

Schneier on Security

[Blog](#) [Newsletter](#) [Books](#) [Essays](#) [News](#) [Talks](#) [Academic](#) [About Me](#)

[Home](#) > [Blog](#)

DNA Matching and the Birthday Paradox

Nice essay:

Is it possible that the F.B.I. is right about the statistics it cites, and that there could be 122 nine-out-of-13 matches in Arizona's database?

Perhaps surprisingly, the answer turns out to be yes. Let's say that the chance of any two individuals matching at any one locus is 7.5 percent. In reality, the frequency of a match varies from locus to locus, but I think 7.5 percent is pretty reasonable. For instance, with a 7.5 percent chance of matching at each locus, the chance that any 2 random people would match at all 13 loci is about 1 in 400 trillion. If you choose exactly 9 loci for 2 random people, the chance that they will match all 9 is 1 in 13 billion. Those are the sorts of numbers the F.B.I. tosses around, I think.

So under these same assumptions, how many pairs would we expect to find matching on at least 9 of 13 loci in the Arizona database? Remarkably, about 100. If you start with 65,000 people and do a pairwise match of all of them, you are actually making over 2 billion separate comparisons ($65,000 * 64,999/2$). And if you aren't just looking for a match on 9 specific loci, but rather on *any* 9 of 13 loci, then for each of those pairs of people there are over 700 different combinations that are being searched.

So all told, you end up doing about 1.4 trillion searches! If 1 in 13 billion searches yields a positive match as noted above, this leads to roughly 100 expected matches on 9 of 13 loci in a database the size of Arizona's. (The way I did the calculations, I am allowing for 2 individuals to match on different sets of loci; so to get 100 different pairs of *people* who match, I need a match rate of slightly higher than 7.5 percent per locus.)

EDITED TO ADD (9/14): The FBI is trying to [suppress](#) the analysis.

Tags: [biometrics](#), [disclosure](#), [DNA](#), [FBI](#), [forensics](#), [secrecy](#)

[Posted on September 11, 2008 at 6:21 AM](#) • [31 Comments](#)

- Two clicks for more privacy: The Facebook Like button will be enabled once you click here. No data is loaded from Facebook until you enable the button. Click the [i] button for more information.

not connected to Facebook

Comments

Matthew • [September 11, 2008 6:41 AM](#)
Two clicks for more privacy: The Tweet button will be enabled once you click here. No data is loaded from Twitter until you enable the button. Click the [i] button for more information. Now add the odds in that more than one suspect in a given case happens to match DNA found at the crime scene and you are back to 1 in a trillion odds.

Matthew • [September 11, 2008 6:41 AM](#)
If you click to activate the share buttons, data will be loaded from a third party, allowing them to track your visit to schneier.com. For more details click the [i] button.

Bernie • [September 11, 2008 7:11 AM](#)

I guess that I am going to have to RTFA because I don't know what it is about just from the extract.

Roy • [September 11, 2008 7:12 AM](#)

They keep leaving out that 1 mismatch in N loci is proof of a mismatch. Somehow a few failures to match gets ignored and the innocent gets convicted.

Sparky • [September 11, 2008 7:17 AM](#)

Search

Powered by [DuckDuckGo](#)

Go

☐ Blog ☐ Essays ☒ Whole site

Subscribe



About Bruce Schneier



I am a [public-interest technologist](#), working at the intersection of security, technology, and people. I've been writing about security issues on my [blog](#) since 2004, and in my monthly [newsletter](#) since 1998. I'm a fellow and lecturer at Harvard's [Kennedy School](#), a board member of [EFF](#), and the Chief of Security Architecture at [Inrupt, Inc.](#) This personal website expresses the opinions of none of those organizations.

