



After Genetic Privacy: an Interview with Yaniv Erlich

HEATHER DEWEY-HAGBORG

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

heather@deweyhagborg.com

DATE RECEIVED:

June 10, 2015

DOI:

10.15200/winn.142767.70892

ARCHIVED:

March 29, 2015

CITATION:

Heather Dewey-Hagborg, After Genetic Privacy: an Interview with Yaniv Erlich, *The Winnower* 2:e142767.70892, 2015, DOI: [10.15200/winn.142767.70892](https://doi.org/10.15200/winn.142767.70892)

© Dewey-Hagborg This article is distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.



In 2013, Yaniv Erlich's genetics lab at MIT (now at Columbia) called the entire possibility of genetic anonymity into question when they discovered the identities of DNA donors by cross-referencing their genetic data with publicly available information from genealogy databases. Their article "Identifying Personal Genomes by Surname Inference"⁽¹⁾ published in *Science* created a stir across privacy and medical research communities.

*In 2013, Yaniv Erlich's **genetics lab** at MIT (now at Columbia) called the entire possibility of genetic anonymity into question when they discovered the identities of DNA donors by cross-referencing their genetic data with publicly available information from genealogy databases. Their article "Identifying Personal Genomes by Surname Inference"⁽¹⁾ published in Science created a stir across privacy and medical research communities.*

Heather Dewey-Hagborg: In your own words, can you give us a brief explanation of the study? What did you do and what did it mean to you?

Yaniv Erlich: We showed that it is possible in some cases to infer the surnames of males from their allegedly de-identified DNA samples. In most societies, a male receives his surname from his father, who received his surname from his own father and so on. Now, since males receive their Y

chromosome from their father and the father of their father, this process creates a correlation between surnames and y chromosomes.

Our technique exploits this correlation to identify the surname of individuals and uses open genetic genealogy databases to infer the right surname. Surnames are strong identifiers. Correctly inferring them dramatically narrows the search space. We specifically showed that if the age and state of the targeted individual are known (HIPAA does not protect these two identifiers), then a surname inference can virtually resolve the identity of the person.

To show that this technique works, we were able to identify with extremely high probabilities close to 50 people that were part of a large scale study, called the 1000 Genomes.

HDH: What were your intentions with the research? I find the project fascinating and inspiring, in part because from an artist's perspective it feels almost like an intervention – a provocative artistic act planted in the context it is meant to critique. And relatedly, how do you get funding for controversial research like this?

YE: We are entering the era of ubiquitous genetic information and there is a lot of excitement in the field about the ability of DNA sequencing to make major strides in precision medicine. Sharing genetic information is crucial to identify the genetic basis of a large spectrum of conditions, including devastating childhood diseases. The reason of our study was to highlight potential gaps in genetic privacy. By that our hope was to start a dialogue between different stakeholders in the community, facilitate the development of technological, ethical, and legal safeguards, and help research participants to make informed decisions.

We were well aware that this was a controversial research. However, computer security has a long history of "Whitehat Hacking", where security gaps are actively mapped and published in order to make the our technology ecosystem more robust. The key point was to be open from the very beginning and to constantly communicate the results to different stakeholders. I was talking about this research in conferences for over a year before the paper was out. We informed the leadership of the 1000 Genomes project about our results in close to real time after we obtained them. We let the NIH know about our study and even delayed our manuscript by a few weeks to give them time to assess their response. I also consulted with key people in the field that previously studied other genetic privacy loopholes to learn from their experience how to communicate the risk.

Looking back two year after this research has been published – I feel that we achieved our goals.

Funding was not a problem – my lab at MIT was fortunate to be supported by a generous gift of a wonderful couple, Paul and Andrea Heafy, that enabled many of our studies and allowed us to be creative.

