Security Risks to Third-Party Genetic Genealogy Services

Peter Ney, Luis Ceze, Tadayoshi Kohno
Direct-to-Consumer (DTC) Genetic Testing and Analysis

- Genetic Interpretation
  - Health, Ethnicity, Relative Prediction, ...

- Raw Genetic Data

- DTC Testing Company
  - 23andMe
  - AncestryDNA
  - MyHeritage
  - FamilyTreeDNA
Direct-to-Consumer (DTC) Genetic Testing and Analysis

Genetic Interpretation
Health, Ethnicity, Relative Prediction, ...

Raw Genetic Data

23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

3rd-Party Genetic Service

DTC Testing Company
Direct-to-Consumer (DTC) Genetic Testing and Analysis

Genetic Interpretation
Health, Ethnicity, Relative Prediction, ...

Raw Genetic Data

DTC Testing Company
23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

Research Focus

Genetic Interpretation
Health, Ethnicity, Relative Prediction, ...

3rd-Party Genetic Service
Third-Party Genetic Genealogy Services

Alice's Genetic Data

Relative Matching
Bob is Alice's Sibling
Frank is Alice's 2nd-Cousin

Genetic Genealogy Database

Bob  Carol
Dan  Frank

1M+
Relative Matching Algorithms

- Long shared segments of DNA are indicative of recent shared ancestry
- More and longer shared segments mean a closer relationship
- Relative matching algorithms try to identify these shared segments between users
Prior Attacks Against Genetic Genealogy Services: Identity Inference

Goal: identify the source (person) of an anonymous DNA sample or genetic data
Prior Attacks Against Genetic Genealogy Services: Identity Inference

**Step 1**

- Anonymous DNA sample or genetic data
  - Research Dataset
  - Crime Scene

Process sample and construct genetic files

DTC Genetic Data (Unknown)
Prior Attacks Against Genetic Genealogy Services: Identity Inference

Step 2

Malory

Unknown Genetic Data

Relative Matching
Carol is a grandmother
Frank is a cousin

Genetic Genealogy Database

Bob
Carol
Dan
Frank

... 1M+
Prior Attacks Against Genetic Genealogy Services: Identity Inference

**Step 3:** Combine the relatives with other sources of information like genealogies to identify the source of the sample or data

---

**Law enforcement**
- 100+ samples identified from crimes and unknown remains
- Suspected Golden State Killer

**Anonymous research data**
- Ex: 1000 Genomes Data (Erlich et al. Science. 2018)
**Attack 1: Extract Genetic Markers from Other Users**

Malory

Bob

Artificial or Manipulated Genetic Data

Relative Matching Queries

Matching Segments and Visualizations

Genetic Genealogy Database

Bob

Carol

Dan

Frank

1M+
Attack 2: Forge Genetic Relationships

Genetic Genealogy Database

Artificial or Manipulated Genetic Data

Malory is Bob’s second cousin

Malory is Bob's second cousin

Bob
Carol
Dan
Frank

1M+
Case Study on GEDmatch

- GEDmatch runs the largest third-party DTC genetic genealogy service
  - Over 1.2 millions files have been uploaded
- Used extensively by law enforcement
  - Used to solve Golden State Killer case
  - Government contracting (Parabon Nanolabs)
    - Unidentified remains (DNA Doe Project)
- Identity inference attacks demonstrated on GEDmatch (*Erlich et al. Science. 2018*)
- Goal is to evaluate the feasibility of these new attacks on GEDmatch
Experimental Setup

Account 1
Normal User

Account 2
Adversary

GEDmatch

Experimental Genetic Profiles

Artificial data

Relative Matching Queries

Relative Results and Visualizations
Ethics of Data Uploads and Queries

- Uploaded all data to a sandboxed “Research” setting so that the uploaded files would not interact with real GEDmatch users.
- Only ran queries with and analyzed results from data that we uploaded.
  - GEDmatch lets you target relative matching queries against specific data files.
- ToS allowed artificial data uploads if:
  - (1) Intended for research.
  - (2) Not used to identify anyone in the database.
- IRB determined that research was exempt from review because the experimental data was derived from public sources with no identifiers.
Attack 1: Extract Genetic Markers from Other Users

