INSTITUTE FOR DEFENSE ANALYSES

Probability of Identifying a Target from Human Genetic Datasets

Ashley Farris
Robert Cubeta

June 2024

IDA Product 3000645

INSTITUTE FOR DEFENSE ANALYSES
730 E. Glebe Rd
Alexandria, VA 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses under contract HQ0034-19-D-0001, project AI-6-5283, "Biology Data Risk Assessment Methodology" for the Director, Science & Technology Exploitation and Analytics, Maintaining Technology Advantage (MTA), Office of the Under Secretary of Defense, Research & Engineering. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The authors wish to thank Jeff Grotte, Catriona Mitchell, and Carly Cox for their thoughtful reviews, and the publications team Florestine Purnell and Amberlee Mabe-Stanberry for their efforts in polishing the final product.

For More Information:
Mr. Robert L. Cubeta Project Leader
rcubeta@ida.org, 703-575-4681
Ms. Jessica L. Stewart, Director, SFRD
jstewart@ida.org, 703-575-4530

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Product 3000645

Probability of Identifying a Target from Human Genetic Datasets

Ashley Farris
Robert Cubeta

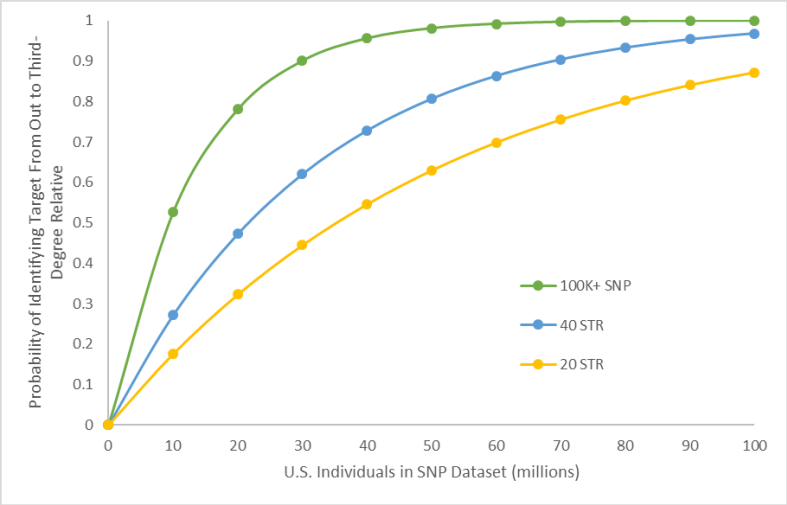
This page is intentionally blank.

Executive Summary

This paper describes a case study of potential privacy and national security risks associated with the aggregation and sale of human genetic data by characterizing how U.S. genetic data can be used by strategic competitors to identify individuals. The Director, Science & Technology Exploitation and Analytics, Maintaining Technology Advantage (MTA), Office of the Under Secretary of Defense, Research & Engineering requested that the Institute for Defense Analyses (IDA) develop this paper.

Since 2018, genetic datasets such as short tandem repeat (STR) and single nucleotide polymorphism (SNP) have been "matched" using a phenomenon called linkage disequilibrium[1] to identify individuals as suspects for violent crimes. In the figure below, we calculated the probability of identifying a target at several specific SNP dataset sizes, from 100 SNP profiles to 6.6 million SNP profiles. To contextualize the process of identifying an individual from matched genetic data, the results model an estimate to the following question: What is the probability of identifying a person by matching a collected genetic sample with a SNP record of either themselves or a family member as a function of SNP dataset size?

[1] Linkage disequilibrium is a measure of the likelihood that two genes are inherited together. This phenomenon allows researchers to predict the probability an individual might possess genotype A, if they already know that individual possesses genotype B. Given two disparate datasets such as SNP and STR results, linkage disequilibrium can enable record matching even across datasets that test for different marker types.



Probability of Identifying a Target from Their Genetic Data or that of a Relative Out to the Third-Degree Using Various Genetic Tests as a Function of SNP Dataset Size

The above overall probabilities consist of three separate events: 1) the target individual or at least one of their relatives is in the SNP dataset record; 2) the collected and analyzed genetic sample can be matched with the SNP record of the target or at least one of their relatives (first-, second-, or third-degree relatives such as parents, siblings, grandparents, aunts/uncles, grand aunts/uncles, or first cousins); and 3) genealogical investigation can be used to generate a family tree that enables the identification of the target.

While the matched STR and SNP data may have beneficial applications for law enforcement, use of genetic datasets to identify individuals for other purposes (such as genetic surveillance) raises both privacy and national security concerns. This study only focused on the use of genetic data for estimating reidentification probabilities. Additional research would be required to assess other types of data that may be used for reidentification alone or in conjunction with genetic data, including health records, social media profiles, mobility data, and multiomic data. This analysis contributes to the broader discussion about potential policy decisions relating to the privacy and national security concerns associated with the acquisition of U.S. persons sensitive personal data, including genomic data, by strategic competitors.

Table of Contents

This page is intentionally blank.

1. Introduction

Since 2018, genetic datasets (e.g., the FBI's CODIS[2] database and direct-to-consumer genetic testing databases) have been matched using probabilistic methods such as linkage disequilibrium for the purposes of identifying individuals as suspects and victims involved in violent crimes.[3] While this technique may have beneficial applications for law enforcement, use of genetic datasets to identify individuals for other purposes raises both privacy and national security concerns.

IDA supports the Director, Science & Technology Exploitation and Analytics, Maintaining Technology Advantage (MTA), Office of the Under Secretary of Defense, Research & Engineering. The MTA directorate's mission is to maintain a Department of Defense (DOD) technology advantage by balancing protection efforts with technology advancements to maintain leadership and technology superiority of critical and emerging technologies throughout the technology development lifecycle. MTA collaborates closely with the National Security Innovation Base (NSIB)—to include the Military Services, other DOD offices, the U.S. defense industry, and the U.S. academic/research enterprise—to identify and implement best practices, policies, mechanisms, strategies, and standards that protect U.S. technological advantage, foster U.S. technological development, and mitigate exploitation by strategic competitors. In 2022, MTA asked IDA to:

1. Develop a repeatable methodology to assess the national security risk posed by strategic competitor acquisition of U.S. biological datasets either alone or when combined with other data, and
2. Apply the methodology to representative case studies illustrating both the threat and risk of strategic competitor acquisition of U.S. biological data to facilitate messaging across the DOD and broader National Security audiences.

IDA's methodology assesses risk as the product of 1) the likelihood of a strategic competitor successfully achieving a user-specified application of a given dataset, and 2) the resulting consequence to a user-specified operation of interest. This operation of

---

[2] CODIS refers to the Combined DNA Index System, a computer software program that operates the U.S. national, state, and local databases of DNA profiles from convicted offenders, unsolved crime scene evidence, and missing persons.

[3] Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi, "Identity inference of genomic data using long-range familial searches," *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

interest need not be a specific military operation. Consequence to other national security activities such as intelligence activities or economic competitiveness are also considered. The methodology and results of this analysis are described in IDA papers P-33619[4] and P-33456[5].

However, the methodology used in these previous papers requires specification of a dataset and its characteristics (e.g., dataset size, types of data, population that data is derived from). There remained a need to understand how human genetic dataset characteristics (e.g., dataset size and type of genetic data) could more broadly impact the consequence resulting from dataset acquisition. MTA asked IDA to analyze how U.S. genetic data can be used to identify individuals. This analysis can inform potential policies, enforcement actions, and rulemaking concerning the privacy and national security concerns associated with the acquisition of U.S. persons sensitive personal data, including genomic data, by strategic competitors. A detailed methodology section is also included to explain the approach, assumptions, and limitations used to generate the results.

