent suspect will match the crime scene DNA type S?
  - given that the type was observed at the crime scene,
  - given the available population database D, which doesn't have S. Let the size of D be  $n-1$ .



### Probability comments

- Probability (of a match)
  - is a summary of information we have
  - Does not involve unknown information.
- information we have:
  - Population sample
  - Crime stain
    - Relevant: observations at crime, in population sample
    - Irrelevant: it's name S
- Good:  $Pr(\text{random match} | \text{data about S})$
- Bad:  $Pr(\text{random match} | \text{name of S})$

### $Pr(\text{match})$ – analysis

- Construct the *ExtendedDatabase* of size  $n$  by including the crime stain S (*condition on S*).
  - ExtendedDatabase has  $a \approx kn$  singletons:  $S=S_0, S_1, S_2, S_3, \dots, S_{a-1}$
- Innocent suspect arrested, with haplotype T.
- We want  $Pr(\text{match}) = Pr(T=S)$ .
  - Same as  $Pr(T=S_i)$  for any  $i$ . (Same information/evidence, so same probability)
    - Same unrelatedness to innocent suspect.
- Obtain in 3 steps.

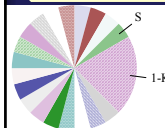


### $Pr(\text{match})$ – 3 part calculation

Assume T is type of innocent suspect

A	T is in ExtendedDatabase	$Pr(A)=1-\kappa$
B	$T=S_i$ for some singleton $S_i$ in the ExtendedDatabase	$Pr(B A) \leq \kappa$
C	$T=S (=S_0)$	$Pr(C B\&A)=1/m\kappa$

}  $1/n$



$$Pr(C) = Pr(C\&B\&A) = Pr(C|B\&A) \cdot Pr(B|A) \cdot Pr(A) \leq (1-\kappa)/n.$$

So ...  $Pr(T=S) \approx (1-\kappa)/n$

- Imagine  $\kappa=90\%$ . Then  $Pr(T=S) \approx 1/10n$ .
- $LR = 1/Pr(T=S) \approx 10n$  is the odds against a random match, the strength of evidence against a matching suspect.
- $1/(1-\kappa)$  – equal to 10 in this example – is the *inflation factor*, the factor by which the matching LR exceeds the simple counting rule estimate.

## Review – wrong question

- ask statistician:
  - “some event seen 0/1000. Frequency?”
    - “some event” ignores the science
    - “0/” ignores the crime scene
    - “frequency” presupposes the wrong question
- statistical answer: “less than 3/1000”
- garbage in, garbage out

$$LR = 1/Pr(T=S) \\ \approx n/(1-\kappa)$$

## Summary

- Test is the innocent suspect
  - probability that an *innocent suspect* would match the crime scene type
- Probability is not frequency
  - (inference from data; no confidence intervals)
- Condition on the crime scene type
  - (toss into database. No more “0 count”).
- Sample frequency may not approximate probability
  - LR can be  $\gg$  sample size



(our new garden sculpture)

The end