We need to share genetic information

Hereditary Spastic Paraparesis (Erlich et al.)

Joubert syndrome (Endevson et al.)

Hemifacial Microsomia (Zielinski, .., & Erlich)

Genetic Privacy
Vulnerability research

Intro.

Methodology

The Venter case

Anonymous datasets

Summary

Intercom

Fingerprint reader

Me

IT department of a major bank
Correlation between Y-chr and surnames

www.ysearch.org:

Labeled chart showing genetic data for Smith and Erlich surname matches.

Genetic privacy
The main idea

A systematic study: can we recover the identity of anonymous genomic datasets?
Databases of interest

140,000 publicly accessible surname-Ychr records

www.smgf.org

www.ysearch.org
How to find surnames?

Estimating the **time** to most recent common ancestor

\[
t_i \quad \text{Target in db}
\]

\[
t_i' \quad \text{Surname}
\]

\[
t_i' \quad \text{i-th record in db}
\]
Empirical test to determine the probability of recovering a US surname

For US Caucasian males:
- **12%** Successful recoveries
- **5%** Wrong recoveries
- **83%** Unknown
Distribution of inferred surnames

Most of the inferred surnames are relatively rare
Triangulate individuals with metadata

Age

State

Surname

The median of age+state+surname is 12 males.
Putting it all together: the Venter case

We got a surname from whole-genome sequencing data.

lobSTR: A short tandem repeat profiler for personal genomes

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

1Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 2Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; 3Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel

ySearch

Try it yourself: bit.ly/find_craig
Getting to Craig Venter

Searching for:
1. Venter
2. California
3. Born in 1946
4. Male

In USSearch.com

Two matches, including:

1. J Craig Venter
   John Venter
   Jcraig Venter

   La Jolla, CA
   Carlsbad, CA
   Mountain View, CA
   La Mirada, CA
   Arlington, VA

   Claire Fraser
   Heather Kowalski
   Melanie Wranaker
   Robert Fraser

   Gsac
   Guardian
   New England
   Wire Products Inc

   More Jobs
Can we identify anonymous genomes?
1000 Genomes cases

Found an obituary that has the exact description of the pedigree

Probability of a random match < 5x10^-9

*Some of the details in this slide were modified to respect the identity of the family
Beginner’s luck?

Breaching the privacy of close to 50 CEU samples.
Our study

Identifying Personal Genomes by Surname Inference

Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich

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...
The hitchhiker guide to genome hacking

Routes for breaching and protecting genetic privacy

Yaniv Erlich\textsuperscript{1} and Arvind Narayanan\textsuperscript{2}

Abstract | We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these data sets is vital for progress in biomedical research. However, a growing concern is the ability to protect the genetic privacy of the data originators. Here, we present an overview of genetic privacy breaching strategies. We outline the principles of each technique, indicate the underlying assumptions, and assess their technological complexity and maturation. We then review potential mitigation methods for privacy-preserving dissemination of sensitive data and highlight different cases that are relevant to genetic applications.
On Jim Watson’s *APOE* status: genetic information is hard to hide


---

Alzheimer’s disease

Barack Obama is the President
The path forward

Redefining Genomic Privacy: Trust and Empowerment

Yaniv Erlich¹*, James B. Williams², David Glazer², Kenneth Yocum³, Nita Farahany⁴, Maynard Olson⁵, Arvind Narayanan⁶, Lincoln D. Stein⁷,⁸, Jan A. Witkowski⁹, Robert C. Kain³

Key points: transparency, reputation system, compensation
Acknowledgements

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

Funding:

Andria and Paul Heafy