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, ...

DTC Testing Company
23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

Raw Genetic Data
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

Research Focus

DTC Testing Company
- 23andMe
- AncestryDNA
- MyHeritage
- FamilyTreeDNA

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

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

Genetic Genealogy Database
Bob
Carol
Dan
Frank

1M+
Research Questions

1) Given the popularity of genetic genealogy services, what security and privacy issues might exist? Can these be demonstrated on a real service?

2) How does the design of a genetic genealogy service impact security? What might be done to make them more secure?
Prior Attacks Against Genetic Genealogy Services: Identity Inference

Goal: identify the source (person) of an anonymous DNA sample or genetic data

Anonymous DNA sample or genetic data

Research Dataset  Crime Scene
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

Relative Matching
Carol is a grandmother
Frank is a cousin

Genetic Genealogy Database

Unknown Genetic Data

Malory
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*)
Hypothesis #1: Can We Extract Raw Genetic Markers from Other Users in a GG Database?

Matching Segments and Visualizations

Artificial or Manipulated Genetic Data

Relative Matching Queries

Genetic Genealogy Database

Malory

Bob

Carol

Dan

Frank

1M+
Hypothesis #2: Can We Generate Artificial Relatives for Other Users in a GG Database?

Genetic Genealogy Database

Artificial or Manipulated Genetic Data

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 on GEDmatch

**Account 1**
Normal User

**Account 2**
Adversary

GEDmatch

- **Experimental Genetic Profiles**
  - $\times 5$

- **Artificial data**
  - $\times n$

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 let’s you target relative matching queries against specific data files
- ToS allowed artificial data uploads if:
  - Intended for research
  - 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
Generating DTC Data Files for Experimentation

- Include ~500,000-700,000 genetic markers throughout the genome (called SNPs)
- No standardization (each company is slightly different)
- Plain text CSV with 4 fields
  - SNP identifier
  - Chromosome #
  - Index within chromosome
  - DNA bases

<table>
<thead>
<tr>
<th>#</th>
<th>rsid</th>
<th>chr</th>
<th>pos</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs548049170</td>
<td>1</td>
<td>69869</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>rs13328684</td>
<td>1</td>
<td>74792</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs9283150</td>
<td>1</td>
<td>565508</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs116587930</td>
<td>1</td>
<td>727841</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs12184325</td>
<td>1</td>
<td>754105</td>
<td>CC</td>
<td></td>
</tr>
<tr>
<td>rs12567639</td>
<td>1</td>
<td>756268</td>
<td>AA</td>
<td></td>
</tr>
<tr>
<td>rs114525117</td>
<td>1</td>
<td>759036</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs12124819</td>
<td>1</td>
<td>776546</td>
<td>AA</td>
<td></td>
</tr>
<tr>
<td>rs12127425</td>
<td>1</td>
<td>794332</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs79373928</td>
<td>1</td>
<td>801536</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>rs72888853</td>
<td>1</td>
<td>815421</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>rs7538305</td>
<td>1</td>
<td>824398</td>
<td>AC</td>
<td></td>
</tr>
<tr>
<td>rs28444699</td>
<td>1</td>
<td>830181</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>rs116452738</td>
<td>1</td>
<td>834830</td>
<td>GG</td>
<td></td>
</tr>
</tbody>
</table>
Generating DTC Data Files for Experimentation

### DTC Genetic Data Files (23andMe v5 SNP-chip)

<table>
<thead>
<tr>
<th>rsid</th>
<th>chr</th>
<th>pos</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs548049170</td>
<td>1</td>
<td>69869</td>
<td>TT</td>
</tr>
<tr>
<td>rs13328684</td>
<td>1</td>
<td>74792</td>
<td>GG</td>
</tr>
<tr>
<td>rs9283150</td>
<td>1</td>
<td>565508</td>
<td>GG</td>
</tr>
<tr>
<td>rs116587930</td>
<td>1</td>
<td>727841</td>
<td>GG</td>
</tr>
</tbody>
</table>

### Whole genome sequence & variant data

<table>
<thead>
<tr>
<th>rsid</th>
<th>chr</th>
<th>pos</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs548049170</td>
<td>1</td>
<td>69869</td>
<td>TT</td>
</tr>
<tr>
<td>rs13328684</td>
<td>1</td>
<td>74792</td>
<td>GG</td>
</tr>
<tr>
<td>rs9283150</td>
<td>1</td>
<td>565508</td>
<td>GG</td>
</tr>
<tr>
<td>rs116587930</td>
<td>1</td>
<td>727841</td>
<td>GG</td>
</tr>
</tbody>
</table>
Generating DTC Data Files for Experimentation

