Security Risks to Third-Party Genetic Genealogy Services

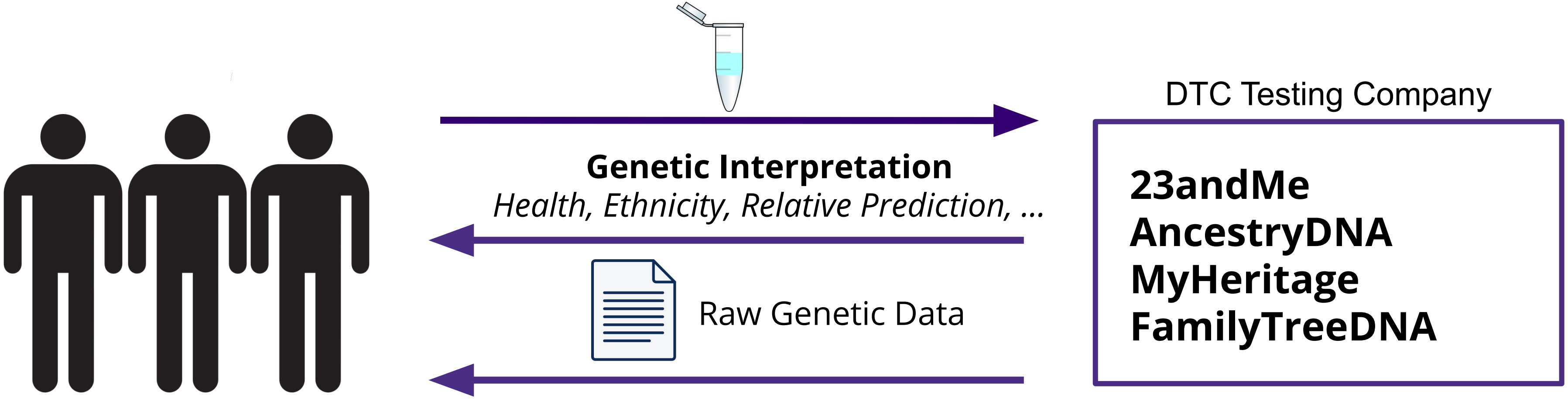
Peter Ney, Luis Ceze, Tadayoshi Kohno

BE BOUNDLESS

SECURITY & PRIVACY RESEARCH LAB UNIVERSITY OF WASHINGTON

MISL

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



Genetic Interpretation

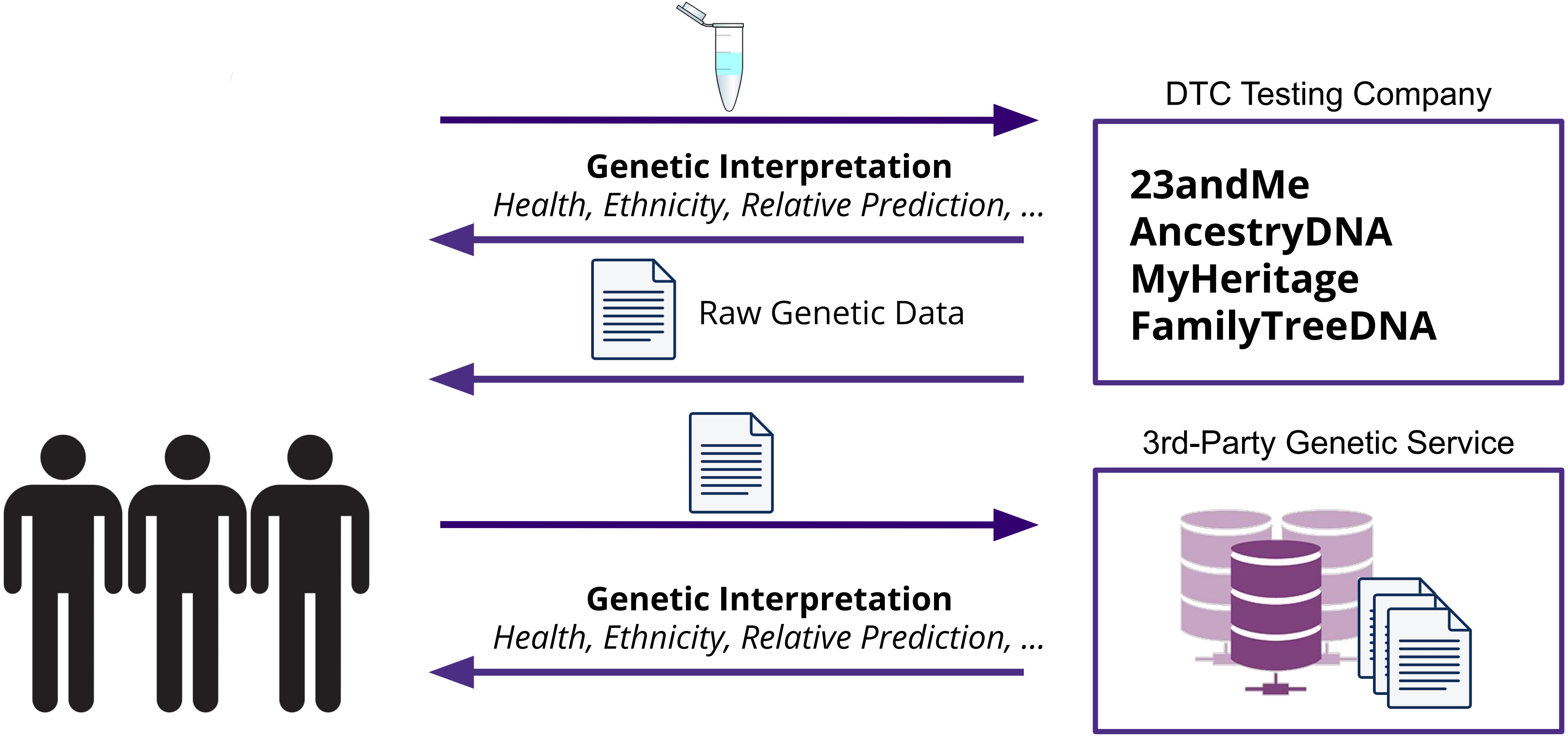
Health, Ethnicity, Relative Prediction, ...

Raw Genetic Data

DTC Testing Company

23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

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



DTC Testing Company

23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

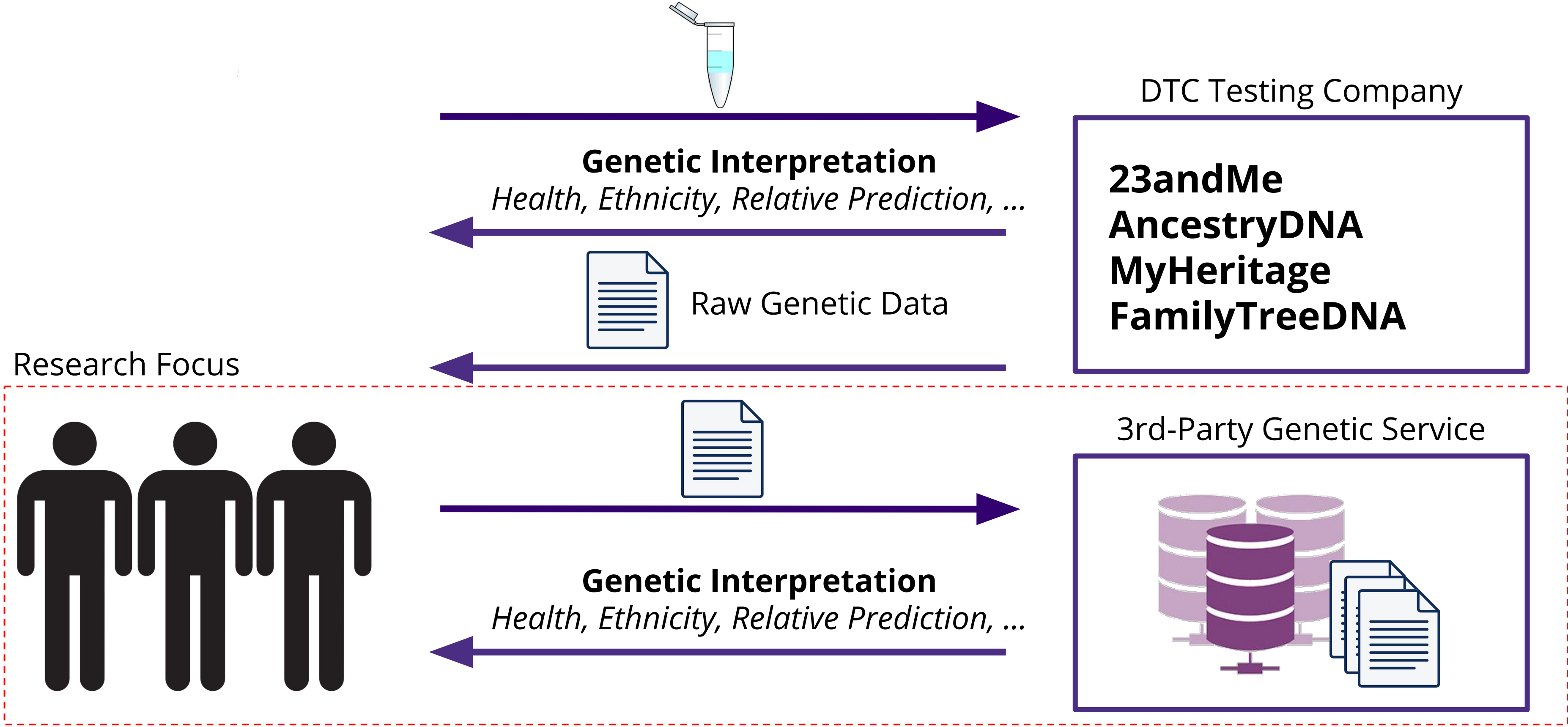
Genetic Interpretation
Health, Ethnicity, Relative Prediction, …

