

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

### Introduction

Evidence: A trait (e.g. DNA, fingerprint, fiber) is shared between crime scene and suspect. That's easily explained if the suspect is the donor, or it could be a coincidence. "How big a coincidence?" is a basic question, for it quantifies the strength of the evidence linking suspect to crime scene.

In assessing the amount of coincidence one would have to accept to believe the suspect is unrelated to the crime scene, we consult a reference population sample. If the crime scene trait is common in the sample then intuitively it is common in the population and the evidence is weak. It is natural to assume that the sample frequency estimates the population frequency. But if the sample frequency is very small, especially if it is zero (trait never seen before), this intuition breaks down. Clearly the evidence is stronger, but how strong? What to do?

An important example of the problem under discussion is DNA haplotype evidence, in particular a Y chromosomal haplotype.

[1] The subject of this talk is to find "the right number to define the evidentiary value of rare haplotype evidence," a puzzle that I've been fascinated by for at least a dozen years. In fact I gave an early version of this talk with different formulas, different approach but similar result in Berlin ten years ago. I'll concentrate on Y-haplotypes – YFiler in fact – for my examples, but the principles apply as well to mtDNA.

[2] **The rules of genetics are indeed simple** – at least the ones I'm worried about. But I think if we examine them we can quickly see some scientific consequences that people often ignore in favor of mere statistics in dealing with the question that I'm trying to answer here.

[3] Understanding Y haplotypes

Therefore let's take a few minutes to consider the scientific context before looking at the forensic analysis.

[4] Mark Jobling made a comment a few years ago that stuck in my head. **All men are related**. This is obvious; nonetheless you'll often hear the expression "population sample of unrelated men". But especially when dealing with non-recombining haplotypes we should not speak so carelessly. We are told that all living men can trace back to a common Y chromosome about 60000 years ago, so called "Adam".

[5] Identity by descent or by state?

So let's try to examine this question. Suppose that two men have the same Yfiler haplotype. Would you guess that they are connected without mutation through a patrilineage, or that they have coincidentally ended in the same place through different mutations?

[6] Y-haplotype lineage

Let's draw a picture. We start with Adam, who has a number of sons, grandsons, and so on. Seven or eight generations are shown so far. Time is moving to the right. And it occasionally happens in the course of creating offspring that mutation occurs, which I indicate by a change in color. There. There's another one. There's one on the southern branch of the family and another one down there. You see what's happened. We have this rare phenomenon of a convergent mutation. Two men from different branches are both yellow by coincidence.

[7] Convergent Y mutation

Why is convergent mutation rare? A YFiler haplotype is, nearly enough, a collection of 17 integers which we can think of as a lattice position in 17-space. Mutation is therefore analogous to a well-studied mathematical problem called a random walk, in this case a 17-dimensional one, because mutation is normally a plus or minus one repeat step in one of the 17 coordinates. Suppose we consider the lineage connecting two men. There can't be convergent mutation unless there is some mutation, so suppose there are two mutations between them. The first mutation is something, and the second mutation may undo the first but there is only 1/34 chance of that. So with two mutations there is a 3% chance of convergent mutation. If more mutations separate the men, even four of them, then the chance that some two cancel the other two is very small. Consequently if you have identical y-filer haplotypes they are very likely related

without mutation.

#### [8] Convergence experiment

I did an experiment to try to quantify that as follows. I grew a Yfiler population of 90000 men simulating mutation and other parameters realistically so as to mimic available data for a major population. The random matching rate for pairs of my simulated men is 1/9000, just as it is for US Caucasians. As I evolved the population I tagged my data in such a way that I could answer this question: When two men do match by state, what's the chance that they also match by descent? It turns out the answer is 33/34 – an intriguing number. Let me repeat: Identity of haplotypes is, 97% of the time, identity by descent, without intervening mutation.

#### [9] Time to diverge

That fact suggests that identical men are not too distantly related. Let's try to estimate the time scale. STR loci mutate at an average rate of once per 350 meioses. The mutation rate for a Yfiler haplotype is 17 times higher because there, each of 17 loci has the opportunity to mutate, so 5% per generation. Suppose that we have 4 generations per century. That means that two men with a common ancestor a century ago are 3<sup>rd</sup> cousins, separated by 8 meioses, and the chance of no mutation between them is easily calculated to be 70%. With a 1000-year-ago common ancestor, on the other hand there are probably multiple mutations and it's under 2% to have none.