Programming Tools

- Standard bioinformatics tools (e.g., samtools) to process variant files
- Python scripts to parse genetic data files, modify SNPs, process web files, and run attack algorithms

Dataset

- Sample size for testing was small (5 target files) and all 23andMe files. Choose this to limit impact on the GEDmatch service.
- 1000 Genomes data came from same sub-population
Relative Matching on GEDmatch

Chromosome 7

- Long shared segments of DNA are indicative of recent shared ancestry
- More and longer shared segments means a closer relationship
- Relative matching algorithms try to identify these shared segments between users
- GEDmatch uses proprietary algorithms to identify matching DNA segments
Populated User Account with Genetic Data Files

Uploaded Genetic Data Files

- Q5620531
  - Margaret B
- BN861059
  - Paul B.
- PU6714417
  - Mary B
- AR5198750
  - Rebecca R.
- NU6088065
  - Robert J

You have not uploaded any GEDCOM (genealogy) resources
Relative Matching on GEDmatch

Direct relative matching query between two users

Easily scrape the query results and visualizations

Coordinates of IBD Segments

Chromosome Visualization

Largest segment = 85.3 cM
Total Half-Match segments (HIR) = 256.5 cM (7.151 Pct)
Estimated number of generations to MRCA = 2.9
5 shared segments found for this comparison.
485673 SNPs used for this comparison.
70.276 Pct SNPs are full identical
Comparison took 0.257 seconds.
CPU time used: 0.044 cpu seconds...
Hypothesis #1: Can We Extract Raw Genetic Markers from Other Users in a GG Database?

Artificial or Manipulated Genetic Data

Matching Segments and Visualizations

Relative Matching Queries

Genetic Genealogy Database

Bob

Malory

Carol

Dan

Frank

1M+
GEDmatch Visualizations and Segments

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

Matching algorithms and visualizations were proprietary so it was necessary to run a number of experiments to figure out how they were working.

Regular file  Modified data file
Matching algorithms and visualizations were proprietary so it was necessary to run a number of experiments to figure out how they were working.

Hypothesis

1) At high resolution these pixels seemed to correspond to individual markers

2) Many markers seemed to be missing

3) Results not phased

Regular file  Modified data file

- GT == TG
- GG == TG
- GG == TT
Matching algorithms and visualizations were proprietary so it was necessary to run a number of experiments to figure out how they were working.

### Hypothesis

A section of chromosome is considered a shared segment if the files match on a single base for a run of consecutive markers.

<table>
<thead>
<tr>
<th>rsid</th>
<th>chr</th>
<th>pos</th>
<th>genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs548049170</td>
<td>1</td>
<td>69869</td>
<td>TT</td>
</tr>
<tr>
<td>rs13328684</td>
<td>1</td>
<td>74792</td>
<td>GG</td>
</tr>
<tr>
<td>rs9283150</td>
<td>1</td>
<td>565508</td>
<td>GG</td>
</tr>
<tr>
<td>rs116587930</td>
<td>1</td>
<td>727841</td>
<td>GG</td>
</tr>
<tr>
<td>rs3131972</td>
<td>1</td>
<td>752721</td>
<td>GG</td>
</tr>
<tr>
<td>rs12184325</td>
<td>1</td>
<td>754105</td>
<td>CC</td>
</tr>
<tr>
<td>rs12567639</td>
<td>1</td>
<td>756268</td>
<td>AA</td>
</tr>
</tbody>
</table>

**Regular file**

**Modified data file**
Genetic Extraction Experiments with Marker Visualizations

Ran attack 5 times (one for each experimental file)

Collected visualizations from Chrome browser (20 comparisons x 22 autosomes = 440 per attack)

Process visualizations with python scripts implementing a mastermind-like algorithm to infer which markers went with which pixels
Genetic Extraction Experiments with Marker Visualizations

Fill in the gaps using a statistical technique called genetic imputation. Relied on a publicly available genetic imputation service run by the Sanger Institute.
Genetic Extraction Experiments with Marker Visualizations

In total we were able to extract an average of 92% of the genetic markers with 98% accuracy from the 5 test file.