[Contact Info](#)

Related Entries

[Vulnerability Disclosure in the Age of AI](#)

[FBI's 2025 Internet Crime Report](#)

[Bypassing On-Camera Age-Verification Checks](#)

[Smart Glasses for the Authorities](#)

[FBI Extracts Deleted Signal Messages from iPhone Notification Database](#)

[Proton Mail Shared User Information with the Police](#)

Featured Essays

[Four Ways AI Is Being Used to Strengthen Democracies Worldwide](#)

[The CrowdStrike Outage and Market-Driven Brittleness](#)

Of course, these odds aren't really relevant if the DNA evidence is used to prove the guild or innocence of a suspect, for who there already is other (circumstantial) evidence.

When you start using DNA to *find* the suspect, *and* prove his guild, things start to go wrong.

BTW, the 1.4 trillion searches are based on a very naive algorithm. With some clever form of representation and list sorting, you would probably be able to reduce it to a search for matches between consecutive items.

jeff • [September 11, 2008 7:19 AM](#)

Unfortunately, all of the math is based on the false assumption that the loci are statistically independent. I realize that the assumption makes the math easier, but that's no excuse. Within genetically similar populations (races, tribes, families, etc.) all of the odds change.

While I support this use of DNA, people shouldn't be citing odds that have no basis in reality. Its a misuse of mathematics.

Jeff

Sparky • [September 11, 2008 7:20 AM](#)

@Roy: that might be because you might get mismatched loci because of damaged crime scene samples.

It's a bit like regular fingerprinting, you don't have to match every single like, because you're always comparing two imperfect prints

John Moore • [September 11, 2008 7:43 AM](#)

They should be using loci chosen by population geneticists, in other words specific loci within the general population and there's a reason they use 13 markers as a minimum number to match an individual. One of the regions is likely microsatellite. Also, if one thinks about it, one is studying a unique population — those human beings who are incarcerated in prison for various reasons. There may be an unconscious sample selection bias going on within the criminal justice labs due to the fact that these people are already previous or past convicts. If this subpopulation has many markers in common such as SNPs, and these markers aren't ignored, then one would get a very high frequency of matches, but the dataset would still be biased. Are you screening this individual to see if he committed a current crime or whether he has been incarcerated at some point for a past crime? Since the data gathered is mostly from convicts, one is working with a biased data set from the start. If one looks for common patterns within the biased data set, one proves what one is looking for. This is pointed out in *The Mismeasure of Man* by Gould and it's bad science.

xxx • [September 11, 2008 7:48 AM](#)

While I support this use of DNA, people shouldn't be citing odds that have no basis in reality. Its a misuse of mathematics.

Which is exactly the point being made — FBI regularly misuses mathematics to misrepresent to juries what exactly does "defendant's DNA was found at the crime scene" actually mean. They're trying *hard* to give the impression that DNA matching is infallible in practice, which is simply not true.

Roy • [September 11, 2008 8:06 AM](#)

Kary Mullis, who won the Nobel for inventing the polymerase chain reaction that made DNA analysis practical, insisted on a large number of loci — 24, if memory serves — for forensic identification.

Garbage In • [September 11, 2008 8:09 AM](#)

So what are the odds of a false negative? So many criminals convicted of serious crimes have been exonerated based on DNA evidence, could any of them have been guilty?

noah • [September 11, 2008 8:23 AM](#)

@xxx but the odds of someone becoming a suspect AND matching on 13 loci is still so small as to not be relevant. If you just go looking through the whole population to find a match to DNA found at a crime scene, and have no corroborating evidence, then that's another story. I doubt that will ever

[How Online Privacy Is Like Fishing](#)
[How AI Will Change Democracy](#)
[Seeing Like a Data Structure](#)
[LLMs' Data-Control Path Insecurity](#)
[AI and Trust](#)
[The Value of Encryption](#)
[The Eternal Value of Privacy](#)
[Terrorists Don't Do Movie Plots](#)

[More Essays](#)

Blog Archives

[Archive by Month](#)