Raw Genetic Data

3rd-Party Genetic Service

Genetic Interpretation
Health, Ethnicity, Relative Prediction, …

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



DTC Testing Company

23andMe
AncestryDNA
MyHeritage
FamilyTreeDNA

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

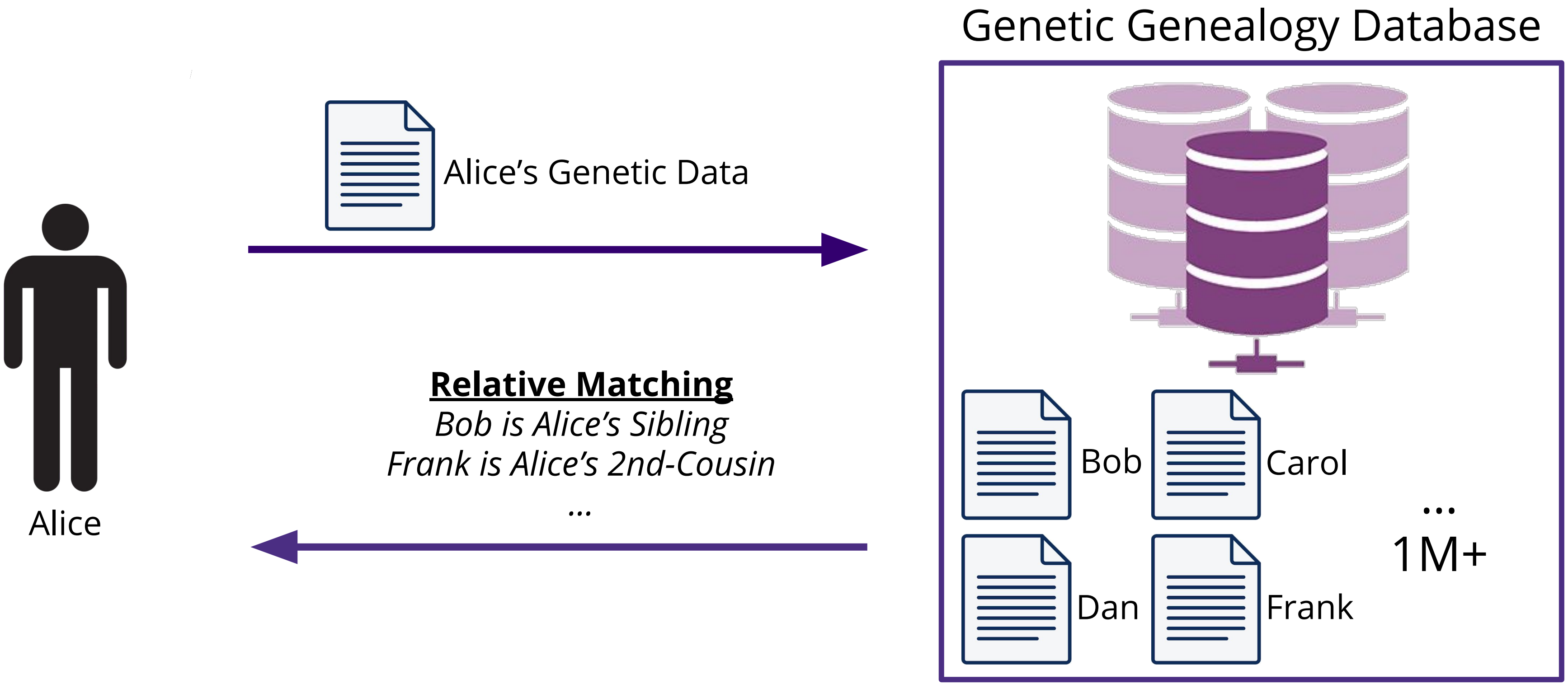
Raw Genetic Data

Research Focus

3rd-Party Genetic Service

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

Third-Party Genetic Genealogy Services



Genetic Genealogy Database

Alice's Genetic Data

Alice

Relative Matching

Bob is Alice's Sibling

Frank is Alice's 2nd-Cousin

…

Bob

Carol

Dan

Frank

…

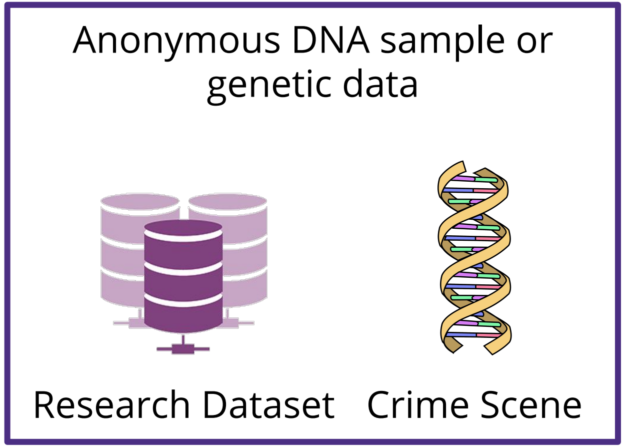
1M+

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



Anonymous DNA sample or genetic data

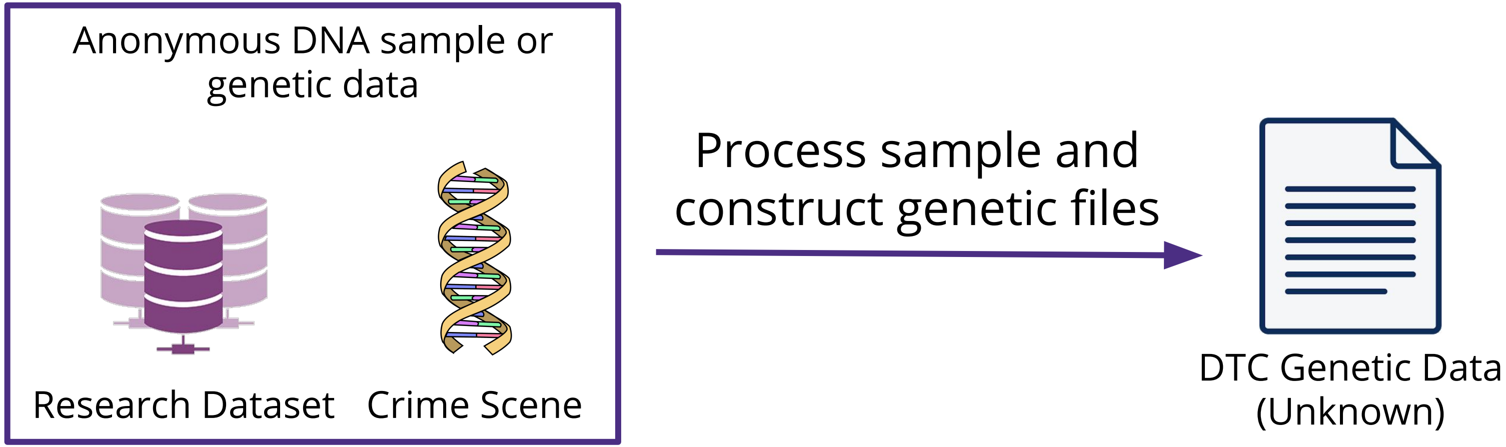
Research Dataset

Crime Scene

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

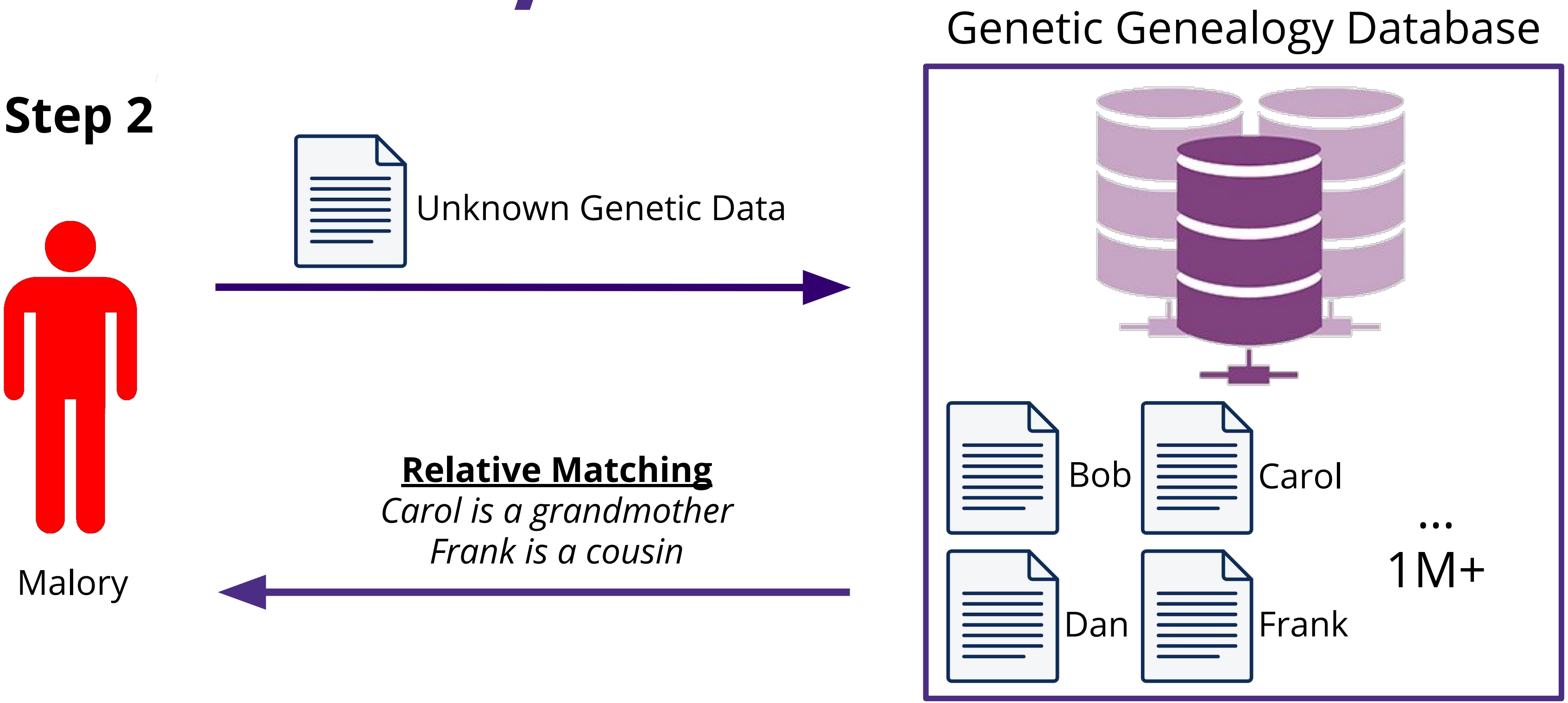
Research Dataset

Crime Scene

Process sample and construct genetic files

DTC Genetic Data (Unknown)

Prior Attacks Against Genetic Genealogy Services: Identity Inference



Step 2

Unknown Genetic Data

Relative Matching

Carol is a grandmother
Frank is a cousin

Malory

Genetic Genealogy Database

Bob

Carol

Dan

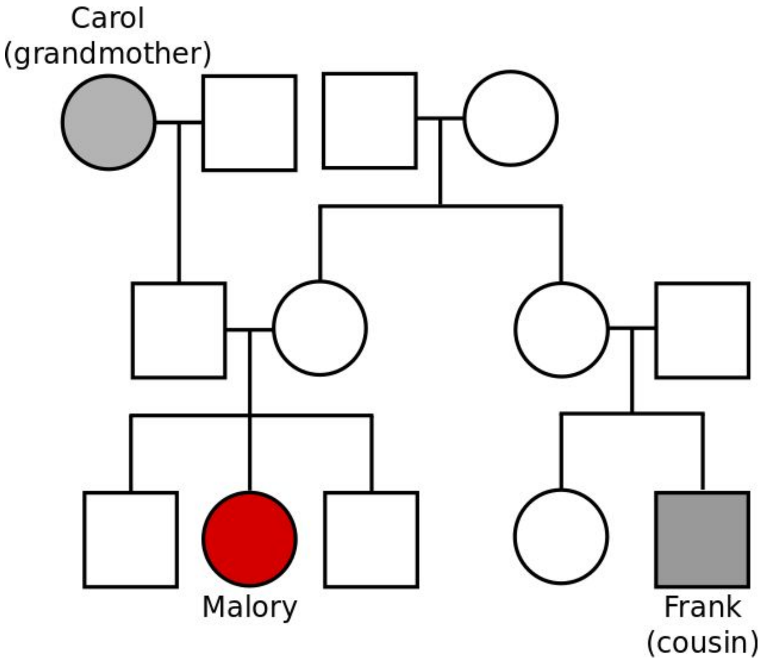
Frank

...