The first round of inference was without error in all runs. All of the error was due to the statistical inference of missing SNPs (imputation).

There was a small difference in which SNPs could be recovered but stayed mostly consistent.
Genetic Extraction with Matching Segments

Long run of heterozygous markers will always produce a matching DNA segment against any person because SNPs only have two possible bases (bi-allelic).
Genetic Extraction with Matching Segments

Single homozygous marker

Malicious Data

| A | C | T | G | C | A | T | C | G |
+---+---+---+---+---+---+---+---+---|

Target

+---+---+---+---+---+---+---+---+---|

Presence or absence of a DNA segment can be used to infer individual markers in any target. 
Validated this attack on multiple markers with similar approach as before.
Hypothesis #2: Can We Generate Artificial Relatives for Other Users in a GG Database?

Genetic Genealogy Database

Artificial or Manipulated Genetic Data

Malory is Bob’s second cousin

1M+
Experimenting with Artificial Relatives

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

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

Target

Known

Artificial

Generate

Relative Matching

Forge segments and relationships.
Experimenting with 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. Validated on GEDmatch.
2) Gather DNA sample surreptitiously and sequence it.
3) Adversary wants to forge relative for themselves.
Experimenting with Artificial Relatives

Marker Extraction Attack

GEDmatch

Modified genetic files to appear like relatives

Expected relative prediction returned
Experimentation Artifacts Borrowed from the Community?

- Mostly not
- Big challenge was finding good datasets for experimentation
  - Very little public data is available from direct-to-consumer testing sources
  - No standards or documentation on DTC file formats
- Required to make most of the experimental pipelines from scratch
Reproducing Results?

- Replicated part of prior methods to generate DTC files from variant data
  - Code was not easily available and had to be written from scratch
- Other groups have partially replicated these attacks both on GEDmatch and in simulation. *Edge and Coop. ELife*. 2020.
Failed / Unsuccessful Experiment: Disrupting Identity Inference
Failed / Unsuccessful Experiment: Disrupting Identity Inference

2nd-Cousin (artificial)  2nd-Cousin

[Family tree diagram]
Failed / Unsuccessful Experiment: Disrupting Identity Inference

Search occurs on wrong branch of tree

Falsely predicted relatives

2nd-Cousin (artificial)  2nd-Cousin
Failed / Unsuccessful Experiment: Disrupting Identity Inference

- How do you run experiments that take genealogies / family trees into account?
- Family tree data is available
  - 1M+ person trees meant for research
- Tried to run simulations to see how easily a random individual could be mis-identified
  - Depends on tree topology and number of relatives in the genetic genealogy database
- Issue: Real inferences are a messy and trees are often wrong (misattributed parentage)
  - Hard to generate convincing experiments
GEDmatch: Matches at ~400 cM and ~200 cM, no shared DNA
Ancestry: ~90% N European, ~10% non-European (~1/8)
Genealogy: The matches’ family trees do not intersect on paper, but match #2’s half-uncle lived in the same town as match #1’s grandparents when match #1’s aunt was conceived, suggesting a non-paternity event (NPE) between these families. That (half-)aunt has a grandson with 1/8 non-European ancestry who is a half-1st cousin to match #1 and a half-1st cousin once-removed to match #2.
Outcome: Abandoned DNA matched to crime scene DNA
Failed / Unsuccessful Experiment: Studies of Other Services

● Strongly considered testing these attacks on other services
  ○ DNA.land: the other major 3rd-party genetic genealogy service

● Big challenge is ToS / ethics considerations
  ○ Different rules about artificial uploads
  ○ No ability to restrict uploads so they don’t affect other users

● May be possible to partially simulate these attacks but results are much less convincing / realistic
Experimental Artifacts?

Release of code and data is in progress. Includes:
- Datasets used in all experiments
- Code to generate and manipulate consumer genetic data files
- Code implementing the extraction algorithms
- Visualizations and other web files to replicate results
What Can be Learned from Your Methodology?

● The use of artificial genetic data sets is a powerful way to query and potentially attack genetic databases.
  ○ Broadly applicable to research in genome privacy
● Good data sets and tooling could make this much easier
● Experimenting with a live service is challenging but important because small design choices make a really big difference
  ○ ToS and ethics are a big constraint on what you can test