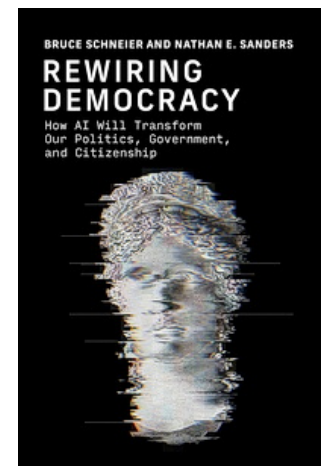
[100 Latest Comments](#)

Blog Tags

[3d printers](#) / [9/11](#) / [A Hacker's Mind](#) / [Aaron Swartz](#) / [academic](#) / [academic papers](#) / [accountability](#) / [ACLU](#) / [activism](#) / [Adobe](#) / [advanced persistent threats](#) / [adware](#) / [AES](#) / [Afghanistan](#) / [AI](#) / [air marshals](#) / [air travel](#) / [airgaps](#) / [al Qaeda](#) / [alarms](#) / [algorithms](#) / [alibis](#) / [Amazon](#) / [Android](#) / [anonymity](#) / [Anonymous](#) / [antivirus](#) / [Apache](#) / [Apple](#) / [Applied Cryptography](#)

[More Tags](#)

Latest Book



[More Books](#)

**I Support
Bloggers' Rights!**

Join **EFF** Now

**DEFEND PRIVACY
Support EPIC**

happen, as the odds are still good that any match would have a good alibi, e.g., living in another region from where the crime occurred, back in prison, dead, etc.

Seth • [September 11, 2008 8:37 AM](#)

noah, if they test precisely 13 loci and someone (who is already a suspect for other reasons, including actual evidence) matches on all, that's a strong indicator. If they test 200 loci and someone matches on 13, that's rather weak evidence.

Bruce, this is the reason people like us don't often get to serve on juries. I've been excluded for knowing arithmetic.

M Welinder • [September 11, 2008 8:39 AM](#)

So what are the odds of a false negative?

Pretty good.

I seem to recall a case where identical twins *both* lost ("won"?) a paternity case.

RonK • [September 11, 2008 8:46 AM](#)

@ Garbage In

So what are the odds of a false negative? So many criminals convicted of serious crimes have been exonerated based on DNA evidence, could any of them have been guilty?

Considering that one way this could happen is if the criminals intentionally have planted DNA evidence from an unrelated party, I'd guess that mathematical ponderings aren't going to help you calculate the odds here.

rich • [September 11, 2008 9:05 AM](#)

@noah

LOL! You have nothing to fear as long as you have an alibi like you're dead or already in prison – too funny.

I hope that was meant as a joke anyhow.

Chris • [September 11, 2008 9:22 AM](#)

The problem with the logic is this statement "aren't just looking for a match on 9 specific loci, but rather on any 9 of 13 loci" In criminal court the loci match has to be exact. In the examples provided if loci 1 matched loci 9 then it would be considered a successful match. For it to be used for identification purposes loci 1 has to match to loci 1 and so on.

Carlo Graziani • [September 11, 2008 9:36 AM](#)

Note that by the standards of the typical criminal judicial process, even a 1% false positive rate is more than acceptable. Compare that rate with what one would expect from tests such as witnesses picking suspects out of a 6-person line-up, or from a stack of 20 photographs, and you get an idea of why the FP rate doesn't really matter.

We're just used to trusting live witnesses more than inanimate ones, so we demand a higher standard of fingerprints, DNA, and the like. Not really justifiably so, in my opinion.

Sparky • [September 11, 2008 10:12 AM](#)

(note: where I live, there is no such thing as a jury trial)

Why is it that the attorneys get to exclude people from a jury anyway? Wasn't the whole idea that a jury is a random sample from the population?

Seth • [September 11, 2008 10:25 AM](#)

Generally, attorneys can exclude anybody who they convince the judge is likely to be biased; in addition, each side usually gets some number of "peremptory" challenges where they can dismiss people without stating a reason.