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



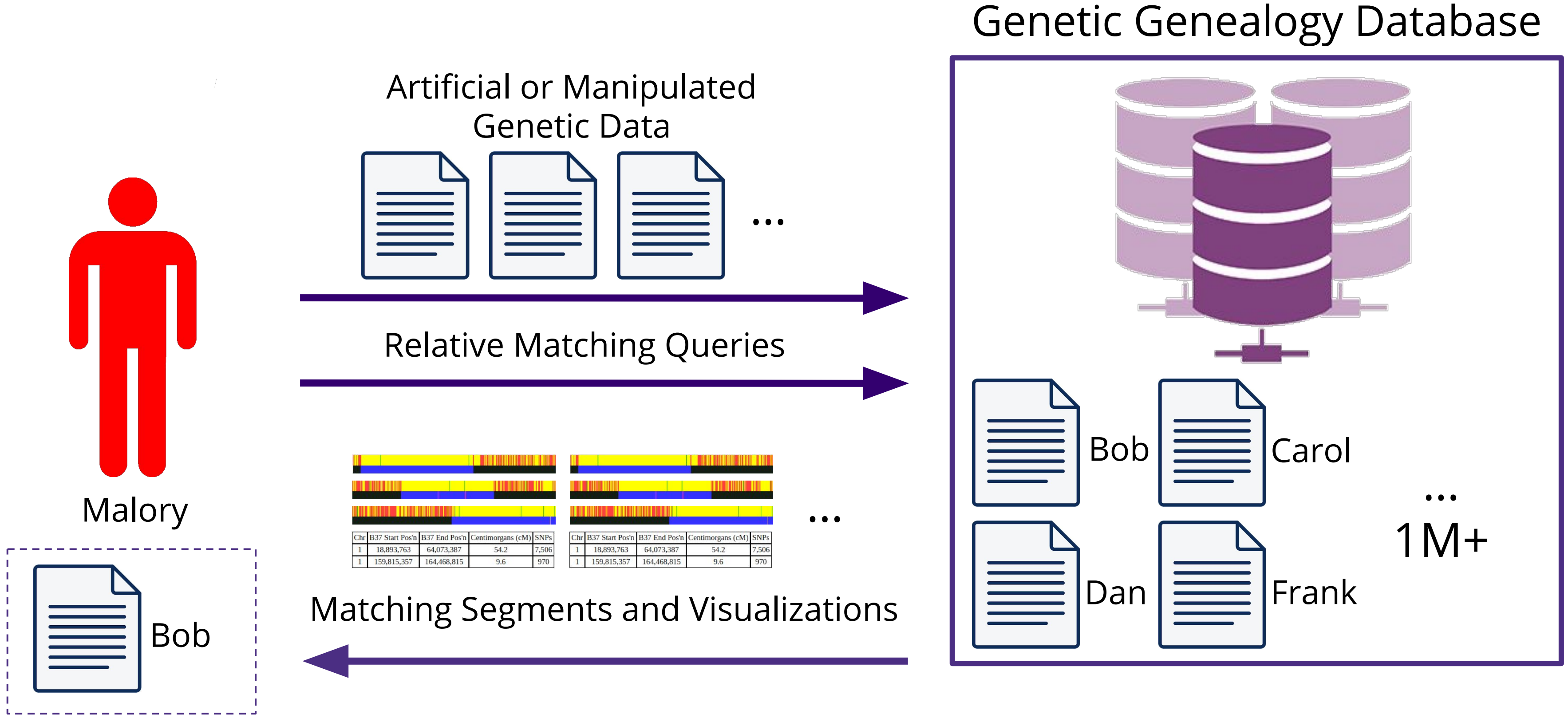
Law enforcement

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

Anonymous research data

- Ex: 1000 Genomes Data (*Erlich et al. Science. 2018)*

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



Genetic Genealogy Database

Artificial or Manipulated Genetic Data

...

Relative Matching Queries

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

...

Matching Segments and Visualizations

Malory

Bob

Bob

Carol

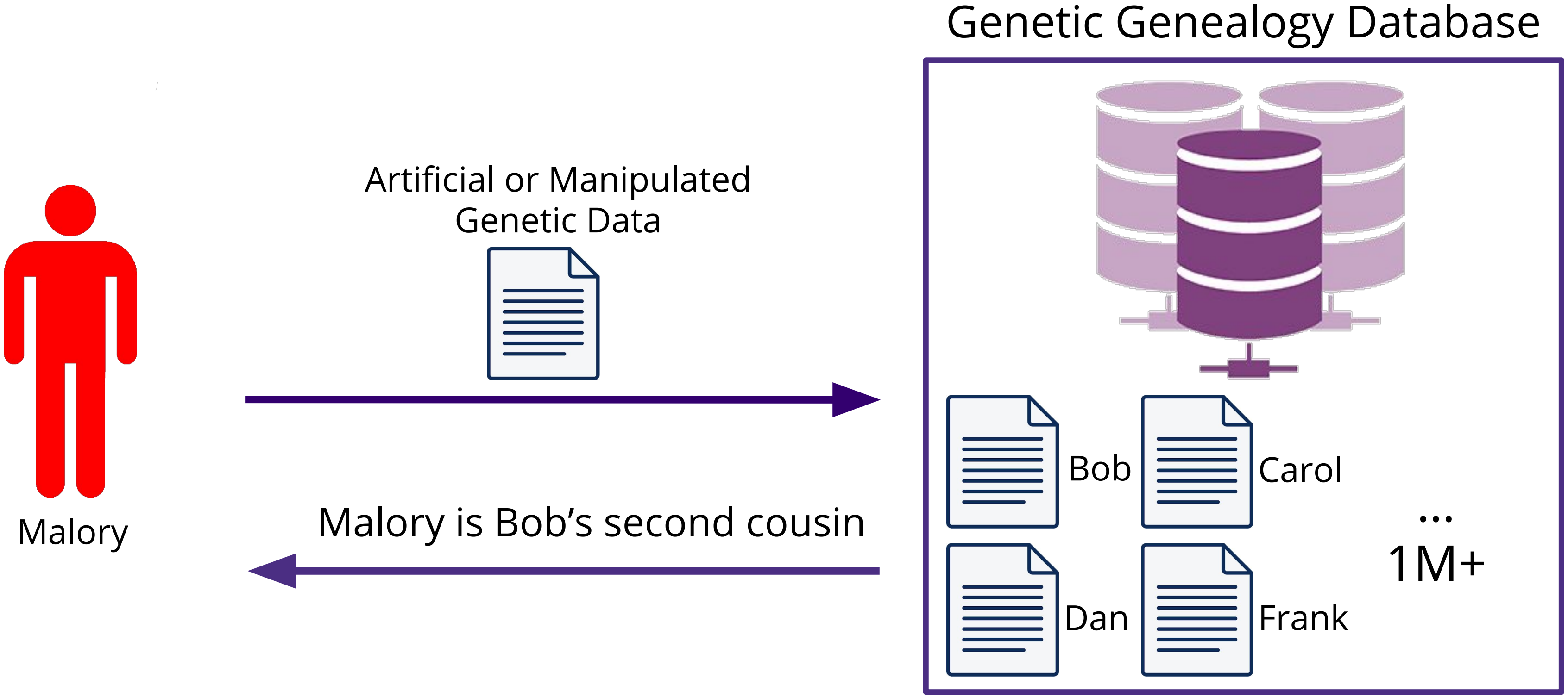
Dan

Frank

...

1M+

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

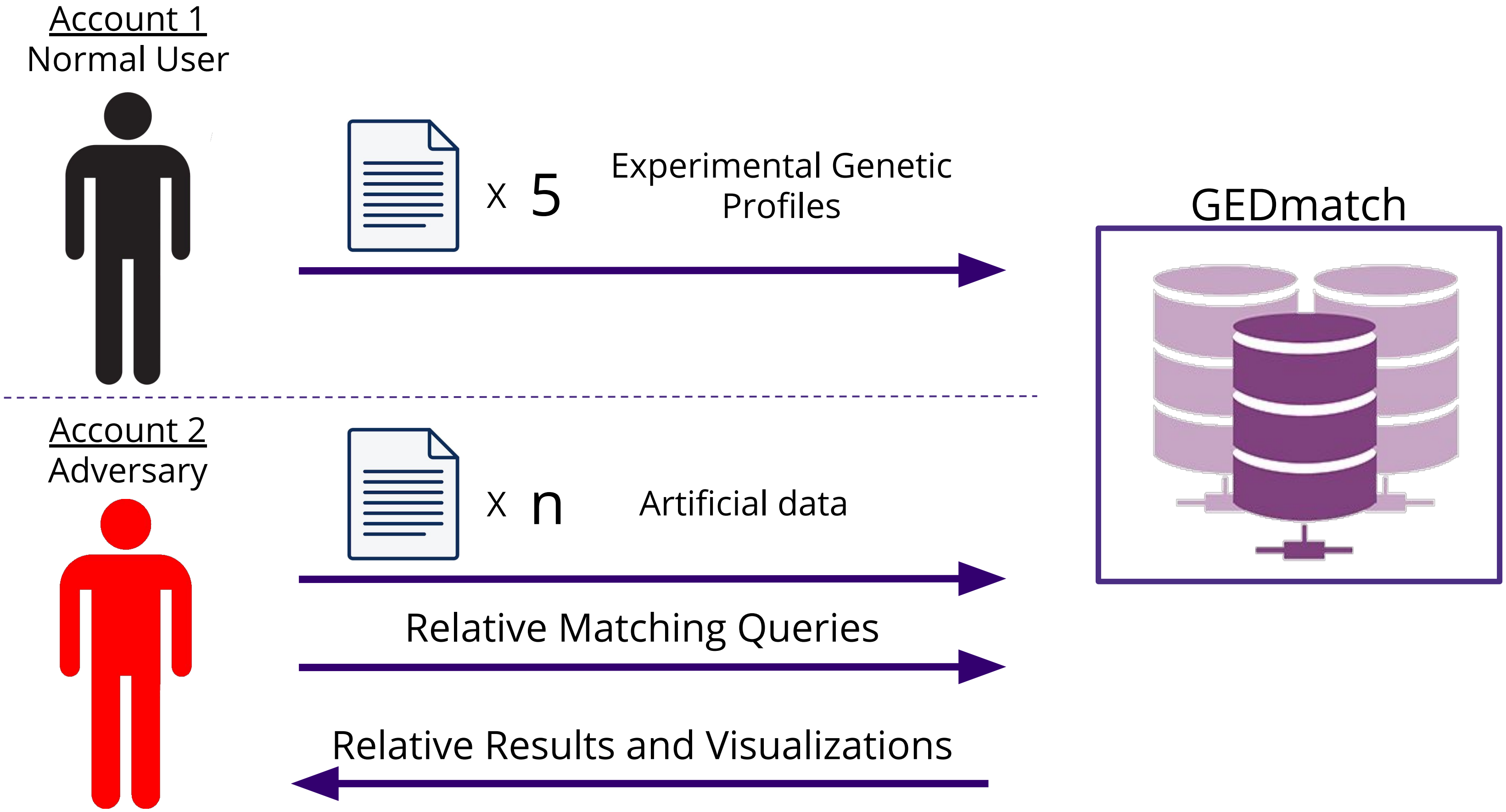

Genetic Genealogy Database
Artificial or Manipulated Genetic Data
Malory
Malory is Bob's second cousin
Bob
Carol
Dan
Frank
...
1M+

