An example of DNA re-identification attack

Vinicius Vielmo Cogo
FCUL, Bioinformatics, May 15, 2013
Main goal

Presentation goal

• Alerting researchers for their responsibility in maintaining the privacy of project participants

The method

• Presenting a step-by-step example of how a simple re-identification attack works and can compromise the privacy of project participants

(The idea is not to teach you how to attack a public project, but only call to reflection on this topic)
Sequenced human genomes around the world

Oct. 2010: 2683

Oct. 2011 (estimated): 30358
The cost per genome is decreasing

Sequencing cost per genome

Number of sequenced genomes
The number of research databases is increasing.
What is private/public data for them (and HIPAA)?

Private:
- Donor name
- Identity number
- Social security number
- Home address

Public:
- DNA sequence
- Age
- State
Which privacy guarantees they provide?

In the participant consent ...

- Projects will **try to maintain** donor privacy, but they **cannot guarantee** it.
Identifying Personal Genomes by Surname Inference

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

Science (339). January 18, 2013
DOI: 10.1126/science.1229566
Goals:

1) **Recover the surname** of sequence donors from 1000 Genome Project

2) **Triangulate the identity** of a sequence donor using his surname, age and state

Based on two facts:

a) Surnames are paternally inherited in most human societies

b) As well as Y-chromosome haplotypes in male individuals
For surname inference:

- Profiling short tandem repeats (STR) on the Y-chromosome
- Querying recreational genetic genealogy databases.
- Obtaining a list of possible surnames for the sequence in question
The method

For surname inference:
- Profiling short tandem repeats (STR) on the Y-chromosome
- Querying recreational genetic genealogy databases.
- Obtaining a list of possible surnames for the sequence in question

What is a STR?
A STR is a small repeating sequence in DNA.

STR Name: **DYS392**
Location: Chromosome Y
Initial Position: 22633873
Final Position: 22633911
Representation: \([\text{TAT}]_n\)

Example of \([\text{TAT}]_5\):
cgac TAT TAT TAT TAT TAT tcga

P.S.:
- FBI uses 13 known STRs distributed in the entire DNA
- This article uses 30 known STRs from chromosome Y
The method

For surname inference:
- Profiling short tandem repeats (STR) on the Y-chromosome
- Querying recreational genetic genealogy databases.
- Obtaining a list of possible surnames for the sequence in question

Triangulate identity:
- Combining surnames with age and state
- Triangulating the identity of the target

P.S.: All attack is done using only free and public Internet resources and data
The tools

Welcome
Much has happened since Y-DNA testing first became available commercially through Family Tree DNA in February of 2000. Many thousands of people have tested to find family connections as well as family origins. Since then, other labs have entered this market, and the number of tested individuals is growing as the use of DNA is becoming more and more accepted as an important tool for family research, enhancing traditional genealogy research methods.

In order to allow people that have tested with the different companies to make their results available for comparison, Family Tree DNA is offering Ysearch as a free public service. We have added several tools that allow you to compare side-by-side different users - the YsearchCompare - as well as generate a Genetic Distance™ Report, and many other features, including the upload of GEDCOM files.

Have you not tested yet? Order your test at Family Tree DNA

Have you not tested with Family Tree DNA? Check this very special deal!

Already have an account? Please upload your GEDCOM. (What is a GEDCOM?)

What next?
- Create a new user
- Search for genetic matches
- Search by last name
- Edit an existing user

Size of the database?
- Surname Count: 85179
- Unique Haplotypes: 94309
- Number of Records: 124366
- Family Tree DNA - 108153
- Rotativo Genetics - 1047
- Oxford Ancestors - 425
- Other - 14641

The Sorenson Database is the foremost collection of genetic genealogy data in the world. Search by DNA results or surname and find your place in the worldwide genetic family tree.

Great News!
We are pleased to announce that Ancestry.com DNA has acquired GeneTree and the DNA related assets from the Sorenson Molecular Genealogy Foundation. We are excited to work with Ancestry.com DNA and continue to advance the field of genetic genealogy. More information to come. Click here for the announcement about this exciting news.
1) Obtain the genome from 1000 Genome Project

2) Profile the Y-STRs with the lobSTR (count the number of repeats on each known STR for the genome in question)

3) Apply the resulting profile in Ysearch

4) Get the list of candidates
The results

For surname inference:

- Authors recovered the correct surname in 12% of cases (based on a threshold of 82% of confidence)
- If wrong surnames are accepted, then it goes to 18%
Triangulate identity:
- A query on U.S. census by year of birth and state results in 60000 U.S. males in 50% of cases.
- Aggregating the surname to the query shrinks to only 12 males
The results

**Triangulate identity:**

- Each surname inference breached the privacy of nearly 16 individuals
- From the number of entries in those tools, one can identify millions of U.S. males
Solutions?

Incorrect

- Stop donating samples to projects
- Stop sharing data

Correct

- Create policies for data sharing
- Educate participants about benefits and risks
- Create proper legislations about the usage of genetic information
Contributions

- Alerts to the problem of re-identification in public sequence databases
- Proves the weakness of current anonymization techniques

Limitations

- Works only for patrilineal relatives (Is it possible for matrilineal relatives if using mtDNA?)
- Authors do not provide an immediate solution to solve the problem
Final remarks

- 12% seems to be a weak result, as it is not the majority

- But it means that from 1092 genomes (of 1000 genome project), 131 participants will never recover their privacy*

- Nearly 2100 people (participant’s relatives) had their privacy breached by this attack

* If one has his password discovered, it can ask to the responsible by the service to change it. We cannot ask nobody to change our DNA.