#### [10] Y-haplotype divergence

So we're starting to get the feeling. If the common ancestor is 100 years ago, identity is usual. If it's 1000 years it's rare and across 10000 years identity has probably never happened; certainly identity by descent never has. Therefore we would expect virtual non-overlap of races that have been isolated from one another for that long a time.

#### Familial Y searching

One last question to hone your intuition. Suppose – maybe via a dragnet – we find a suspect whose Y-haplotype matches a crime scene type, but he has a perfect alibi. How likely is it that his immediate relative is the culprit?

#### Comments on crime-suspect match

The answer may depend on a lot of things. But here are some observations I made from my simulations. The innocent suspect and the true donor are IBD. But there are lots more of those IBD pairs who are a century or two apart than those who are first-degree relatives. The typical distance between IBD men is about 10 generations.

#### [11] probability two random men match

It's easy to look in a population sample (such as the ABI data that you can download from the Internet) and compute the rate of random matches which is very low, around 1/9000 depending on population. Many papers have mentioned this result. However no one seems to notice that this gives a robust figure for the evidential strength of a match. Instead, SWGDAM recommends a much weaker number – unnecessarily weaker – based on confusion about statistics.

#### [12] Y-STR efficacy

Here are some examples of matching probabilities, which could be recommended for criminal casework. But I'm going to present something a little different. With background understanding established we're now moving into part two of the talk, forensics.

#### [13] Probability of a new Y haplotype

The ABI Caucasian Yfiler collection includes 1272 men. 90% of them are what I call "singletons", types that occur just once. Of course, 90% isn't a property of the population; it depends on the database size.

#### [14] Growth of a Y-haplotype database

Let's see what happens as the database grows. If the database were small enough it would be 100% singletons. Even set of 100 men are usually all different. But as the database size increases of course the proportion of singletons – which I call  $\kappa$  – declines. Suppose we grow the database to the whole world? At first I figured that since almost every man has a father, son, or brother there would be no singletons. However, I think that's wrong and because of mutation and continual extinctions there would still be about 11% of singletons.

#### [15] Y-filer population sample data

Now lets look at the Yfiler population data in this chart. I show you the size of each database, the number of singletons, the proportion of singletons. You see the typical proportions are > 90%, so for illustrative

purposes we'll imagine a  $K=0.9$  database. The final column, computed as  $1/(1-\kappa)$ , I call the "inflation factor" for a reason that we'll see shortly.

#### [16] Quiz: Probability of a new type

Let's ask a question: We're building a database, adding one person at a time. What is the probability that the next person sampled has a NEW type? Since the database is 90% singletons, that means that whichever man was last added, it's 90% he's one of those singletons, i.e. was new when added. Therefore, since  $\kappa$  changes only gradually, the same should be true of the next man added. 90%.

An obvious inference or restatement of that is that 90% of the world population is *not* yet represented in the database D.

Or, 10% of the world's population, *is* represented. Remember that: the probability is  $1-\kappa$  that a new observation, such as a crime scene type, will be represented in the database.

#### [17] Crime occurs

A Y haplotype is obtained. The interesting and typical case is that the donor is the criminal (that is, that the crime stain is probative) and that the crime scene type is a new type not found in the database. Let's assume that the database is appropriate – I want to keep this talk focus on the basic question.

#### [18] Suspect matches crime scene

First of all, what is the relevant number? This may seem very obvious: It's the probability that an innocent suspect would match the crime scene type. But in fact this is a controversial slide. That particular phrase took me years to come up with – not because it's perfectly honed word by word (it's not); just to get the idea. "innocent suspect" for example. And the probability is computed given the *available data* of the crime scene type and the population database, that's data. And, general scientific knowledge. The Innocent suspect is the test – very important point to clarify thinking on this issue. P Probability is the issue. All of this seems obvious. data means information that we have "available data" – *of course available* data. Perhaps you remember Col. Jessup – Jack Nicholson's – testimony in *A few good men*. The attorney asks "By danger Colonel, do you mean *grave* danger." "Is there another kind?" Is there another kind of data than available data?" you're thinking. All of these obvious points are controversial.

#### [19] Suspect matches crime scene. Relevant number?

General scientific knowledge. Of course you can use that – like this sort of information. But many people don't. Very superficial analysis of this problem if you don't consider the science, the obvious inferences from genetics such as we looked at at the beginning of this talk.