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

GED
match

Experimental Setup on GEDmatch



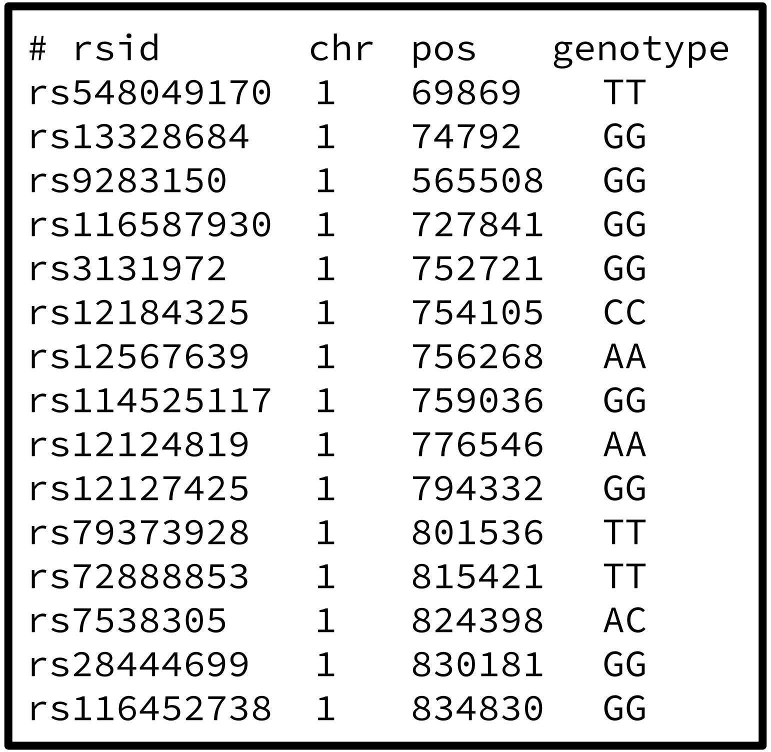
Account 1
Normal User
x 5
Experimental Genetic Profiles
GEDmatch
Account 2
Adversary
x n
Artificial data
Relative Matching Queries
Relative Results and Visualizations

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 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



```
# rsid       chr  pos    genotype
rs548049170  1    69869    TT
rs13328684   1    74792    GG
rs9283150    1    565508   GG
rs116587930  1    727841   GG
rs3131972    1    752721   GG
rs12184325   1    754105   CC
rs12567639   1    756268   AA
rs114525117  1    759036   GG
rs12124819   1    776546   AA
rs12127425   1    794332   GG
rs79373928   1    801536   TT
rs72888853   1    815421   TT
rs7538305    1    824398   AC
rs28444699   1    830181   GG
rs116452738  1    834830   GG
```

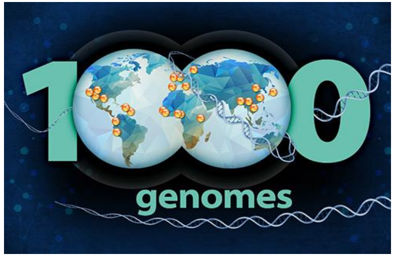
Genetic Data File (GDF)

Generating DTC Data Files for Experimentation

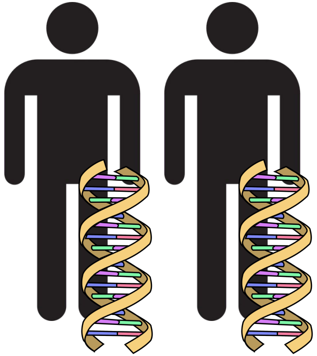
DTC Genetic Data Files
(**23andMe v5 SNP-chip**)

```
# rsid        chr  pos     genotype
rs548049170  1    69869     TT
rs13328684   1    74792     GG
rs9283150    1    565508    GG
rs116587930  1    727841    GG

                 ...
```



Whole genome sequence
& variant data



```
# rsid        chr  pos     genotype
rs548049170  1    69869     TT
rs13328684   1    74792     GG
rs9283150    1    565508    GG
rs116587930  1    727841    GG

                 ...
```

Generating DTC Data Files for Experimentation

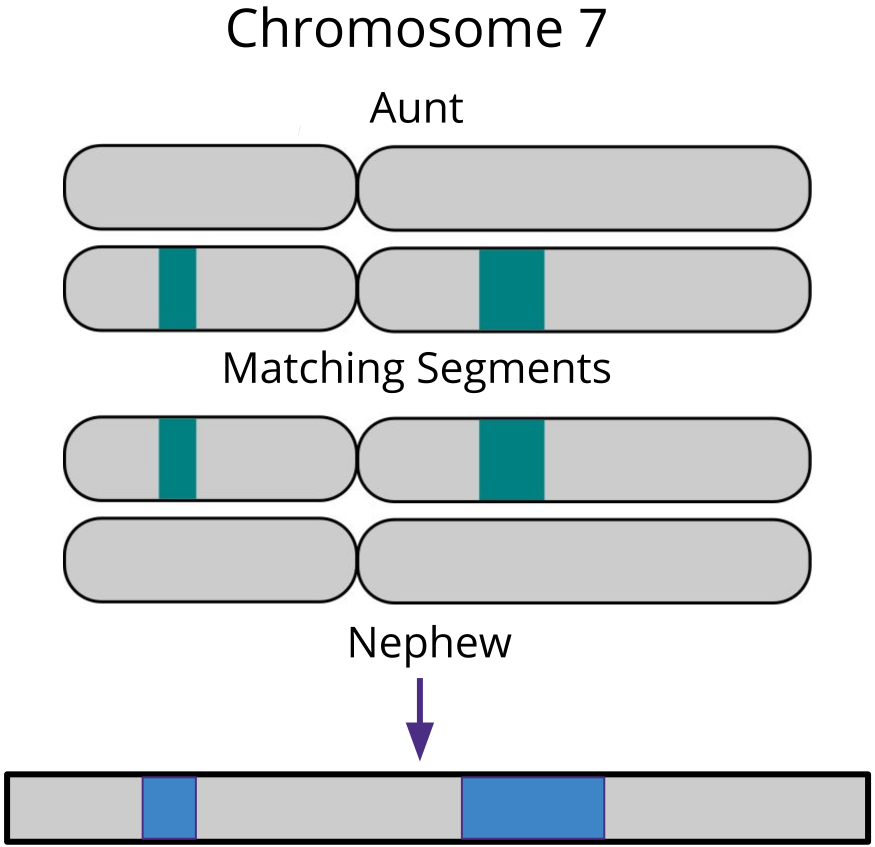
Programming Tools

- 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

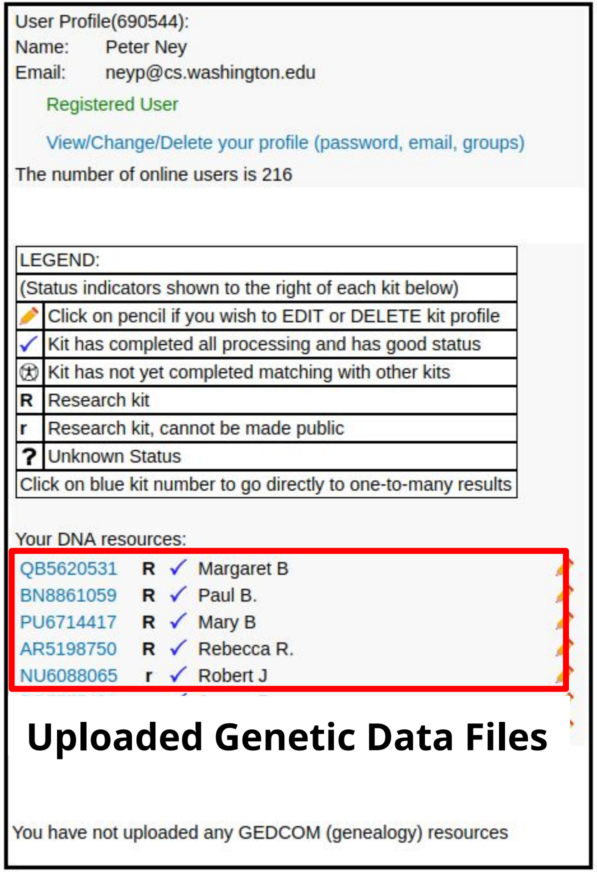
- Sample size for testing was small (5 target files) and all 23andMe files. Choose this to limit impact on the GEDmatch service.
- 1000 Genomes data came from same sub-population

Relative Matching on GEDmatch



- 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



User Profile(690544):
Name: Peter Ney
Email: neyp@cs.washington.edu

Registered User

View/Change/Delete your profile (password, email, groups)

The number of online users is 216

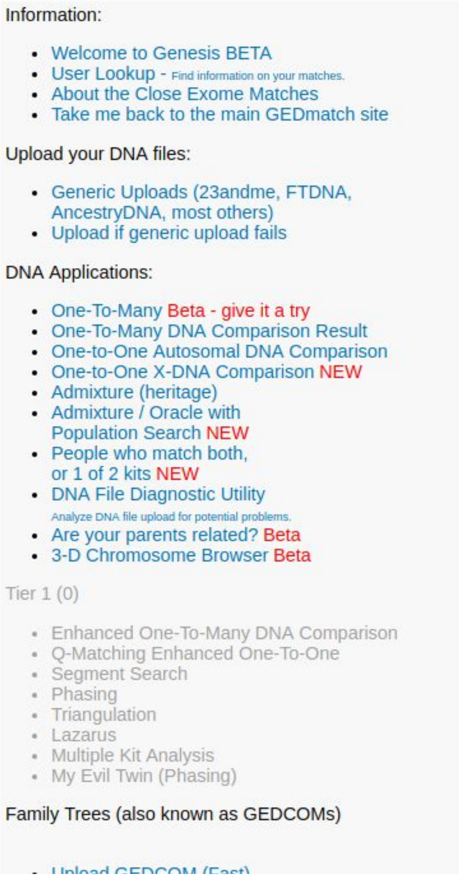
LEGEND:	
(Status indicators shown to the right of each kit below)	
✎	Click on pencil if you wish to EDIT or DELETE kit profile
✓	Kit has completed all processing and has good status
⊛	Kit has not yet completed matching with other kits
R	Research kit
r	Research kit, cannot be made public
?	Unknown Status
Click on blue kit number to go directly to one-to-many results	

Your DNA resources:

QB5620531	R	✓	Margaret B	✎
BN8861059	R	✓	Paul B.	✎
PU6714417	R	✓	Mary B	✎
AR5198750	R	✓	Rebecca R.	✎
NU6088065	r	✓	Robert J	✎

Uploaded Genetic Data Files

You have not uploaded any GEDCOM (genealogy) resources



Information:

- Welcome to Genesis BETA
- User Lookup - Find information on your matches.
- About the Close Exome Matches
- Take me back to the main GEDmatch site

Upload your DNA files:

- Generic Uploads (23andme, FTDNA, AncestryDNA, most others)
- Upload if generic upload fails

DNA Applications:

- One-To-Many Beta - give it a try
- One-To-Many DNA Comparison Result
- One-to-One Autosomal DNA Comparison
- One-to-One X-DNA Comparison NEW
- Admixture (heritage)
- Admixture / Oracle with Population Search NEW
- People who match both, or 1 of 2 kits NEW
- DNA File Diagnostic Utility
  Analyze DNA file upload for potential problems.
- Are your parents related? Beta
- 3-D Chromosome Browser Beta

Tier 1 (0)