---

4 Robert Cubeta et al, *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data and Application to an Agricultural Bioprocessing Case Study*, IDA Paper P-33619 (Alexandria, VA: Institute for Defense Analyses, August 2023).

5 Robert Cubeta et al, *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*, IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, July 2023), TOP SECRET//NO FORN//SI.

2. Identifying Individuals from Acquired Genetic Data

In February 2024, President Biden signed the Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern.[6] This executive order tasks the Attorney General with coordination and consultation with other agencies (including the Department of Defense) to issue regulations that prohibit or restrict transactions of bulk sensitive data. One type of personal data specifically called out in the executive order is human genomic data. There has been considerable interest from the sponsor and stakeholder community on the privacy and national security implications of the acquisition of human genetic data from the U.S. population. This chapter provides background into the topic, a brief discussion of privacy and national security implications of genetic data acquisition, an example of how genetic data could be used for surveillance purposes, a methodology for estimating the probability that an individual can be identified based on the size of an acquired genetic dataset, and the limitations and assumptions used to generate this estimate.

A. Background

There are several types of genotyping assays available. Two common genotyping assays are short tandem repeat[7] (STR) tests and single nucleotide polymorphism[8] (SNP) tests. STR assays are commonly used for forensic identification purposes (such as in the CODIS database), whereas SNP assays are often sold as direct-to-consumer (DTC) kits to the public for the purposes of ancestry, health, and entertainment. As of 2019, it was estimated that over 26 million individuals worldwide had taken a DTC ancestry test.[9] One

---

6 "Executive Order 14117 of February 28, 2024, Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 15421-15430, https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04573.pdf.

7 STRs (also known as microsatellites) are segments of the DNA in which certain patterns of DNA are repeated, usually 5-50 times.

8 SNPs are genomic variants consisting of a substitution of single nucleotide at a specific position in the genome.

9 Antonio Regalado, "More than 26 million people have taken an at-home genetic ancestry test," *MIT Technology Review,* February 11, 2019, accessed August 2, 2023, https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

of the largest DTC companies, 23andMe, has sold over 10 million DNA test kits, mostly to individuals in the United States.[10] Additionally, many of these DTC companies store genetic data and some even resell consumer genetic data to third parties.[11]

B. Privacy and National Security Risks of Genetic Data Acquisition

Human genetic data has several characteristics that make it particularly valuable, including: 1) genetic data can be predictive of health conditions, 2) genetic data for an individual is immutable and cannot readily be changed, and 3) genetic data is shared between relatives. Genetic datasets, particularly when they consist of data from a large number of individuals, can result in a variety of privacy, national security, and economic risks when acquired by other nations.

Several types of privacy concerns can arise from the collection of genetic data. Discrimination on the basis of genetic data is only prohibited under the Genetic Information Nondiscrimination Act of 2008 by health insurance companies and employers, leaving loopholes for insurance companies, such as disability, life, or long-term care insurance, or other organizations. Additionally, malicious actors can combine personally identifiable information with genetic data for use in surveillance, coercion, or manipulation, representing a potential national security concern.[12] According to a 2019 DOD memorandum advising military service members to avoid DTC genetic testing, the scientific community has expressed increased concern that genetic data can be used by parties for "questionable purposes, including mass surveillance and the ability to track individuals without their authorization or awareness".[13]

In addition to privacy concerns, a study commissioned by the American Society of Human Genetics estimated the human genetics and genomics sector of the U.S. economy had an economic impact of $265 billion in 2019.[14] Genetic data has economic value, particularly for pharmaceutical companies seeking to develop new medical treatments. For

---

[10] Rani Molla, "Why DNA tests are suddenly unpopular," *Vox*, February 13, 2020, accessed August 2, 2023, https://www.vox.com/recode/2020/2/13/21129177/consumer-dna-tests-23andme-ancestry-sales-decline.

[11] Scott Thiebes et al, "Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing," *J Med Internet Res* 22, no. 1 (Jan 21 2020), https://doi.org/10.2196/14890, https://www.ncbi.nlm.nih.gov/pubmed/31961329.

[12] National Counterintelligence and Security Center, *Safeguarding Our Future: Protecting Personal Health Data from Foreign Exploitation* (Washington, D.C., 2021): 1.

[13] U.S. Department of Defense, Office of the Secretary of Defense, *Direct-to-Consumer Genetic Testing Advisory for Military Members* (Washington, D.C., 2019): 1.

[14] Simon Tripp and Martin Grueber, "The Economic Impact and Functional Applications of Human Genetics and Genomics," TEConomy Partners, LLC, May 2021, https://www.ashg.org/wp-content/uploads/2021/05/ASHG-TEConomy-Impact-Report-Final.pdf.

example, 23andMe partnered with GlaxoSmithKline (GSK) in 2018 (following a $300 million investment by GSK in 23andMe)[15] to use human genetic data as the basis for "the development of innovative new medicines and potential cures".[16] Human genetic research has an annual U.S. federal government investment of $3.3 billion through research funding.[17] The collection of large, diverse genomic datasets (such as those held by U.S. companies) by other nations can boost their global market share of the genetics and pharmaceutical industries, particularly when there is no reciprocal sharing of health data by these other nations.

C. Calculating the Probability of Identifying a Target from Human Genetic Datasets

As mentioned previously, different genetic datasets can be combined for the purposes of identification. For example, a sample collected from a crime scene and analyzed using either STR or SNP genotyping can be queried against a SNP dataset originally collected by a DTC genetic testing company in an attempt to identify a suspect. If no exact genetic match is present within the SNP dataset, investigators can also determine whether there are partial matches, which could indicate the presence of the suspect's family member within the SNP dataset. While this approach could be beneficial for society for law enforcement, it can also be used for genetic surveillance.

To contextualize this process, we have modeled an estimate of the following question: What is the probability of identifying a person by matching a collected genetic sample with a SNP record of either themselves or a family member as a function of SNP dataset size?

This overall probability consists of three separate events:

1. The target or at least one of their relatives is in the SNP dataset record.
2. The collected and analyzed sample can be matched with the SNP record of the target or at least one of their relatives.
3. Genealogical investigation can be used to generate a family tree that enables the identification of the target.

---

[15] Livescience and Laura Geggel, "23andMe Is Sharing Genetic Data with Drug Giant," *The Scientific American*, July 18 2018, accessed November 2023, https://www.scientificamerican.com/article/23andme-is-sharing-genetic-data-with-drug-giant/.

[16] "GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines," GSK, July 25 2018, accessed November 2023, https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.

[17] Simon Tripp and Martin Grueber, "The Economic Impact and Functional Applications of Human Genetics and Genomics," TEConomy Partners, LLC, May 2021, https://www.ashg.org/wp-content/uploads/2021/05/ASHG-TEConomy-Impact-Report-Final.pdf.

This section describes the methodology the IDA team used. Let *A* refer to the person from whom the genetic sample (GS_A) was collected (the target individual) and let *B* refer to a relative of person *A* whose SNP sample (SNP_B) is included in the SNP database (*A* can be the same person as *B*). The goal of this analysis is to characterize how the probability of matching GS_A to SNP_B varies based on 1) the size of the database that includes SNP_B, 2) the number of loci in GS_A, and 3) the familial relationship between *A* and *B* (i.e., self, sibling, parent, etc.).

We calculated the probability of matching GS_A to SNP_B as the product of three probabilities: 1) the probability that SNP_B is included within the SNP dataset, 2) the probability that GS_A can be matched to SNP_B, and 3) the probability a family tree linking *B* to *A* can be assembled in order to discern the identity of *A*. The following two sections will describe in additional detail the steps and assumptions associated with these three probabilities.