HDH: The obvious question is why anyone should contribute to medical research if it could be used against them? Things that might not even seem possible to know today (like surnames would have seen 3 years ago) could easily arise in the future. It would seem that medical research appears to be knowingly taking advantage of vulnerable populations?

YE: The answer is simple. We are all going to be sick in some point of our life. The people that we love the most are going to be sick.

In those moments, we hope to get the best medical treatment. Medicine will not be able to take advantage of the genetic revolution without massive collection of DNA information from healthy and sick individuals and exchange this information between researchers and clinicians. This is why it is so important to map those privacy issues, plan the right safeguards, and explain the risk and benefits (and there are many benefits) for our research participants.

Of course, everyone should make her or his own decision and we should respect that. As a researcher my role is to provide the best conditions to enable participants share their data.

HDH: I read in the recent Science special issue on privacy(2) that you are working with others to develop trust frameworks in biomedical research – can you explain exactly what you are advocating and what is being developed?

YE: The current regulatory framework was developed decades ago and is heavily biased on protecting participants from researchers. One of the cornerstones of these protections is to de-identify information to retain the scientific value of the data without the ability to harm the originator. But science and especially genetics has dramatically has changed. It is impossible to de-identify DNA. Your work showed that nicely – we leave traces everywhere! Therefore, we suggest building a framework that is based on trust rather than wasting time and energy to somehow protect genetic privacy.

We have many good examples of building trust. Look at peer-to-peer economy (e.g. AirBnB, ebay, Uber). Just a few years ago, it would have been totally bizarre to enter the house or car of a total stranger that you find online*. But the combination of a trusted mediator (e.g. AirBnB), compensation mechanisms, a reputation system, and a code of conduct, it is now possible to establish trust in total strangers for defined tasks. We posit that similar trust centric frameworks can be built for genetic information in order to facilitate participation without false promises of genetic privacy.

(*Needless to say that by using any of these services, you already accept risks that are related to your DNA since you are going to leave it in his house or car)

HDH: In the same article they state that “In August, the National Institutes of Health announced that, starting this month, it expects researchers to obtain informed consent from participants if their DNA, cell lines, tissue, or any other de-identified biological material will be used for research at any point in the future.”

There was no footnote to this statement and I was left wondering just what this means. It seems to imply consent is not obtained today. What is the lay of the land regarding consent currently and how would these new guidelines change it? And further, how would they be enforced?

YE: The Common Rule (the rule that regulates all human subject research in the US) states that if a specimen is part of an existing collection and is de-identified then it is not considered as human subject research. The new policy states that consent is necessary because we cannot treat that as de-identified information.

HDH: Two years after your publication is progress being made? Do you approve with the direction things are moving in outside your own team's R&D?

YE: The main progress was the wide acceptance that genetic privacy of personal genomes is a myth (as you can see in the previous question) Before I published this manuscript, the main criticism was around our ability to truly identify individuals. After it was published, the main criticism became that the this risk is quite known. Sometimes I heard both of these lines of criticism from the same person!

HDH: Are there any specific policy changes you would like to see? Or are there spheres beyond policy you want to see change made in?

YE: The answer for this question is quite long but here is one of the main things that I would like to

see:

We have state and nation-wide frameworks of consent for organ donation, a procedure that is quite unpleasant to think of, and by doing so, you also risk your genetic privacy (your liver has quite a lot of DNA!). Yet, a simple signature is enough to donate your organs (in some counties the system is even opt-out).

Donating DNA after death with medical records can have a tremendous impact on biomedical research. I would like to see some discussion about a policy that will enable people to donate their medical information after death as part of obtaining a driver license.

1. Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. "Identifying Personal Genomes by Surname Inference." *Science* 339, no. 6117 (January 18, 2013): 321–24. doi:10.1126/science.1229566.
1. Couzin-Frankel, Jennifer. "Trust Me, I'm a Medical Researcher." *Science* 347, no. 6221 (January 30, 2015): 501–3. doi:10.1126/science.347.6221.501.