- Enhanced One-To-Many DNA Comparison
- Q-Matching Enhanced One-To-One
- Segment Search
- Phasing
- Triangulation
- Lazarus
- Multiple Kit Analysis
- My Evil Twin (Phasing)

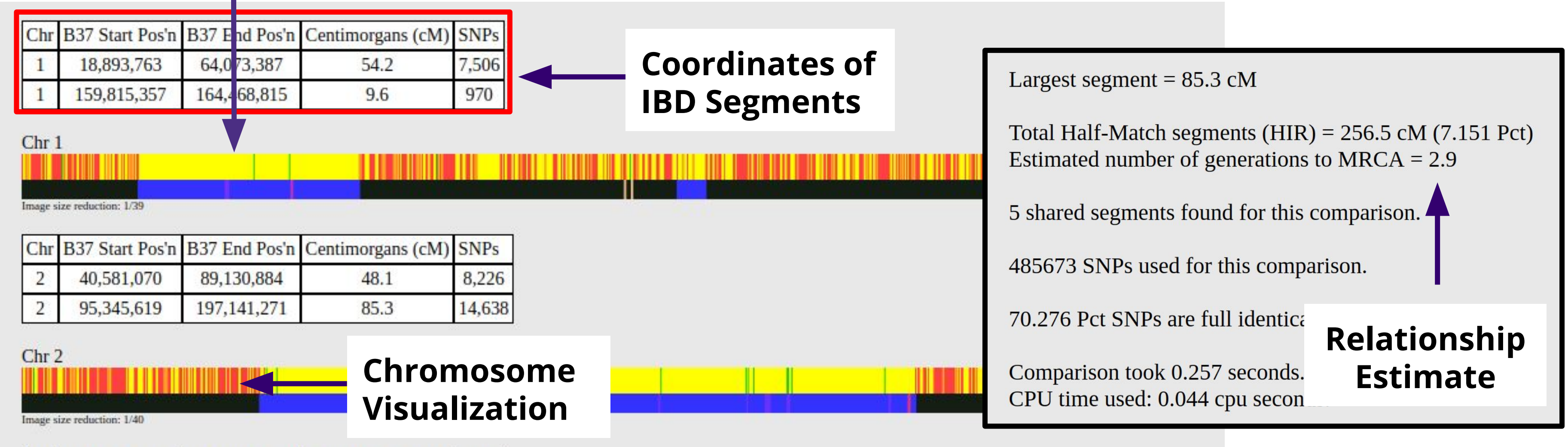
Family Trees (also known as GEDCOMs)

- Upload GEDCOM (Fast)

Relative Matching on GEDmatch

Easily scrape the query results and visualizations

Direct relative matching query between two users



Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Chr 1

Image size reduction: 1/39

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
2	40,581,070	89,130,884	48.1	8,226
2	95,345,619	197,141,271	85.3	14,638

Chr 2

Image size reduction: 1/40

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 identica

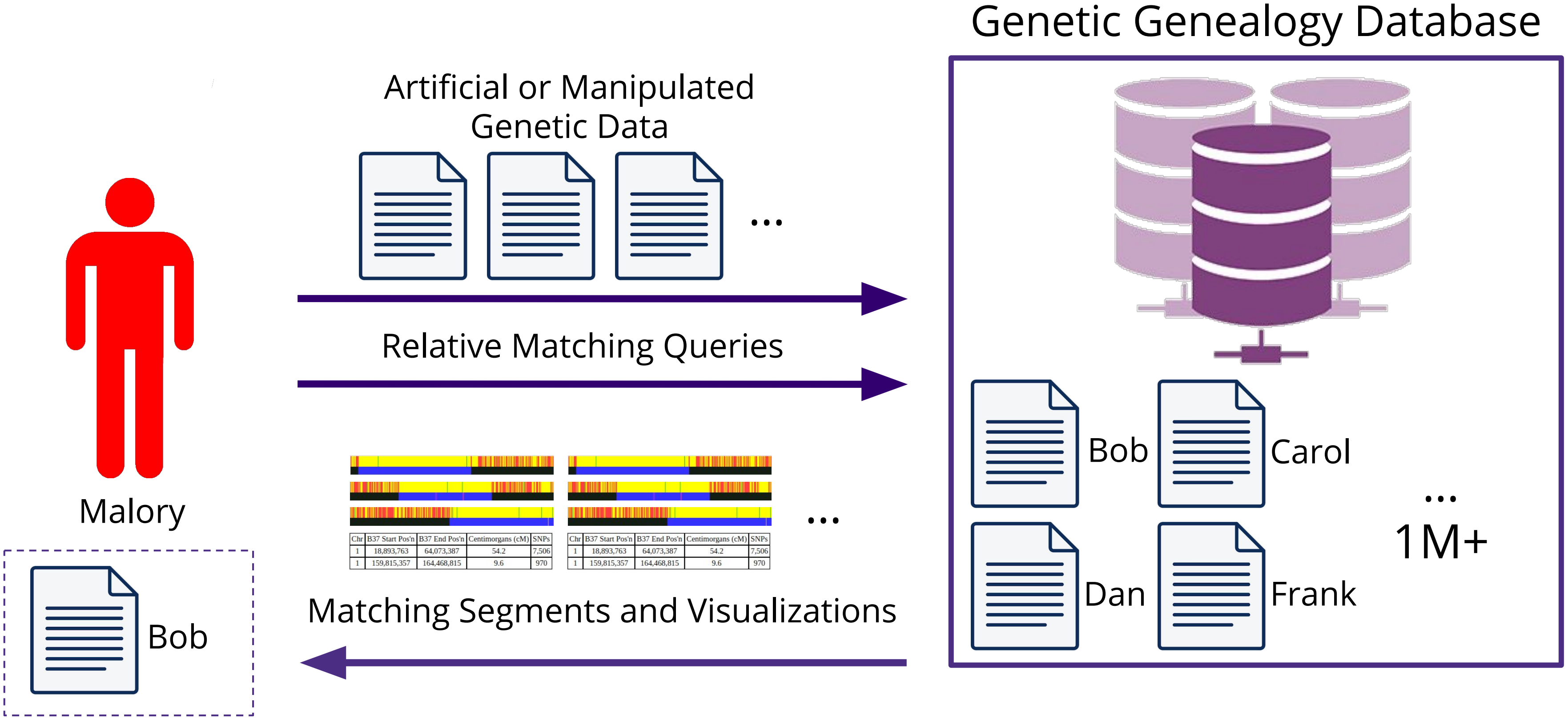
Comparison took 0.257 seconds.
CPU time used: 0.044 cpu secon

Coordinates of IBD Segments

Chromosome Visualization

Relationship Estimate

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



Genetic Genealogy Database

Artificial or Manipulated Genetic Data

...

Relative Matching Queries

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

...

Matching Segments and Visualizations

Malory

Bob

Bob

Carol

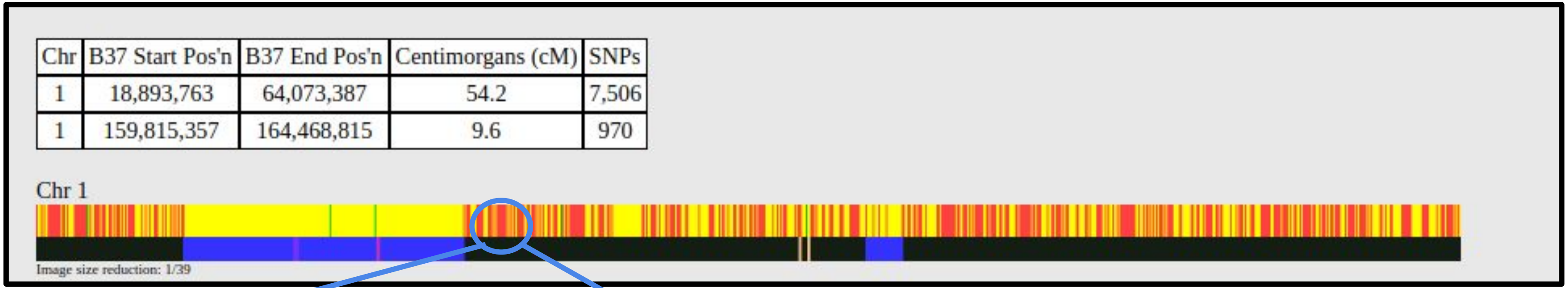
Dan

Frank

...

1M+

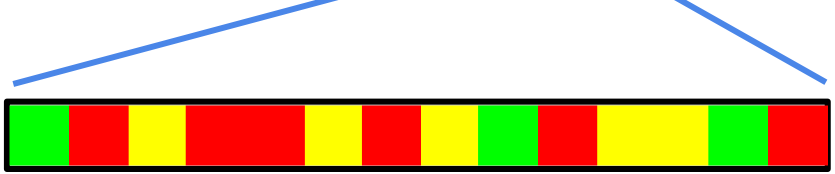
GEDmatch Visualizations and Segments

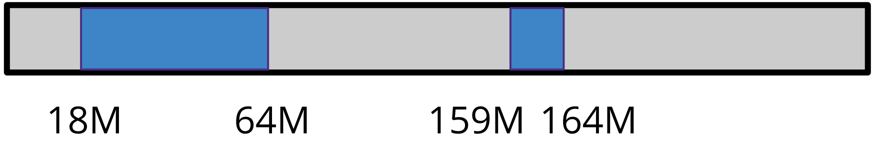


Chr	B37 Start Pos'n	B37 End Pos'n	Centimorgans (cM)	SNPs
1	18,893,763	64,073,387	54.2	7,506
1	159,815,357	164,468,815	9.6	970

Chr 1

Image size reduction: 1/39

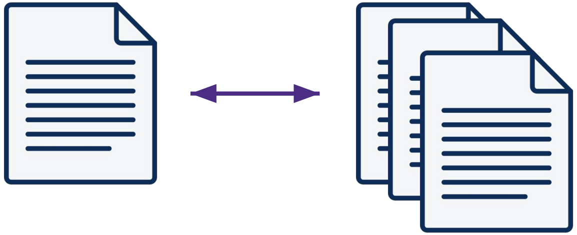




Both visualizations leak information about the underlying DNA markers in other genetic files.

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.

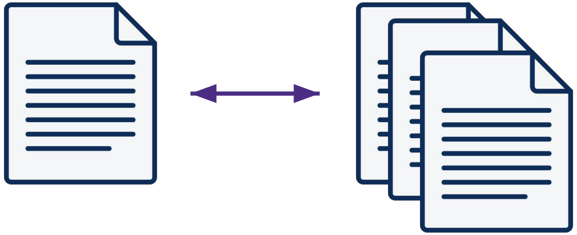


Regular file

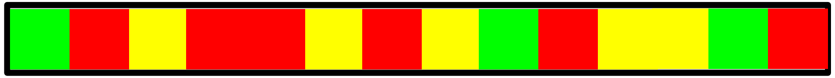
Modified data file

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.