1. Probability that *A* or their relative out to the third-degree exists in a SNP dataset

The first probability, that SNP_B exists within the SNP dataset, depends upon the population coverage, which is the percentage of a population that is included within the SNP dataset. For this study, we made the simplifying assumption that all individuals are equally likely to be present within the dataset representing their population. An example of this could be: if it is known that a SNP dataset includes 2% of all U.S. citizens, the probability of any U.S. citizen existing in the dataset would be 2%. A second example could be that if a dataset is known to include 25% of the residents of the city of Chicago, any resident of Chicago would have a 25% chance of being in the dataset.

In reality, the probability of an individual being included within a SNP dataset depends on multiple factors. For example, people of Northern European backgrounds tend to be over-represented in SNP datasets compared to other ethnicities. Further, about 75% of the MyHeritage dataset is comprised of individuals with a Northern European genetic background.[18] Northern Americans and Europeans, meanwhile, only comprise about 14% of the worldwide population.[19] Also, as of December 2019, members of the U.S. Armed Forces were advised against the use of DTC genetic tests and consequently may be less

---

[18] Yaniv Erlich et al, "Identity inference of genomic data using long-range familial searches," *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

[19] United Nations, Department of Economic and Social Affairs, *World Population Prospects 2022 Summary of Results* (New York, 2022): 5, https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

likely than a member of the general U.S. population to appear in an SNP dataset.[20] With the variability in ethnic population makeup in SNP datasets and without additional data on prevalence of military personnel in DTC genetic datasets, the assumption that all individuals are equally likely to appear in the SNP dataset is necessary because it allows us to estimate the probability that SNP_B is in a SNP dataset based only on the population coverage (i.e., size) of the dataset. Should more granular data become available, calculating these probabilities would be possible.

To estimate the probability that one of *A*'s relatives exists within the SNP dataset, we needed to assume a specific family tree. The average number of children per couple in the United States was 2.4 in 2015[21], which we took to be 2 or 3 children per couple to simplify visualization and explanation of the family tree and to avoid "partial" individuals.[22] As summarized in our assumed family structure in Figure 1, we assume the target has 2 siblings, one of their parents has one sibling and the other parent has two siblings, and half of their grandparents have two siblings while the other half have one sibling. This results in 4 grandparents, 3 aunts/uncles, 2 parents, 2 siblings, 7 first cousins, 6 grand aunts/uncles, and the target (*A*) for a total of 25 individuals in the assumed family tree. In the figure, the gray boxes show the number of each relationship type we assumed and the orange boxes refer to the degree of relationship. The results we present later in this paper are specific to this family structure, therefore the actual probability of any specific person being identified would depend upon their unique family tree (among other factors, see Section E for additional detail).

---

[20] U.S. Department of Defense, Office of the Secretary of Defense, *Direct-to-Consumer Genetic Testing Advisory for Military Members* (Washington, D.C, 2019): 1.

[21] Gretchen Livingston, *Childlessness Falls, Family Size Grows Among Highly Educated Women* (Washington D.C.: Pew Research Center, 2015), 11.

[22] We did calculate the difference between using 2.5 children per generation vs 2-to-3 children per generation and the results were nearly identical, so we decided to use a more easily explainable family tree structure with minimal sacrifices to accuracy.

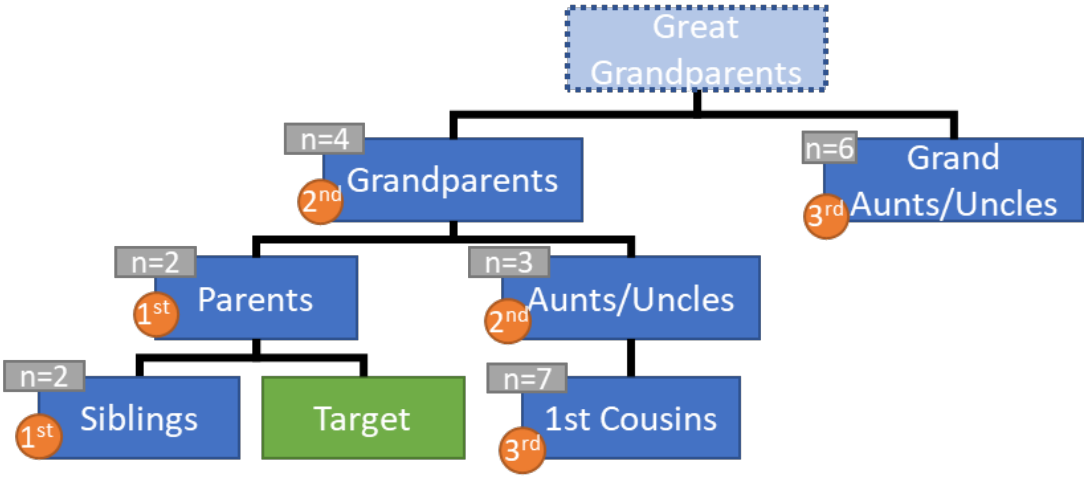


Figure 1. Assumed Family Structure for Up to Third-Degree Relatives

First-degree relatives are those with the closest genetic similarity to an individual (other than identical twins). This group of relatives includes an individuals' parents, siblings, and children. Second-degree relatives include grandparents and aunts/uncles and third-degree relatives include first cousins. For this analysis, we made the assumptions that all couples are monogamous (i.e., no half relatives) and non-incestuous, which are common simplifying assumptions used to estimate kinship, but will cause an underestimation in the probability that a relative of the target is in the SNP dataset. Additionally, we assumed the target's generation is the most recently born generation to have a profile in the SNP dataset, which excludes the target's children, nieces and nephews, and cousins' children. We assumed that individuals under 18 years of age likely will not have their genetic data in the SNP dataset and that the target individual (or any of their relatives from the same generation) does not have children above 18 years of age. We also assumed that the generation of the target's grandparents are the oldest generation that could be in the SNP dataset.

The probability that the SNP dataset includes either *A* or at least one relative of up to third degree is calculated assuming everyone in the United States has an equal chance of being included in the SNP database. This probability is determined based on the number of first-, second-, and third-degree relatives an individual has. Therefore, individual A is assumed to have 24 first-, second-, and third-degree relatives. Calculating the probability that individual A or any one of their first-, second-, or third-degree relatives is in the SNP dataset is done by using the following equation:

$$1 - (1 - \frac{N_{SNP}}{N})^r$$

N = Size of U.S. population; N_{SNP} = size of U.S. population in the SNP dataset; r = individual A + number of relatives

Figure 2 shows the probability of an individual or a relative up to third-degree exists in datasets of different sizes. In this figure, we have varied the value for r to show probabilities of 1) the target existing in the SNP dataset (r = 1), 2) the target or any of their first-degree relatives existing in the SNP dataset (r = 5), or 3) the target or any of their first-, second-, or third-degree relationships existing in the SNP dataset (r = 25). While the probability individual A is within a dataset scales linearly with the size of the dataset, the probability that one of their first-, second-, or third-degree relatives is in the dataset increases at a more rapid rate.