Genetic Genealogy Database

Malory

Artificial or Manipulated Genetic Data

Relative Matching Queries

Matching Segments and Visualizations

1M+

Bob

Carol

Dan

Frank
Both visualizations leak information about the underlying DNA markers in other genetic files.
GEDmatch Visualizations and Segments

<table>
<thead>
<tr>
<th>Chr</th>
<th>B37 Start Pos'n</th>
<th>B37 End Pos'n</th>
<th>Centimorgans (cM)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18,893,763</td>
<td>64,073,387</td>
<td>54.2</td>
<td>7,506</td>
</tr>
<tr>
<td>1</td>
<td>159,815,357</td>
<td>164,468,815</td>
<td>9.6</td>
<td>970</td>
</tr>
</tbody>
</table>

Both visualizations leak information about the underlying DNA markers in other genetic files.
Genetic Extraction via Marker Visualizations

Each pixel corresponds to a single genetic marker (many are missing)
Genetic Extraction via Marker Visualizations

Each pixel corresponds to a single genetic marker (many are missing)

Markers same
Markers half-match
Markers different

Known
Unknown

Relative Matching Queries
Genetic Extraction via Marker Visualizations

**Step 1**
Run 20 relative matching queries against a target and gather visualizations

Malicious Data (known)  
Target User (unknown)
Genetic Extraction via Marker Visualizations

**Step 1**
Run 20 relative matching queries against a target and gather visualizations

- Malicious Data (known)
- Target User (unknown)

**Step 2**
Use mastermind-like algorithm to determine which pixels correspond to specific markers. (Similar to Goodrich. S&P. 2009. DNA sequence extraction via DNA sequence alignment scores.)
Genetic Extraction via Marker Visualizations

Step 3
Combine known artificial genetic markers with visualizations to infer target’s genetic markers

Malicious File

Target File
Genetic Extraction via Marker Visualizations

**Step 3**
Combine known artificial genetic markers with visualizations to infer target’s genetic markers

**Step 4**
Fill in the gaps with genetic imputation (statistical technique)

In total we were able to extract an **average of 92% of the genetic markers with 98% accuracy** from the 5 test files.
Both visualizations leak information about the underlying DNA markers in other genetic files.
**Attack 2: Forge Genetic Relationships**

Malory is Bob’s second cousin

Artificial or Manipulated Genetic Data

Genetic Genealogy Database

Bob  Carol

Dan  Frank

... 1M+
Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction
- Parent/Child: 50%
- 1st cousin: 12.5%
Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction

- Parent/Child: 50%
- 1st cousin: 12.5%

Forge segments and relationships.
Generating Artificial Relatives

Amount of DNA sharing determines the relative prediction
- Parent/Child: 50%
- 1st cousin: 12.5%

Discover target’s genetic profile using:
1) Genetic extraction attacks (shown earlier). *Tested on GEDmatch.*
2) Gather DNA sample surreptitiously and sequence it.
3) Adversary wants to forge relative for themselves.
Why Make Artificial Relatives?

1) “Long lost relative.” Not uncommon in genetic genealogy because of misidentified paternity.

2) Change inferred identity
Why Make Artificial Relatives?

1) “Long lost relative.” Not uncommon in genetic genealogy because of misidentified paternity.
2) Change inferred identity
Why Make Artificial Relatives?

1) “Long lost relative.” Not uncommon in genetic genealogy because of misidentified paternity.

2) Change inferred identity

Open question is how this could affect import inferences, like law enforcement, which is currently an expert driven and manual process.
Poor API and design choices on GEDmatch contributed significantly to the vulnerabilities we uncovered:

- Lack of data authentication / integrity checks
- High resolution visualizations
- Ability to target specific users and direct queries
- Algorithms somewhat vulnerable by design

Responsibly disclosed results to GEDmatch, who modified their visualization algorithms to mitigate data extraction attacks.

Long term changes in the DTC industry, especially data authentication, are needed to prevent attacks via malicious data uploads and ensure long term security.
Consumer genetic genealogy databases have major implications for genetic privacy:
- Used to solve crimes and results are used in court
- Relevant to genetic surveillance and anonymous genetic data
  - 1M+ database: identification is possible by not easily scalable
  - 10M+: identification is simple

Encourage the community to develop methods to make genetic genealogy more secure by design