Hypothesis



1) At high resolution these pixels seemed to correspond to individual markers
2) Many markers seemed to be missing

3) Results not phased

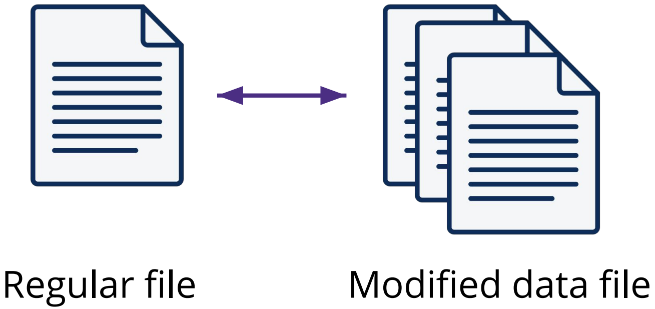
GT == TG 🟩

GG == TG 🟨

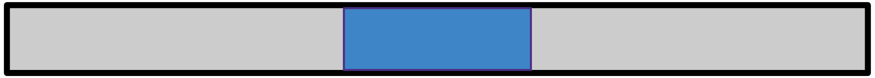
GG == TT 🟥

GEDmatch Visualizations and Segments

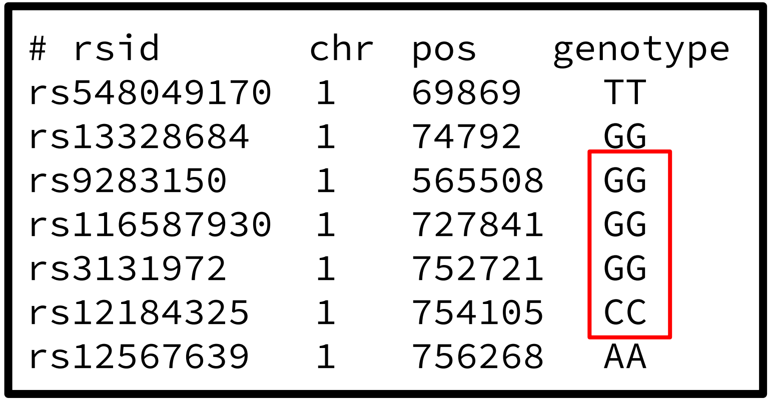
Matching algorithms and visualizations were proprietary so it was necessary to run a number of experiments to figure out how they were working.



Hypothesis



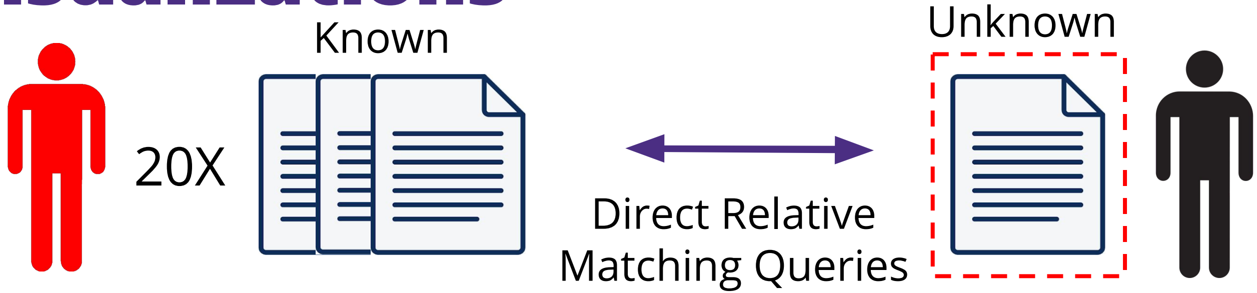
A section of chromosome is considered a shared segment if the files match on a single base for a run of consecutive markers



```
# rsid        chr  pos     genotype
rs548049170   1    69869     TT
rs13328684    1    74792     GG
rs9283150     1    565508    GG
rs116587930   1    727841    GG
rs3131972     1    752721    GG
rs12184325    1    754105    CC
rs12567639    1    756268    AA
```

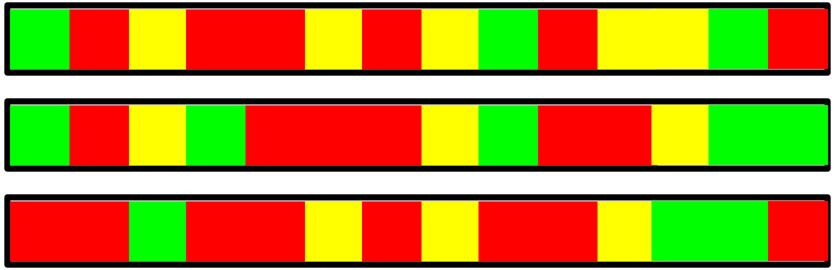
Genetic Extraction Experiments with Marker Visualizations

Ran attack 5 times (one for each experimental file)

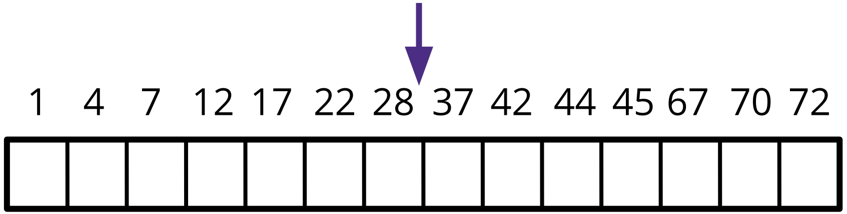


Direct Relative Matching Queries

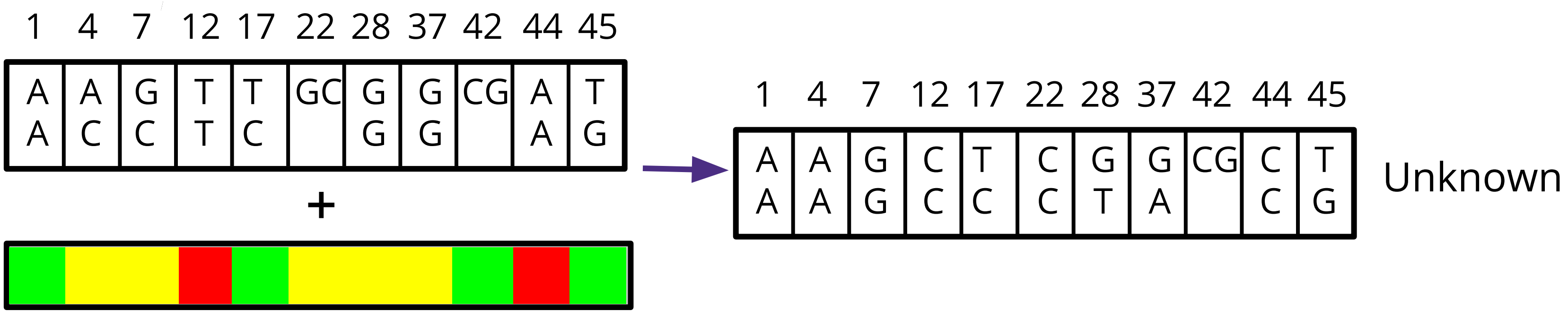
Collected visualizations from Chrome browser (20 comparisons x 22 autosomes = 440 per attack)



Process visualizations with python scripts implementing a mastermind-like algorithm to infer which markers went with which pixels



Genetic Extraction Experiments with Marker Visualizations



Known (from attacker file)

1	4	7	12	17	22	28	37	42	44	45
A	A	G	T	T	GC	G	G	CG	A	T
A	C	C	T	C		G	G		A	G

+

🟩	🟨	🟨	🟥	🟩	🟨	🟨	🟨	🟩	🟥	🟩

1	4	7	12	17	22	28	37	42	44	45
A	A	G	C	T	C	G	G	CG	C	T
A	A	G	C	C	C	T	A		C	G

Unknown

Fill in the gaps using a statistical technique called genetic imputation. Relied on a publicly available genetic imputation service run by the Sanger Institute.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	A	G	C	T	C	G	G	CG	C	T	T	A	T
A	A	G	C	C	C	T	A		C	G	G	C	T

Genetic Extraction Experiments with M

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.

own

Fill in t

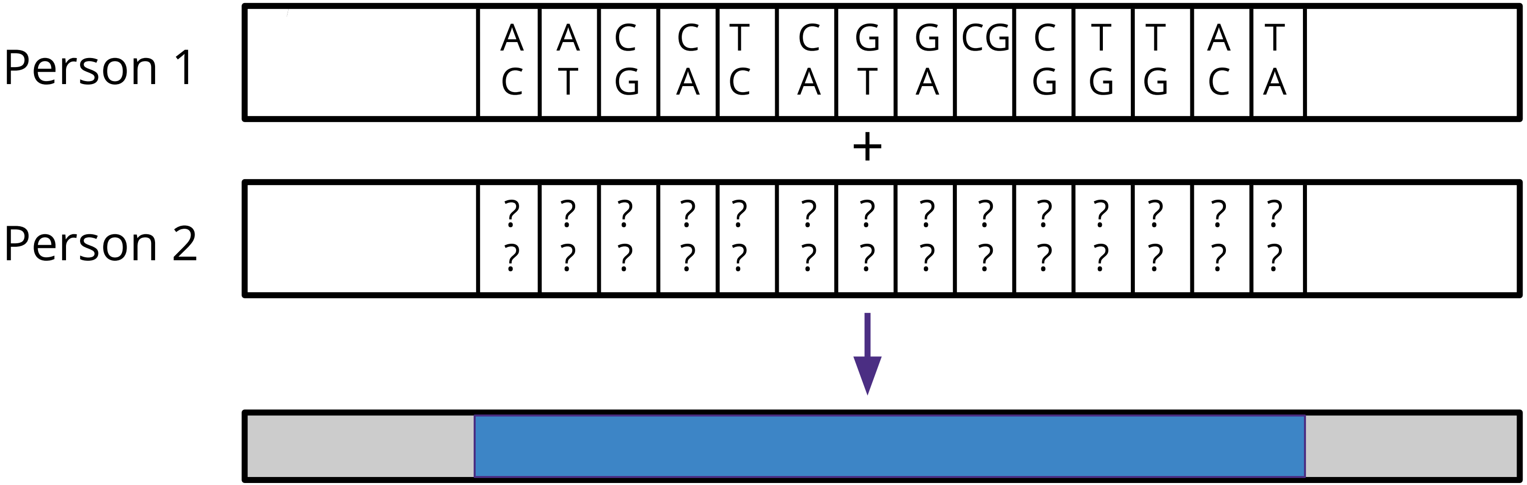
calle

publicly available genetic imputation

service run by the Sanger Institute.