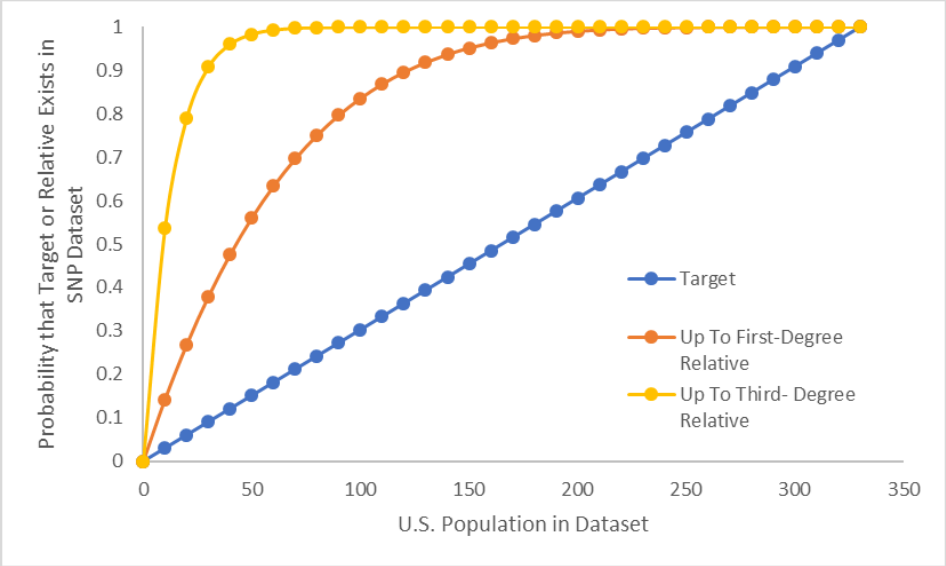


Figure 2. Probability of Individual A or Up to First- or Third-Degree Relatives Existing in SNP Dataset

2. Probability that Linkage Disequilibrium can be used to match GS_A to SNP_B

The second step of the methodology is to determine the probability of matching GS_A to SNP_B given that B is in the SNP dataset. This probability depends on a process called linkage disequilibrium (LD). Linkage disequilibrium describes the phenomenon by which genes, particularly those close to each other on the genome, are inherited together. Due to LD, certain genotype pairs are more or less likely to co-occur and can be used to match loci of two different datasets, even if none of those loci pairs are genotyped together in the same dataset. To determine the probability that LD can be used to match GS_A and SNP_B, we relied upon the results from four academic papers. Edge et al., determined the probability that LD can be used to match an STR sample from an individual to a SNP

sample from the same individual (i.e., matching GS_A and SNP_B when $B = A$).[23] Kim et al.[24], determined the probability that LD can be used to match an STR sample from an individual to a SNP sample from that individual's first-degree relative (i.e., matching GS_A and SNP_B if A is the target and B is a first-degree relative of A).[25] Finally, de Vries et al. and Morimoto et al. determined the probability that LD can be used to match SNP sample (GS_A) to a SNP record (SNP_B) if A is the target and B is a first-, second-, or third-degree relative.[26]

Both Edge et al. and Kim et al. utilized previously reported data from the Human Genome Diversity Panel, using 642,563 SNP loci and 431 non-CODIS STR loci from 872 individuals representative of the worldwide population for their analyses.[27] In both studies, the authors generated figures demonstrating the effect of increasing the number of STR loci tested on the probability of successful LD matching. We used the median accuracy results included in Figure 4C of Edge et al. to determine the probability that LD can be used to match GS_A to SNP_B in the case that B = A (i.e., matching the target to themselves). Likewise, data represented in Figure 4B and 4C of Kim et al., were used to determine the probability that LD can be used to match GS_A to SNP_B in the case that B is a first-degree relative of individual A.

Record matching accuracy values between STR tests and SNP records were not available for second- and third-degree relationships. To estimate these values, we used the median coefficient of kinship values for second- and third-degree relationships for GS_A analyzed with 20 STR (25% for grandparents and aunts/uncles and 12.5% for first cousins). The coefficient of kinship values for parents and siblings corresponded closely with the record matching accuracy values observed by Edge et al. and Kim et al. for first-degree relationships, so we assumed that these values could be used to estimate record matching for second- and third-degree relationships as well. In cases where GS_A was analyzed with STR tests using 40 STR, we calculated the average improvement in record matching accuracy for parents and sibling relationships (1.78) when increasing from 20 STR to 40

---

[23] Michael D. Edge et al, "Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets," *PNAS* 144 no. 22 (2017): 5671-5676, https://doi.org/10.1073/pnas.1619944114.

[24] This paper was from the same lab as Edge et al.

[25] Jaehee Kim et al, "Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci," *Cell* 175 (2018): 848-858, https://doi.org/10.1016/j.cell.2018.09.008.

[26] Jard H de Vries et al, "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy," *Forensic Science International: Genetics* 56 (2022), https://doi.org/10.1016/j.fsigen.2021.102625.

Chie Morimoto et al, "Pairwise Kinship Analysis by the Index of Chromosome Sharing Using High-Density Single Nucleotide Polymorphisms," *PLoS ONE* 11 7 (2016): e0160287, https://doi.org/10.1371/journal.pone.0160287.

[27] This will be mentioned again in the limitations section, but worldwide population diversity does not equal U.S. population diversity, thus these results should be taken as rough estimates.

STR loci and multiplied our coefficient of kinship values by this value to approximate the increase in accuracy for testing with 40 STR in second- and third-degree relationships.

To determine the record matching accuracy when comparing a SNP sample from *A* and the SNP record from *B*, we used published values from de Vries and Morimoto and averaged them together. De Vries et al. used SNP samples from 24 Northern European donors from the Dutch blood bank, which were then tested at 652,027 SNP loci.[28] Morimoto et al. used SNP samples gathered from 1,498 Japanese donors, which were genotyped at 174,254 SNP loci.[29]

Table 1 shows the probability of matching (P_{LD}) for various numbers of STR and SNP test loci from Edge, Kim, de Vries, and Morimoto.

Table 1. Probability that LD Can Be Used to Match GS_A to SNP_B

Relationship	Probability LD can be used to match STR (20 loci) to SNP	Probability LD can be used to match STR (40 loci) to SNP	Probability LD can be used to match SNP to SNP
Individual	95%*	100%*	> 99%***
Parent	45%*	80%*	>99%***
Sibling	50%*	85%*	99%***
Second-degree	25%**	43%**	94%***
Third-degree	12.5%**	22%**	91%***

* These values represent an STR test that assesses 20 or 40 loci. Similar tables can be generated from the dataset presented by Edge and Kim for 5-100 loci. This also assumes that the SNP dataset is of a similar size to the dataset analyzed by Edge and Kim (642,563 loci). A SNP dataset including more loci would likely have higher match probabilities than those presented in this table.

** These values were estimated using coefficient of kinship values.

*** The values in this column represent the median match probabilities between those reported by de Vries and Morimoto, which tested at 652,027 loci and 174,254 loci respectively.

The probability of unsuccessfully matching the target through a specific relative, either because the relative was not in the SNP dataset, or because matching through LD was unsuccessful is calculated using the following formula:

$$P_B = (1 - P_{LD}\frac{N_{SNP}}{N})$$

---

[28] Jard H de Vries et al, "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy," *Forensic Science International: Genetics* 56 (2022), https://doi.org/10.1016/j.fsigen.2021.102625.

[29] Chie Morimoto et al, "Pairwise Kinship Analysis by the Index of Chromosome Sharing Using High-Density Single Nucleotide Polymorphisms," *PLoS ONE* 11 7 (2016): e0160287, https://doi.org/10.1371/journal.pone.0160287.

Note: P_{LD} refers to the probability that LD can be used to match GS_A to SNP_B; P_B refers to the probability of matching to any specific relative (B).

To calculate the overall probability of a match with at least one relative, the P_B values for each relative can be multiplied together and then subtracted from 1, as demonstrated in the following equation for matching with self or a first-degree relative.

$$P_{match} = 1 - (P_{self} * {P_{parent}}^2 * {P_{sibling}}^2)$$

This formula combines the probability that *B* is included within the SNP dataset (Figure 2) and the probability that LD can be used to match GS_A and SNP_B when an STR test used to analyze GS_A assesses 20 or 40 loci or a SNP test used to analyze GS_A assesses 174,254-652,027 loci (Table 1).