#### [20] SWGDAM "Statistical interpretation"

I'm going to briefly compare the approach that's recommended by SWGDAM. They talk about estimating frequency. That's not probability; frequency is something that we don't know. They talk about a confidence interval in trying to estimate the frequency. This is based on a very dry view of statistics that has nothing to do with the scientific context that we're actually in with rare haplotypes. So the result is that they confuse frequency for probability and don't even understand frequency.

#### [21] Relevant question: Pr(match)

Alright. Returning from my polemics to the math. What is the matching probability that a random innocent suspect will match the DNA type given that the type was observed at the crime, given the database which does not have the crime scene type. It's convenient for me to call the database size  $n-1$  ...

#### [22] Probability comments

Let's recall that probability is a summary of information we have, not unknown information. Also, it saves a lot of confusion to realize that the name of the haplotype is not data. If the crime scene type is a previously unobserved type S, the right question is

what's the probability of a random match keeping in mind the data about S.

The confusing question, misleading and hopeless:

what's the probability of a random match to a haplotype if it is named S?

The reason I'm mentioning this is of course it's the last question that people have generally been stuck on

and which leads to the institutional misconception and dead end of fixating on frequency.

### [23] Pr(match) – analysis

The first trick is that you construct an extended database by tossing in the crime stain. The reason for this is simple. We're going to imagine an innocent suspect whose haplotype is T. We want to consider the probability that T matches S, the crime scene type. Well, that probability is the same as this innocent person matching any singleton type S-sub-i in the database because it's the same information so it's the same probability. Same unrelatedness to an innocent suspect. And this is a very important point to dwell on. Of course, these S's have various frequencies. But that's not data, it's hidden information. It's not of our concern when we're computing a probability. The probability for this man to match any of these S's is exactly the same. Identical data, identical probabilities.

Since all the S's have identical status, it's convenient to put S with the others to have them all in the extended database together. Incidentally, by doing that we are in effect conditioning on the crime scene type.

### [24]

So we can now obtain the answer in 3 steps.

Step A: Is the man's type in the database at all. As I commented a few slides ago, the chance of that is  $1-\kappa$

Step B: If he's in the database, since  $\kappa$  of the database is singletons his pro-rata chance to match among the singletons is  $\kappa$ .

Step C: If he matches some singleton, then since the number of singletons is  $\alpha n$ , the the chance to match the interesting one is  $1/\alpha n$ .

I do break this into 3 steps because I like that step C emphasizes the equal probability for matching any of the singletons. But one could easily collapse B & C and simply say that if he's in the database of size  $n$  his pro-rata chance to match a singleton is  $1/n$ .

Whichever way, multiplying everything together gives a matching probability [of  $(1-\kappa)/n$ ] as shown.

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

The probability of a match is  $(1-\kappa)/n$ . So what does this mean?

If  $\kappa=0.9$  that means the probability is about  $1/10n$ . People have said that the strength can't be more than the size of the database. No, in this case it's 10 times bigger. That's why I call 10 the "inflation factor", the factor by which the LR exceeds the size of the database.

### [26] Review– wrong question

Comparing this with the statistical method, the history is somebody asked a statistician "If some event was seen zero times in a thousand. What's the frequency?" The statistician naively accepts the question as given – statisticians don't ask prying questions "Gee what's your context, what are you asking about, how can I help?" but other than that it's the asker's fault. "some event" ignores the science, "zero over" fails to condition, and "frequency" is the wrong question. So the statistician reasonably says "less than  $3/1000$ ". That's a fair answer to the wrong question.

### [27] Summary

Finally here are the messages that I'd like you to take away. Just four points

The test is an innocent suspect. We're after the probability that an innocent suspect would match the crime scene type.

The remaining three points I listed in Berlin 10 years ago.

Probability not frequency and there's a huge difference. Until you've freed your mind from the institutionalized misconception that the question is one of frequency, you cannot solve this problem.

Conditioning on the crime scene type: even this point has not been generally appreciated. Some people say "condition on the crime scene type" and they don't do it.

The intuition that sample frequency approximates probability is not correct when traits are mostly very rare – which evidenced by a lot of singletons. Then, the LR can significantly exceed the sample size.

[28] The end

I'll leave you with our nice sculpture which my wife had made.