The first makes sense; the second avoids a lot of hassle over whether someone is sufficiently prejudiced to exclude.

Anonymous • [September 11, 2008 10:25 AM](#)

@sparky

FYI, its Guilt not Guild

kaszeta • [September 11, 2008 10:45 AM](#)

I was actually on a murder jury years ago when I lived in Minnesota, and I remember two things about the DNA evidence:

1. Listening to cross-examination of the DNA sequencing technician is one of the most boring things I've ever experienced, and
2. The false positive rate they quoted along with the DNA evidence was "approximately 1 in 250", consistent with what they say in this article. I remember killing a lot of time in my sequestered hotel room playing around with the probabilities to convince myself that that the 1 in 250 number was reasonable, and what that implied.

Derick • [September 11, 2008 12:12 PM](#)

The more I read the comments the more I believe that none of us know what we are talking about.

DNA testing has been highly scrutinized in courts and the way the evidence has to be presented is very specific. This is a huge improvement over fingerprints which still have an air of infallibility around them. At least with DNA evidence the odds are presented.

NickFadz • [September 11, 2008 1:16 PM](#)

@M Welinder

Actually, that was a false positive, not a false negative. Both twins were identified as being the father. One is the father and the other isn't creating a false positive.

<http://abcnews.go.com/TheLaw/LegalCenter/story?id=3195632>

kiwano • [September 11, 2008 1:55 PM](#)

@rich:

noah's comment must've been a joke. i mean when was the last time that already being in jail was a credible alibi?

Bryan Feir • [September 11, 2008 2:31 PM](#)

@kiwano:

What, you've never seen Two-Way Stretch?

<http://www.imdb.com/title/tt0054417/>

TB • [September 11, 2008 11:35 PM](#)

I hang out with way too many molecular geneticists. I've always enjoyed getting their goat by proposing that the only appropriate use of genetics in the courtroom is to prove the negative. I never bother to try this with pop-gen kids because they tend to know the difference.

Craigh Hughes • [September 12, 2008 1:09 PM](#)

Surely the odds of a particular individual matching some human DNA profile cannot be narrower than 1 in the-total-population-of-history, which is probably on order of 10^{10} people. So saying anything

stronger than “chance is 1 in 10 billion” when referring to the human population is a fallacy. *someone* had to match, and there are only on order 10 billion someone to choose from!

greg • [September 15, 2008 7:13 AM](#)

I am in fact working as a population genetics postdoc.

First one must not assume that courts are up with the facts. They are run by lawyers, not scientists and really have little to know idea when it comes to numbers espically statistics. You only have to look at some of the lead evidence used by the FBI that is now pretty well debunked now to see how courts don't deal well with this.

The problem often comes down to the “the probability of what?” Note the probability of a match is incomplete. Whats the Null? What matches what? My DNA to DNA found at the crime scene.. where I work? This is the problem of proper hypothesis testing.

A good example of this is what locus vary with ethnic origin. Its quite different, and as such if a particular race (say Maori people in NZ) have a much higher chance of matching another person from that same race than 2 Europeans matching each other (I can go the other way round too, like for some African groups).

The second is what kind of Judaical bias you prefer? Is it better to let 100 guilty people go free rather than one innocent person be locked up?

Most western countries are biased *against* locking up or mistreating innocents in principal (Perhaps not in practice).

I will reiterate what some have said above. If you data mine a DNA database, the current loci are not good enough since this is not what they are designed for. But as weight of evidence from a small list of suspects (selected *without* knowledge of the DNA) it works well and perhaps even intuitively. A suspect with other evidence that was found *after* the DNA “match” was discovered is less concrete, and givening odds like 1:1000000 are complete rubbish in this case. At least without proper statical hypothesis.