3. Probability that a family tree linking person B to person A can be assembled

To determine the identity of person *A*, after a match is made between GS_A and SNP_B, a family tree needs to be constructed to determine the relationship between person *A* and person *B*. For example, if person *B* was person *A*'s first cousin, it would not initially be clear which cousin *A* might be. Additional data types, such as census data, vital records (e.g., marriage, birth, and death certificates), the Social Security Death Index, and newspapers.com have been compiled into searchable databases available via subscription through services such as Ancestry.com and public social media accounts can be used to glean additional information.[30] Even with this, the process can be challenging.

Genealogy is a time-consuming, sometimes expensive step in the forensic genetic genealogy process. Cost estimates for forensic genetic genealogy are typically in the range of $65,000 per case; $15,000 of this cost is associated with laboratory analysis and genealogical research with the remainder being used for investigative costs.[31] To identify person *A*, person *B* needs to be definitively identified, then a common ancestor between the two must be determined.[32] Then the family tree must be built "forward" in time from

[30] Ellen M. Greytak, CeCe Moore, and Steven L. Armentrout, "Genetic Genealogy for Cold Case and Active Investigations," *Forensic Science International* 299 (2019): 107.

[31] Ray Wickenhauser, "Investigative Genetic Genealogy: Current Status and Future Potential." *Forensic Science International: Synergy 3* (2021).

[32] Ellen M. Greytak, CeCe Moore, and Steven L. Armentrout, "Genetic Genealogy for Cold Case and Active Investigations," *Forensic Science International* 299 (2019): 108.

the common ancestor out to person A.[33] This process is not always successful, even after tens or hundreds of hours of research.

To capture the probability of successfully assembling a family tree, we conducted a literature search for publications attempting to quantify the success rate of identification following a match between GS_A and SNP_B. We found two studies[34] and a news article[35] that informed our values for this probability, the results of which are summarized in Table 2. For cases that provided only values in centimorgans (cM), we assumed a degree of relationship based on the range of cM shared between different relationship types.[36]

Table 2. Forensic Genetic Genealogy Case Summaries

Reference	Relationship Description from Paper	Assumed Degree of Relationship*	Solved	Time to Solve
Thompson	Parent or sibling	1	Yes	3 hours
Aldhous	Parent or sibling	1	Yes	N.L.
Ertürk	1550 cM	2	Yes	N.L.
Aldhous	First cousin	3	Yes	< 2 hours
Ertürk	360 cM	4	Yes	N.L.
Ertürk	280 cM	4	Yes	N.L.
Thompson	Second cousin	5	Yes	50-100 hours
Ertürk	240 cM	5	No	N.L.
Ertürk	170 cM	5	Yes	N.L.
Ertürk	140 cM	5	Yes	N.L.
Aldhous	Second cousin	5	Yes	Within “hours”
Aldhous	Second cousin	5	No	N.L.
Aldhous	Second cousin	5	No	N.L.

[33] Ellen M. Greytak, CeCe Moore, and Steven L. Armentrout, “Genetic Genealogy for Cold Case and Active Investigations,” *Forensic Science International* 299 (2019): 107.

[34] Jim Thompson, Tim Clayton, John Cleary, Maurice Gleeson, Debbie Kennett, Michelle Leonard, and Donna Rutherford, “An Empirical Investigation into the Effectiveness of Genetic Genealogy to Identify Individuals in the UK,” *Forensic Science International: Genetics* 26 (2020): 102263.

Mine Su Ertürk, Colleen Fitzpatrick, Margaret Press, and Lawrence M. Wein, “Analysis of the Genealogy Process in Forensic Genetic Genealogy,” *Journal of Forensic Sciences* 67 (2022): 2218-2229.

[35] Peter Aldhous, “We Tried to Find 10 BuzzFeed Employees Just Like Cops Did for the Golden State Killer,” *BuzzFeed News*, April 9, 2019, https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy.

[36] Living DNA, “What does my relationship prediction mean,” last updated in 2020, https://support.livingdna.com/hc/en-us/articles/360013536560-What-does-my-relationship-prediction-mean-.

Reference	Relationship Description from Paper	Assumed Degree of Relationship*	Solved	Time to Solve
Aldhous	Half second cousin	6	Yes	N.L.
Ertürk	120 cM	6	Yes	N.L.
Ertürk	120 cM	6-7	No	N.L.
Ertürk	100 cM	6-7	No	N.L.
Ertürk	90 cM	6-7	No	N.L.
Ertürk	80 cM	6-7	No	N.L.
Ertürk	80 cM	6-7	Yes	N.L.
Ertürk	80 cM	6-7	No	N.L.
Ertürk	70 cM	6-7	No	N.L.
Ertürk	70 cM	6-7	No	N.L.
Ertürk	60 cM	6-7	Yes	N.L.
Ertürk	60 cM	6-7	Yes	N.L.
Thompson	Third cousin	7	Yes	50-100
Aldhous	Third cousin	7	No	30+
Aldhous	Third cousin	7	No	N.L.
Aldhous	>third cousin	7+	No	N.L.
Thompson	Third to fourth cousin	7-9	No	100+
Thompson	Third to fourth cousin	7-9	No	100+
Thompson	>fourth cousin	9+	No	100+
Thompson	Fourth cousin	9	No	100+
Thompson	Fourth cousin	9	Yes	50-100
Thompson	>fourth cousin	9+	No	100+
Thompson	>fourth cousin	9+	No	100+
Aldhous	>fourth cousin	9+	No	N.L.

*If the assumed degree of relationship could not be determined (e.g., was noted as “greater than third cousin” or “between third and fourth cousin”, it was excluded from our determination of identification probability.

N.L. means the time to identify was “not listed” in the referenced article.

In our literature review, 100% of targets in cases where there was a match with up to a third-degree relative were able to be identified, thus we assumed the probability for successful generation of a family tree to be 1 when examining relatives out to the third-degree. Not all of the cases were associated with a time to identification, but from the data

we gathered, shorter identification times were associated with closer relationships between *A* and *B*.

D. Probability of Identifying a Target from Human Genetic Datasets

Using the methodology described in Section C, Figure 3 depicts the probability of identifying a target (*A*) from a collected genetic sample analyzed using an STR or SNP assay (examining 20 or 40 STR loci or hundreds of thousands of SNP loci) with a SNP record of either the target, the target or one of their first-degree relatives, or the target or one of their relatives up to the third degree as a function of SNP dataset size. Additionally, in Appendix A, we expand this figure to also include the probability of matching the genetic sample of the target to the SNP record of fourth- and fifth-degree relatives, but with additional caveats.

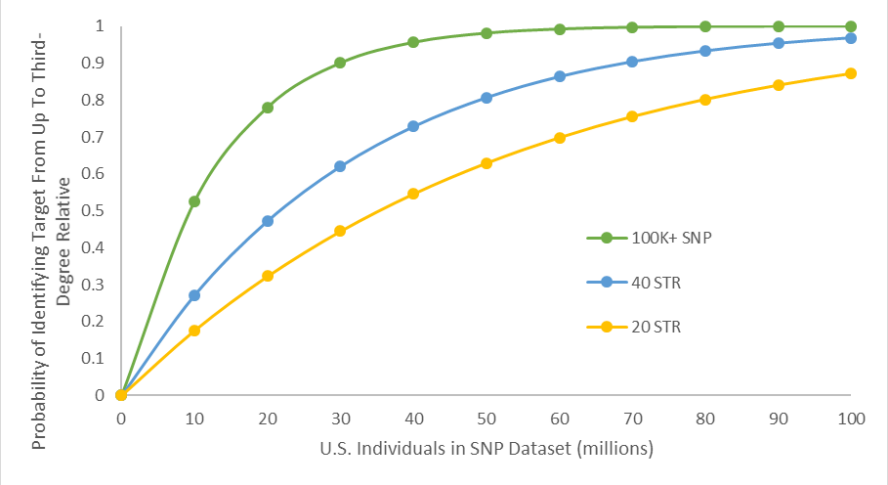


Figure 3. Overall Probability of Identifying a Target (*A*) from Their Genetic Data or that of Up to a Third-Degree Relative Using Various Genetic Test Types as a Function of SNP Dataset Size

When assuming the probability of assembling a family tree to identify *A* is 1 for up to third-degree relatives, the probability of a successful identification depends primarily upon the population coverage of the SNP dataset (i.e., the proportion of a population contained within the dataset compared to the overall population) and the number of STR or SNP loci (i.e., positions on the genome) that are being tested.