A	A	G	C	C	C	T	A		C	G	G	C	T

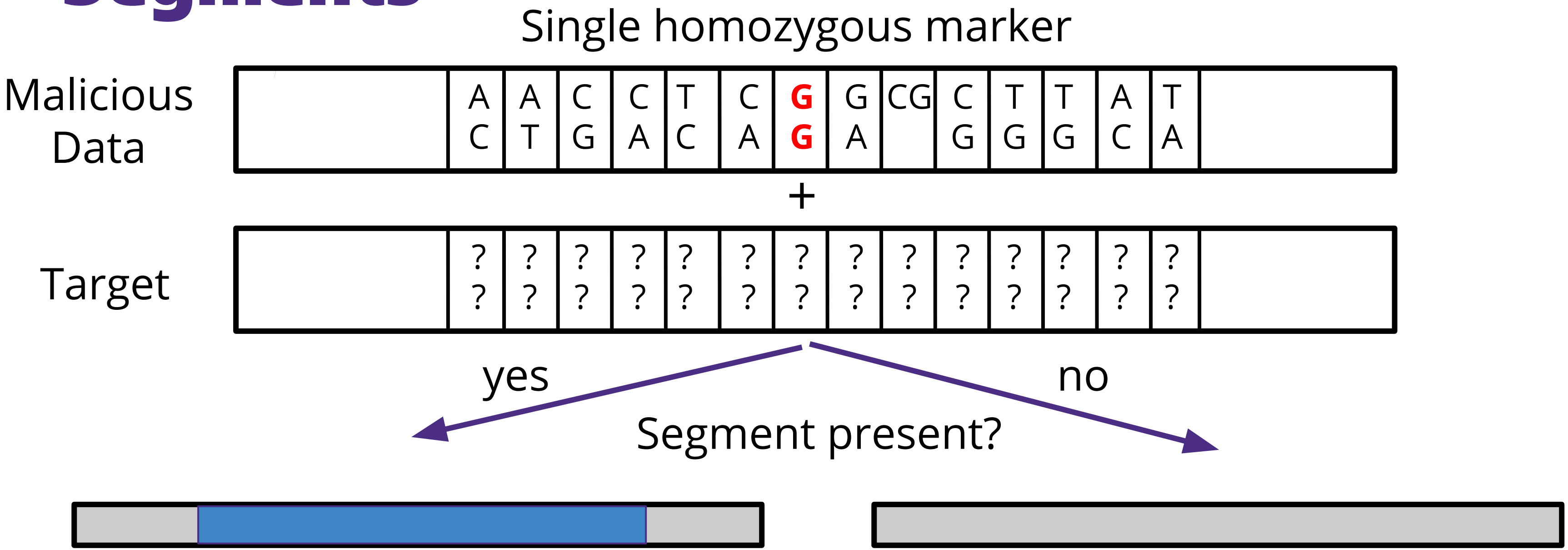
Genetic Extraction with Matching Segments



Person 1		A C	A T	C G	C A	T C	C A	G T	G A	CG	C G	T G	T G	A C	T A	
								+								
Person 2		? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	? ?	

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	A	C	C	T	C	**G**	G	CG	C	T	T	A	T
	C	T	G	A	C	A	**G**	A		G	G	G	C	A

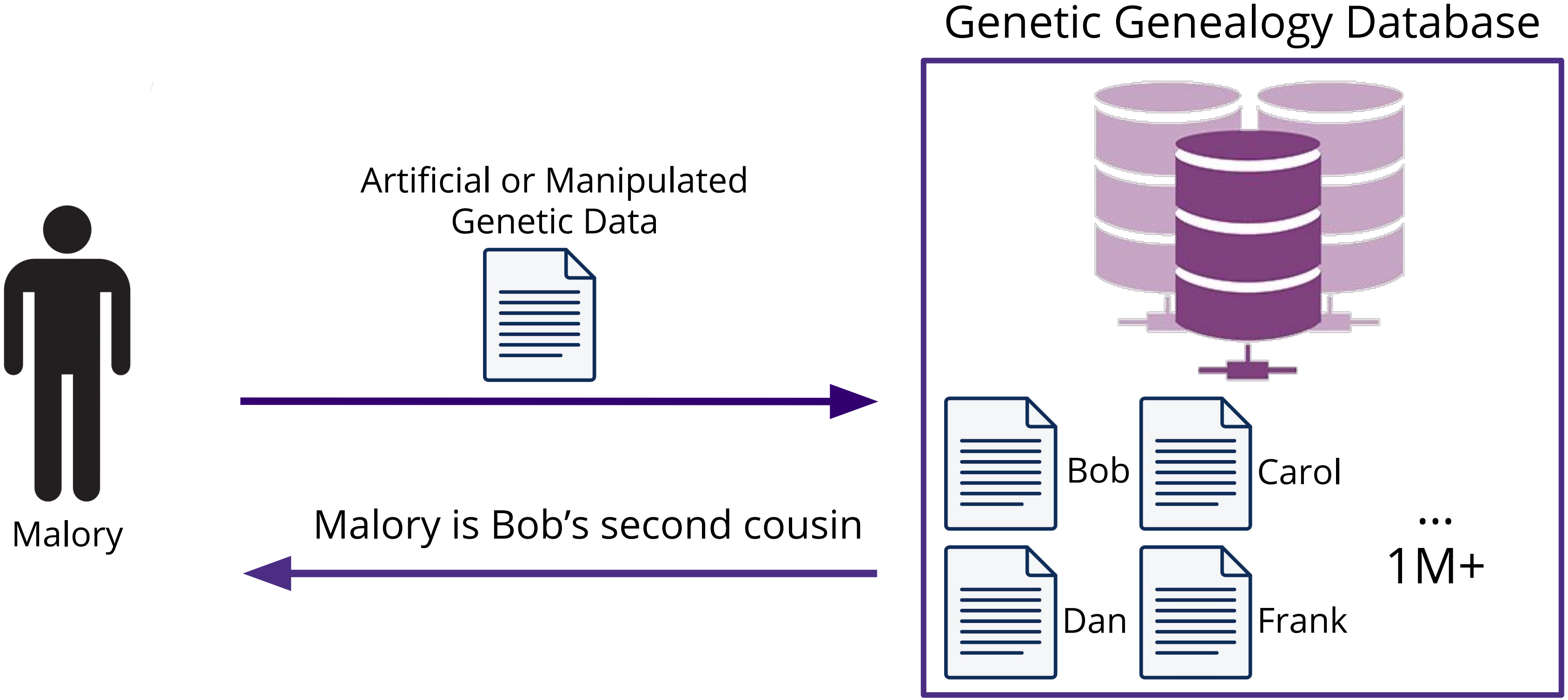
+

Target	?	?	?	?	?	?	?	?	?	?	?	?	?	?
	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Segment present?

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.*

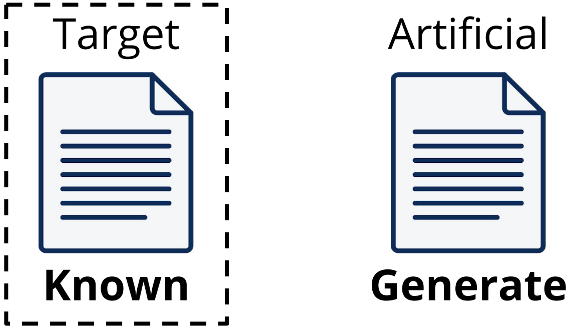
Hypothesis #2: Can We Generate Artificial Relatives for Other Users in a GG Database?


Genetic Genealogy Database
Artificial or Manipulated Genetic Data
Malory
Malory is Bob's second cousin
Bob
Carol
Dan
Frank
...
1M+

Experimenting with Artificial Relatives

Amount of DNA sharing determines the relative prediction

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



Target

Known

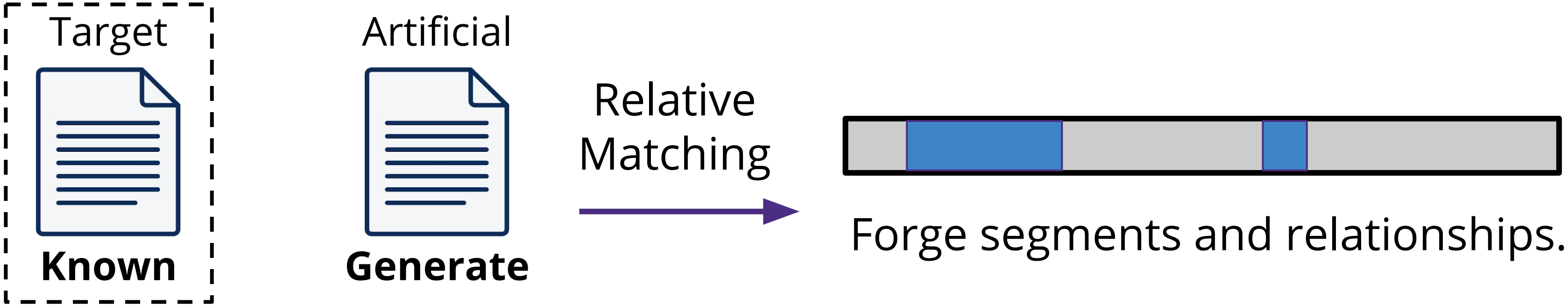
Artificial

Generate

Experimenting with Artificial Relatives

Amount of DNA sharing determines the relative prediction

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



Target

Known

Artificial

Generate

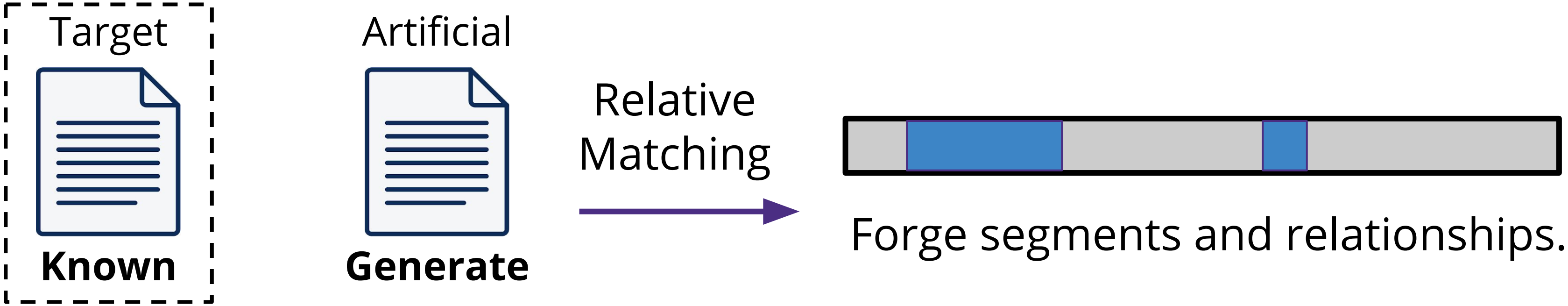
Relative Matching

Forge segments and relationships.

