Genome Hacking

Yaniv Erlich
Whitehead Institute for Biomedical Research

Twitter: @erlichya
Public data is important for genetic studies. To make this endeavor sustainable, we must proactively map risks.
Co-segregation between Y-chr and surnames

Intro.
Risk assessment
The Venter case
Anonymous datasets
Summary

www.ysearch.org:

Displaying User ID: CEEPG

Haplogroup: Unknown
Last name: Erlich
Variant spellings: Erlich

5/13/13 Yaniv Erlich
Exploiting genetic genealogy databases

Found on the Web, With DNA: a Boy’s Father

By Rob Stein
Washington Post Staff Writer
Sunday, November 13, 2006

Like many children whose mothers used an anonymous sperm donor, the 15-year-old boy longed for any shred of information about his biological father. But, uniquely, this resourceful teenager decided to try exploiting the latest in genetic technology and the sleuthing powers of the Internet in his quest.

By submitting a DNA sample to a commercial genetic database service designed to help people draw their family tree, the youth found a crucial clue that quickly enabled him to track down his long-sought parent.

"I was stunned," said Wendy Kramer, whose online registry for children trying to find anonymous donors of sperm or egg helped lead the teenager to his father. "This had never been done before. No one knew you could get a DNA test and find your donor."

While welcomed by advocates of children trying to locate anonymous donors, the case -- apparently the first of its kind -- has raised alarm among sperm banks and some medical ethicists. They are concerned it might start a trend that could violate the privacy of thousands of sperm donors and discourage future ones.

An anecdote?
The main idea – a **systematic study**

Can we recover the identity of anonymous sequencing datasets using public resources?
Empirical test: what is the probability to recover a surname?

Expectation for US Caucasian males from middle and upper class: 12% Successful recoveries
The Venter case

- We got a surname from whole genome sequencing data

lobSTR: short tandem repeat profiler for personal genomes

Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich

Summary

Try it yourself: bit.ly/craig_venter_haplotype_updated
Can we identify anonymous personal genomes?
Recovering the identities of CEU individuals

1000 Genomes

EMBL-EBI

10 CEU genomes

Found an obituary that has the exact description of the pedigree

Google

Winfield Utah

8 Surname predictions with Utah ancestry

Probability of a random match < 5x10^{-9}
Beginner’s luck?

In total:
5 successful surname recoveries
Breaching the privacy of close to 50 CEU samples.

- Successful surname recovery (targeted individual)
- Person tested by genetic genealogy service (source)
- Patrilineal line from source to target

Breaching the privacy of close to 50 CEU samples.
Summary

Our approach:

- No experimental work involved.
- The identifying information propagates via deep genealogical ties.
- The attack completely relies on public resources.

Testing close to 1000 Y-STR haplotypes, demonstrating complete identification of Venter and close to 50 CEU individuals.
IMHO, recommendations

1. Consent:
   - Be honest about risks. Be honest about benefits.

2. Multi-tier approach:
   - Give participants options for data sharing.

3. Proactive approach:
   - Keep mapping risks. Friendly hacking is far better than a real one.

4. Technical solutions:
   - We did not explore those enough. Much more to do here.
Acknowledgements

Melissa Gymrek (HST – Harvard/MIT)
Amy McGuire (Baylor)
David Golan (Tel-Aviv University)
Eran Halperin (Tel-Aviv University)

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,1,2,3,4 Amy L. McGuire,5 David Golan,6 Eran Halperin,7,8,9 Yaniv Erlich1*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Funding:
Andria and Paul Heavy
Jim and Cathy Stone

Open Access (with FREE registration)