Table 3 provides estimates of match probability for specific SNP dataset sizes from 100 to 6.6 million SNP profiles (approximately 2% of the U.S. population).

Table 3. Calculated Probabilities Identifying Target (*A*) from Themselves, or Up to a Third-Degree Relative Using Various Genetic Test Types

		Type of Collected Genetic Sample		
Number of SNP Profiles in Dataset	% of US Population	20 STR loci (forensic)	40 STR loci (forensic)	100K+ SNP loci (DTC)
100	0.00003%	<0.1%	<0.1%	<0.1%
1,000	0.0003%	<0.1%	<0.1%	<0.1%
10,000	0.003%	<0.1%	<0.1%	<0.1%
100,000	0.03%	0.19%	0.31%	0.73%
1,000,000	0.3%	1.91%	3.10%	7.11%
3,300,000	1%	6.16%	9.89%	21.66%
6,600,000	2%	11.97%	18.84%	38.77%

Finally, we wanted to explore how the probability of identification changes when just considering the target, considering up to first-degree relatives, or up to third-degree relatives. Figure 4 shows the overall probability of identification when GS_A is analyzed with an STR test examining 20 loci when considering only whether the target is identified from their own SNP data, considering whether the target is identified by their own SNP data or that of one of their first-degree relatives, or considering whether the target is identified from with their own SNP data or that of up to their third-degree relatives. Figure 5 and Figure 6 display analogous cases when the GS_A is analyzed using an STR test examining 40 loci or a SNP test examining over 100,000 loci, respectively.

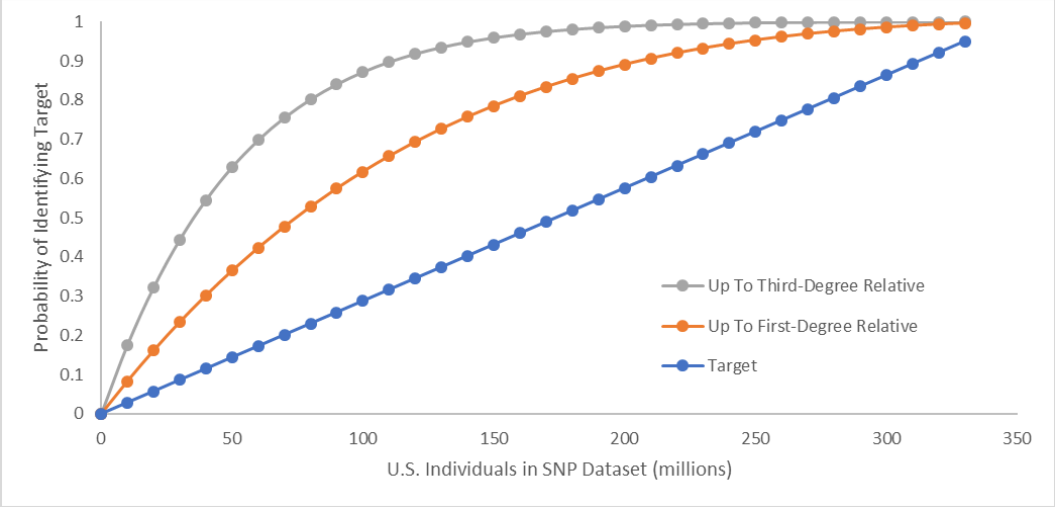


Figure 4. Overall Probability of Identifying Target (*A*) from Collected STR Sample (20 Loci) from Various Relationship Types

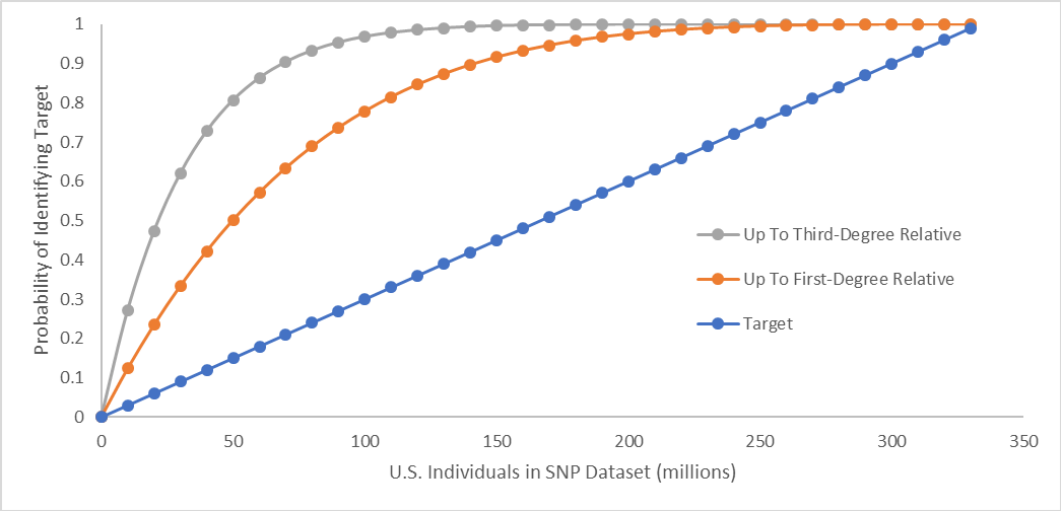


Figure 5. Overall Probability of Identifying Target (*A*) from Collected STR Sample (40 Loci) from Various Relationship Types

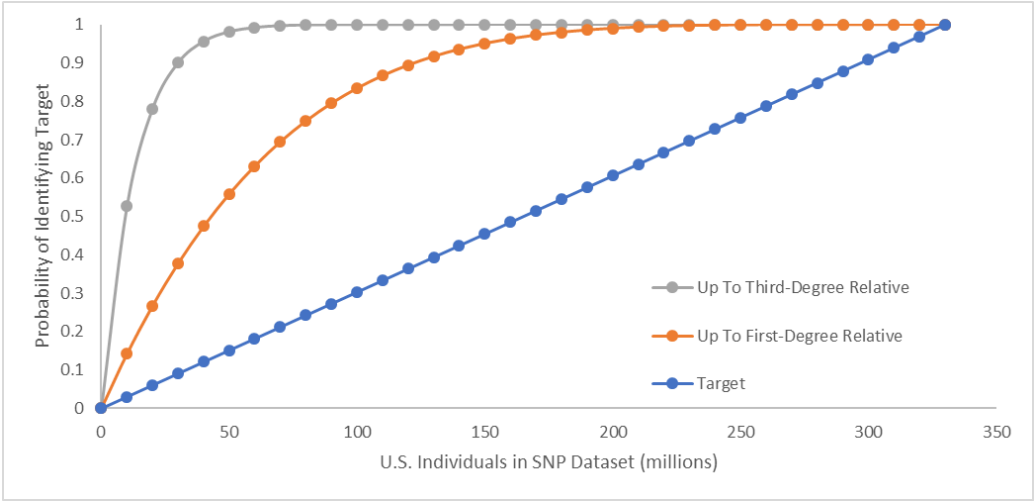


Figure 6. Overall Probability of Identifying Target (*A*) from Collected SNP Sample (100K+ Loci) from Various Relationship Types

As demonstrated in these figures, the relationship between probability of identification and SNP dataset size when only considering the target matching with themselves is a linear relationship, whereas the probabilities increase at a faster rate with SNP dataset size when relatives are considered.