Experimenting with Artificial Relatives

Amount of DNA sharing determines the relative prediction

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



Target

Known

Artificial

Generate

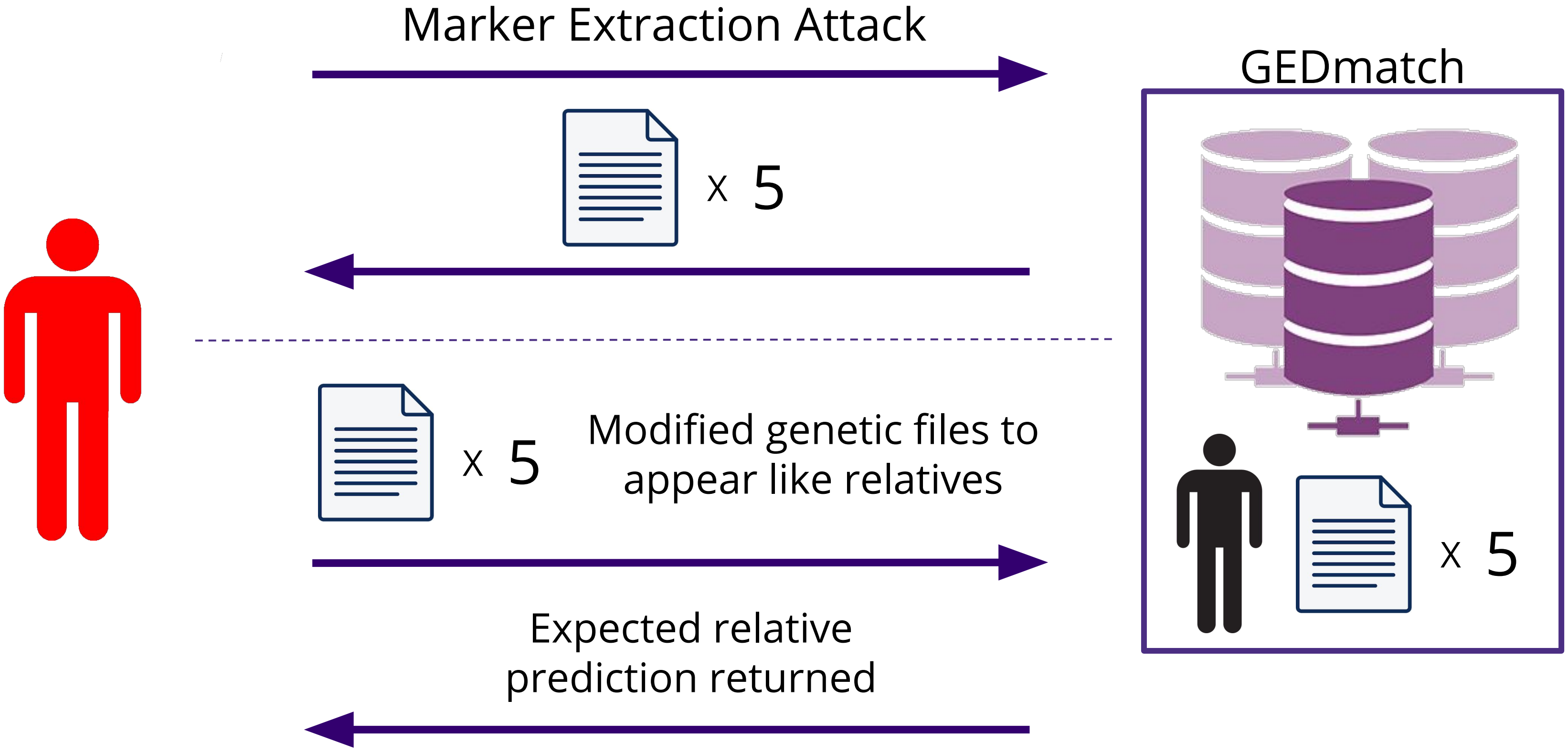
Relative Matching

Forge segments and relationships.

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

x 5

x 5

Modified genetic files to appear like relatives

x 5

Expected relative prediction returned

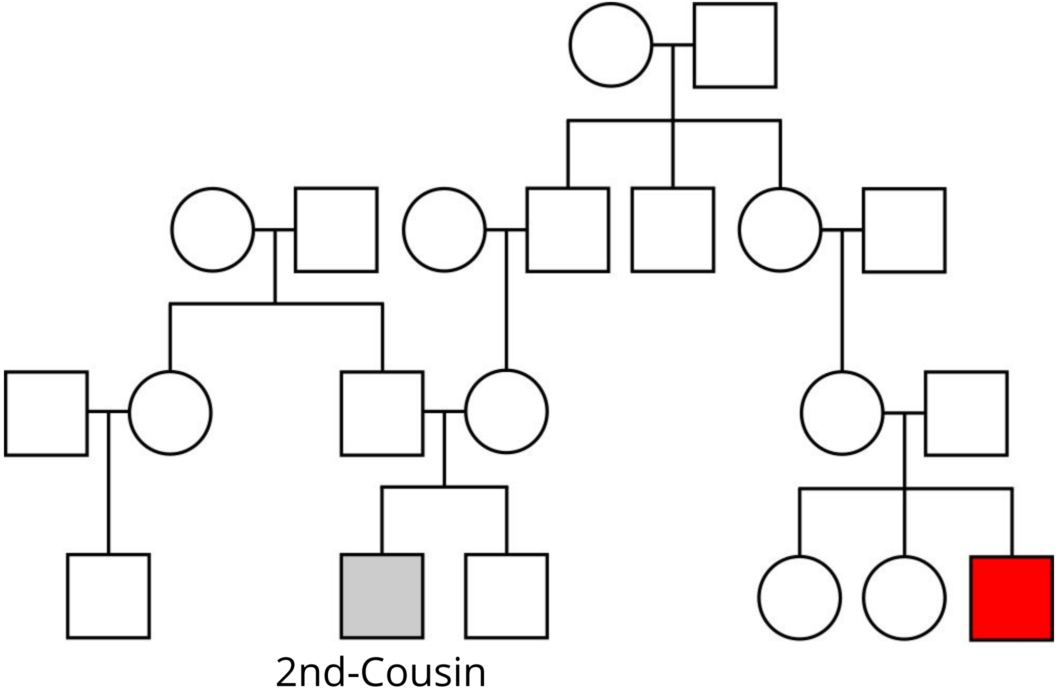
Experimentation Artifacts Borrowed from the Community?

- 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

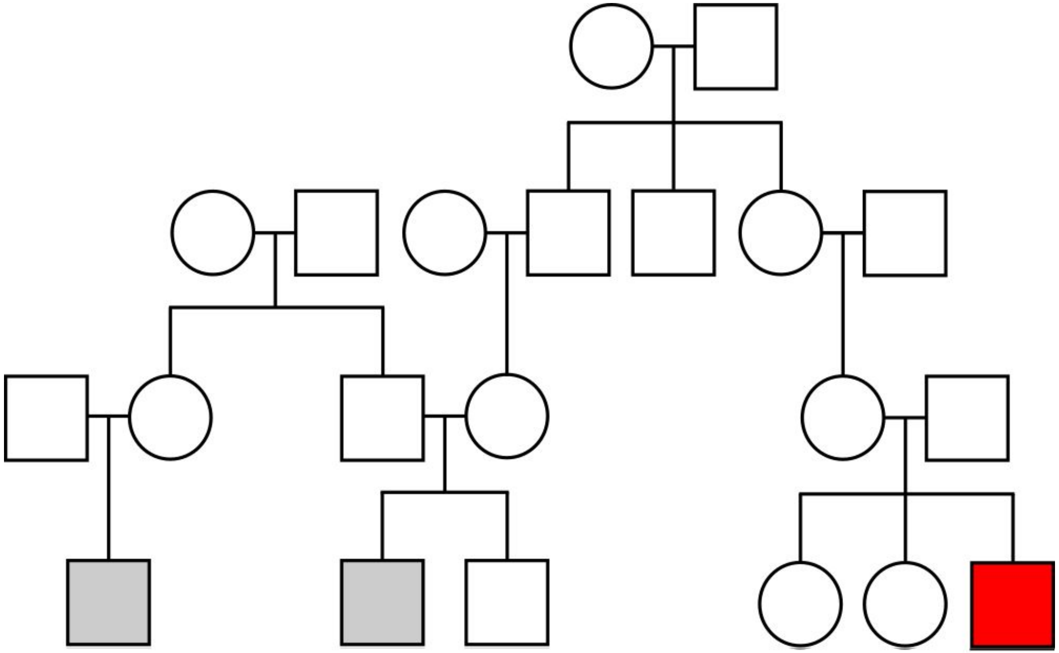
Reproducing Results?

- 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

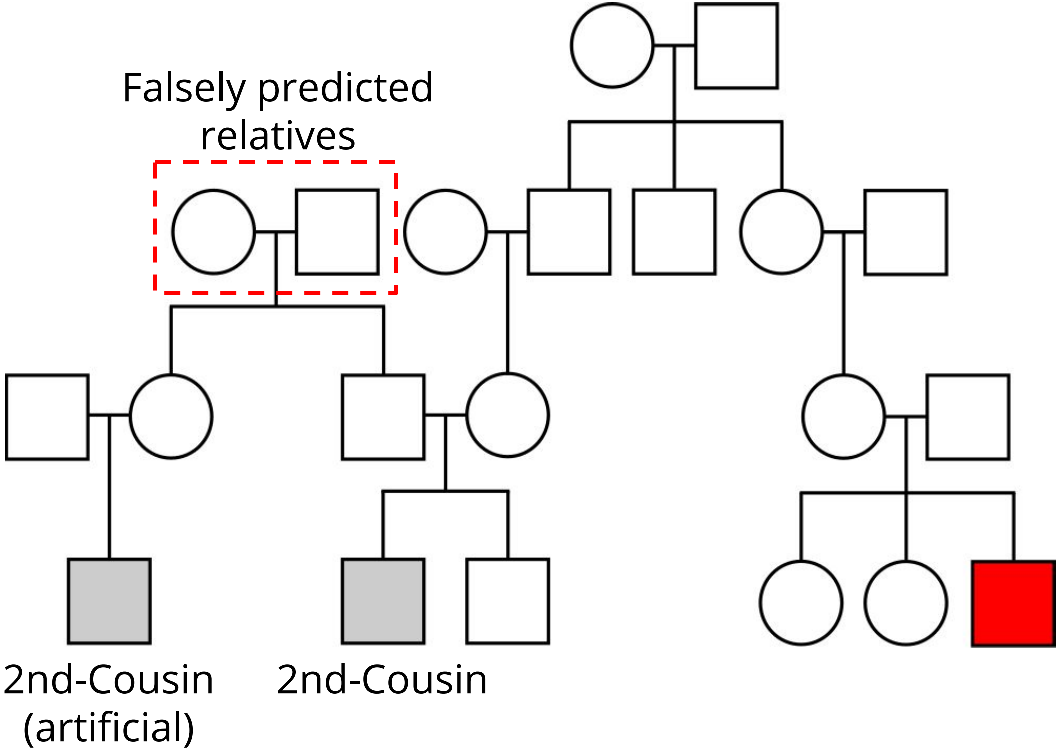


2nd-Cousin (artificial)

2nd-Cousin

Failed / Unsuccessful Experiment: Disrupting Identity Inference

Search occurs on wrong branch of tree



Failed / Unsuccessful Experiment: Disrupting Identity Inference

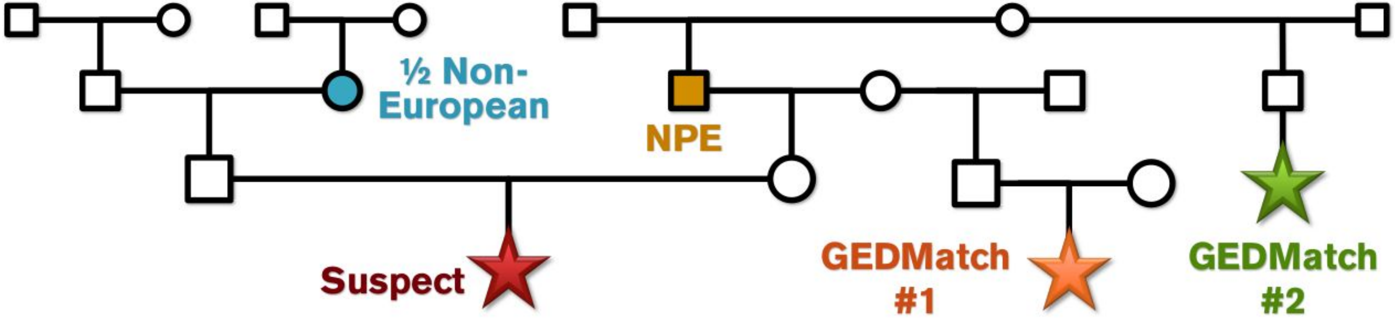
- How do you run experiments that take genealogies / family trees into account?
- Family tree data is available
  - 1M+ person trees meant for research
- Tried to run simulations to see how easily a random individual could be mis-identified
  - Depends on tree topology and number of relatives in the 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