Filias Cupio • [September 15, 2008 10:40 PM](#)

@Chris:

No, they aren't matching locus 1 from the crime scene to locus 9 of the suspect. It is:

“We compared the suspect to the crime scene sample, and they matched at loci 1, 3, 4, 5, 8, 9, 10, 11 and 13” I.e. there was a mismatch or inconclusive match at 2, 6, 7 and 12. There are 715 ways of chosing 9 loci out of 13, so there are 715 different ways you could get a “9 out of 13” match.

wm • [September 17, 2008 4:02 AM](#)

@noah: “but the odds of someone becoming a suspect AND matching on 13 loci is still so small as to not be relevant. If you just go looking through the whole population to find a match to DNA found at a crime scene, and have no corroborating evidence, then that's another story. I doubt that will ever happen, as the odds are still good that any match would have a good alibi, e.g., living in another region from where the crime occurred”

You're being too optimistic. Take a look at this article:

http://www.forensic-evidence.com/site/EVID/EL_DNAerror.html

Here the guy had an alibi, lived 200 miles away, was too ill to have committed the crime (it doesn't go into details in the article, but if it's the case I'm thinking of, the burglar broke in through a small, high window, whereas this guy couldn't walk in through his own front door on a bad day), and was picked out of a DNA database with no other evidence against him.

None of this counted in his favour, though — he was in jail for months until his lawyer got another DNA test done, on more loci, which failed to match.

(In this case, the original match was only 6 loci, but since this was “a 1 in 37 million probability”, obviously “it had to be him”.)

Required disclaimer:

The views expressed above are entirely those of the writer and do not represent the views, policy or understanding of any other person or official body.

I believe the FBI is presently testing 20 loci.

Yes, there are sometimes degraded samples where it is not possible to test all 20 loci.

There is a huge difference between the following two scenarios:

(1) the sample was degraded so only X of 20 loci could be compared, and they all matched

(2) the sample was not degraded, all 20 loci were compared and only X matched

In case (1), this could be the actual suspect, and the odds of it being the actual suspect will depend on how many loci match, but if, say, you match on 13 of 20 the person in CODIS is still way more than 99.99999999% likely to be the source of the crime scene DNA

In case (2) this is definitely NOT the suspect. No prosecutor would proceed with a case against someone who had a definite mismatch on even one locus.

Some states allow familial matches.

If the crime scene STR profile is uploaded to CODIS and matches a profile on at least one of the two markers on every locus, it is extremely likely that the person whose profile is in CODIS is either the parent or the child of the person whose DNA was found at the crime scene.

In that case, they do additional testing to rule out false positives: both Y-DNA (if both samples are from males) and mitochondrial DNA (mtDNA).

A mismatch on either would exclude the CODIS sample as a false positive.

Also, in terms of the birthday paradox, we're not talking about the odds of "any two people in Arizona having the same profile." We're talking about the odds of a *particular* individual person's DNA being one of those that has a false positive match. Surely that probability is much much lower?

 [Subscribe to comments on this entry](#)

Leave a comment

[Blog moderation policy](#)

[Login](#)

Name

Email

URL:

☐ Remember personal info?

Fill in the blank: the name of this blog is Schneier on _____ (required):

Comments:

Allowed HTML `<a href="URL">` • `<em>` `<cite>` `<i>` • `<strong>` `<b>` • `<sub>` `<sup>` • `<ul>` `<ol>` `<li>` • `<blockquote>` `<pre>`
Markdown Extra syntax via <https://michelf.ca/projects/php-markdown/extra/>

Preview ☐

Notify me of follow-up comments by email.

☐

Notify me of new posts by email.

Submit

[← *Mythbusters* Episode on RFID Security Nixed](#)

[Turning off Fire Hydrants in the Name of Terrorism](#)

⇒

Sidebar photo of Bruce Schneier by Joe MacInnis.
[Powered by WordPress](#) [Hosted by Pressable](#)