Finally, we determined the impact on the overall probability of identification when person *A* is known not to be in the SNP dataset. Figure 7 shows the calculated probability of identification using up to a third-degree relative assuming that the target is not in the SNP dataset. When P_{LD} is lower (as in the 20 STR case), assuming the target is not in the SNP dataset makes a larger difference in probability compared to when P_{LD} is higher (as in the 100,000+ SNP case). In all cases, however, there is a non-zero probability of

identification from datasets of a comparable size to those collected by companies such as 23andMe, even if person *A* is assumed to not be in the SNP dataset. This highlights the importance of considering genetic data protection more broadly than discouraging individual consumers from taking DTC tests, as having one family member of first-, second-, or third-degree relatives in a SNP dataset makes that entire group of people more vulnerable to genetic surveillance or exploitation.

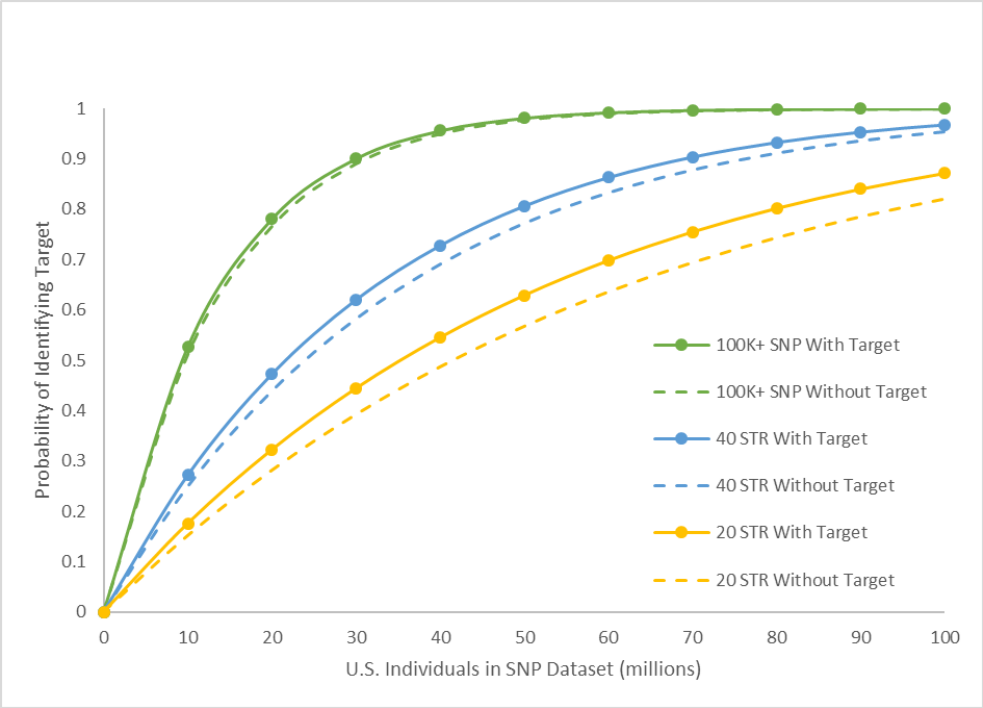


Figure 7. Overall Probability of Identifying Target from Up to Third-Degree Relatives When Target is Assumed to Not Be in the SNP Dataset

E. Assumptions and Limitations

A few assumptions and limitations are inherent with the approach used to estimate the identification probability and are necessary to interpret Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7. The magnitude of the uncertainty inherent in these results and the sensitivity of the results to the following assumptions is unclear without further analysis:

1) A specific family tree was modeled based on an "average" U.S. individual, where 2 or 3 children per couple was assumed. This resulted in 4 first-degree family members, and 24 first-, second-, and third-degree family members. Additionally, we assumed all couples were monogamous and non-incestuous.

2) We assumed that all people within the United States are equally likely to be represented in the SNP dataset. In reality, this is unlikely to be the case, as DTC

SNP datasets tend to overrepresent Americans of Northern European ancestry[37] and certain populations (e.g., active-duty military) have been advised against taking these tests.[38]

3) The probabilities of matching STR or SNP data to DTC SNP records were developed based off peer reviewed publications[39] that analyzed a dataset representing the genetic diversity of populations other than the U.S. population.[40] Ideally, as our audience is primarily geared towards U.S. government and industry stakeholders, the underlying data would be more representative of the population within the United States. However, the IDA team was unable to find data that fit those requirements. Of note, Edge et al. did examine the impact of genetic diversity among populations on the matching accuracy and found that the average match scores will differ between genetic populations with more or less genetic diversity (with less genetically diverse populations having higher overall match scores). The overall U.S. population is genetically diverse, though the worldwide population is not necessarily representative of the population distribution found in the United States.

4) Edge et al. and Kim et al. provided ranges of probabilities for P_{LD} depending on which genetic loci were selected. For this analysis, we modeled the median probability values.

5) There were only limited instances in the literature quantifying the success rate of generating a family tree following a successful genetic match to a relative.[41] In

---

[37] Yaniv Erlich et al, "Identity inference of genomic data using long-range familial searches," *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

[38] U.S. Department of Defense, Office of the Secretary of Defense, *Direct-to-Consumer Genetic Testing Advisory for Military Members* (Washington, D.C., 2019): 1.

[39] Michael D. Edge et al, "Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets," *PNAS* 144 no. 22 (2017): 5671-5676, https://doi.org/10.1073/pnas.1619944114.

Jaehee Kim et al, "Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci," *Cell* 175 (2018): 848-858, https://doi.org/10.1016/j.cell.2018.09.008.

Jard H de Vries et al, "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy," *Forensic Science International: Genetics* 56 (2022), https://doi.org/10.1016/j.fsigen.2021.102625.

Chie Morimoto et al, "Pairwise Kinship Analysis by the Index of Chromosome Sharing Using High-Density Single Nucleotide Polymorphisms," *PLoS ONE* 11 7 (2016): e0160287, https://doi.org/10.1371/journal.pone.0160287.

[40] The STR to DTC SNP record matching probabilities were based off of the genetic diversity of the worldwide population, whereas the SNP to DTC SNP record matching probabilities were based off a Japanese population and a Dutch population.

all cases where a match was a third-degree relative of the target or closer, the target could be identified successfully; when times were reported, they were under 5 hours. For this estimation, we assumed that this would hold for all cases where a third-degree relative or closer could be matched, but there is a large degree of uncertainty with this assumption due to the limited number of data points. In other words, we assume that the probability of identification is 100% for all relationships out to third degree.

For these reasons, we present the results of this analysis as an estimate of one potential capability associated with acquisition of U.S. genetic data. We aimed to provide the sponsoring office with an illustrative case that could be used to demonstrate to stakeholders how the size of a SNP dataset may influence the probability of identifying a target.

Additionally, results should be considered a rough estimate of an "average" U.S. citizen, rather than exact figures that convey the probability of identifying any specific individual. Factors that could alter the probability of identifying a specific individual include: 1) the size and structure of that individual's family tree,[42] 2) the genetic diversity of the population the target belongs to,[43] 3) the overall representation of the target's genetic

---

[41] Jim Thompson, Tim Clayton, John Cleary, Maurice Gleeson, Debbie Kennett, Michelle Leonard, and Donna Rutherford, "An Empirical Investigation into the Effectiveness of Genetic Genealogy to Identify Individuals in the UK," *Forensic Science International: Genetics* 26 (2020): 102263.

Mine Su Ertürk, Colleen Fitzpatrick, Margaret Press, and Lawrence M. Wein, "Analysis of the Genealogy Process in Forensic Genetic Genealogy," *Journal of Forensic Sciences* 67 (2022): 2218-2229.

Peter Aldhous, "We Tried to Find 10 BuzzFeed Employees Just Like Cops Did for the Golden State Killer," *BuzzFeed News*, April 9, 2019, https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy.

[42] Erlich et al, modeled the size of a SNP dataset needed to have a relationship of third cousin or closer within the dataset for 99% of a population and found that 2% population coverage (assuming all individuals in the population are equally likely to be in the SNP dataset) was sufficient. However, the failure rate of genealogical investigations for relationships more distant than first cousins is not negligible and the time and cost associated with these investigations increases dramatically the more distant the matched relative.

[43] A second cousin match investigated by Aldhous failed because of the low genetic diversity within the group; they were incapable of establishing a family tree due to the interrelatedness of the population. Additionally, low genetic diversity populations can inflate the match probability, making it more difficult to distinguish between actual matches and non-matches (leading to a potentially higher level of false positives or false negatives, depending upon cutoff values).

Peter Aldhous, "We Tried to Find 10 BuzzFeed Employees Just Like Cops Did for the Golden State Killer," *BuzzFeed News*, April 9, 2019, https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy.

Michael D. Edge et al, "Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets," *PNAS* 144 no. 22 (2017): 5671-5676, https://doi.org/10.1073/pnas.1619944114.

population in DTC SNP datasets,[44] and 4) the availability and accessibility of non-genetic genealogical data.[45]

It is important to note that identification is possible, not only though first-, second-, and third-degree relatives, but also through more distant relatives. Indeed, law enforcement officials have used the SNP data of second cousins, third cousins, and other relatives to identify perpetrators of violent crimes.[46] The benefit of expanding genetic searching to more distant matches is that there is a greater likelihood of finding a partial match in the SNP dataset (increasing the probability in the first step); however, more distant relatives have lower genetic similarity to a target (decreasing the probability in the second step) and generating a family tree to identify a target from a second cousin or more distant relative was not always possible in the literature we reviewed[47] (decreasing the probability in the third step). Additionally, assembling these family trees can be time and cost prohibitive,[48] particularly for an application such as genetic surveillance where answers may be desired in a more rapid timeframe. Appendix A describes the quantification of the probability of identifying a target from more distant relatives, which may be useful for some applications where timeliness (on the order of days) is not crucial.

F. Conclusions

While there remains a need to be able to openly share data for scientific and health research, the protection of human genetic data has been a question or topic of interest for the last several years, with various groups raising concerns about impacts on privacy and national security. This paper determined the probability of identifying a target from human genetic datasets to inform potential policies, enforcement actions, and rulemaking concerning the privacy and national security concerns associated with the acquisition of U.S. persons sensitive personal data, including genomic data, by strategic competitors.

One of the notable takeaways from this study is the amount of time needed for the genealogical step of identifying a person, particularly when investigating family members

---

[44] For example, 75% of the MyHeritage dataset consists of individuals with primarily Northern European genetic backgrounds. Yaniv Erlich, et al., "Identity inference of genomic data using long-range familial searches," *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

[45] This can include factors such as: a family member uploading a family tree, publicly accessible social media profiles, most family members were born in countries with accessible public records (birth, death, and marriage).

[46] Yaniv Erlich et al, "Identity inference of genomic data using long-range familial searches," *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

[47] 57% of cases with a second cousin match were successful and 25% of cases with a third or fourth cousin match were successful.

[48] Second cousin matches took between 50-100 hours to generate a family tree, when successful. Third cousin and beyond matches took 50-100+ hours to generate family trees, when successful.

more distantly related than first cousins. However, if this process could be automated or more publicly available records made accessible, time and cost could become less substantial factors in the ability to identify individuals from distant relatives. Another finding that could alter the results of this study is the lack of available data on matching and identifying. Additional studies examining the probability of matching distant relatives and the time and likelihood of assembling a family tree would enable us to update our estimates with data we have greater confidence in.

Another type of data that may warrant additional exploration and consideration of protection may be publicly accessible vital records, social media profiles, and other supplementary datasets that can be used to accomplish this genealogical step. Notably, public records are likely to increase over time, as generations grow up with access to social media and online records, which could alter the probability of successful identification. Continued adoption of artificial intelligence to extract and synthesize meaningful data from diverse data sources such as the open web and publicly available information may lower the time and cost to generate a family tree in the future, resulting in a decreased amount of time to identify a target from human genetic datasets without the adoption of additional protections. While open access to data has numerous benefits to society, unintended dual use of this data should be considered when designing protection efforts.

Finally, it is clear that policies encouraging individuals in sensitive positions (such as active-duty military personnel) to avoid DTC genetic tests is insufficient in protecting those individuals' privacy, as these individuals may be identified through family members whose behavior they have little control over. More comprehensive protection measures of large human genetic datasets may be necessary to preserve the privacy of U.S. persons and mitigate national security concerns associated with the acquisition of U.S. persons' sensitive personal data, including genomic data, by strategic competitors. Additionally, there are other methods of identification of persons that do not rely upon genetic data, including mobile phone data. This analysis did not investigate these other methods or compare the efficacy or resource intensity of genetic identification to other methods. To fully understand the risk genetics could pose to privacy or national security, a full characterization of alternative methods that can be used either by themselves or in combination with genetics could be beneficial.

Appendix A. Identifying Individuals from Genetic Data of Distant Relatives

While matching the genetic sample from person *A* (GS_A) to the SNP record of person *B* (SNP_B) of out to third-degree relatives will have a higher probability of being able to successfully use LD, it is also possible to perform similar GS_A to SNP_B matching to more distant relatives, such as fifth-degree relatives (e.g., second cousins). As any individual is more likely to have more second cousins than they do parents or siblings, widening the scope to more distant relatives has the benefit of increasing the probability that a relative exists in the SNP dataset without increasing the population coverage included within that dataset. However, the genetic similarity between these distant relatives is lower than that of first-, second-, and third-degree relatives. Moreover, assembling a family tree for relatives more distant than first cousins took researchers days or weeks as opposed to hours and was not always successful.[49]

In this appendix, we will explore the probability of identification considering up to fifth-degree relatives, expanding our discussion from Chapter 2 to include fourth-degree (first cousins once removed) and fifth-degree relatives (second cousins and first cousins twice removed) to illustrate how distant relationships can influence the probability of identification. Figure A-1 shows the family structure we assumed for this analysis, including the number of individuals we considered for each relationship type.

---

[49] Jim Thompson, Tim Clayton, John Cleary, Maurice Gleeson, Debbie Kennett, Michelle Leonard, and Donna Rutherford, "An Empirical Investigation into the Effectiveness of Genetic Genealogy to Identify Individuals in the UK," *Forensic Science International: Genetics* 26 (2020): 102263.

Mine Su Ertürk, Colleen Fitzpatrick, Margaret Press, and Lawrence M. Wein, "Analysis of the Genealogy Process in Forensic Genetic Genealogy," *Journal of Forensic Sciences* 67 (2022): 2218-2229.

Peter Aldhous, "We Tried to Find 10 BuzzFeed Employees Just Like Cops Did for the Golden State Killer," *BuzzFeed News*, April 9, 2019, https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy.

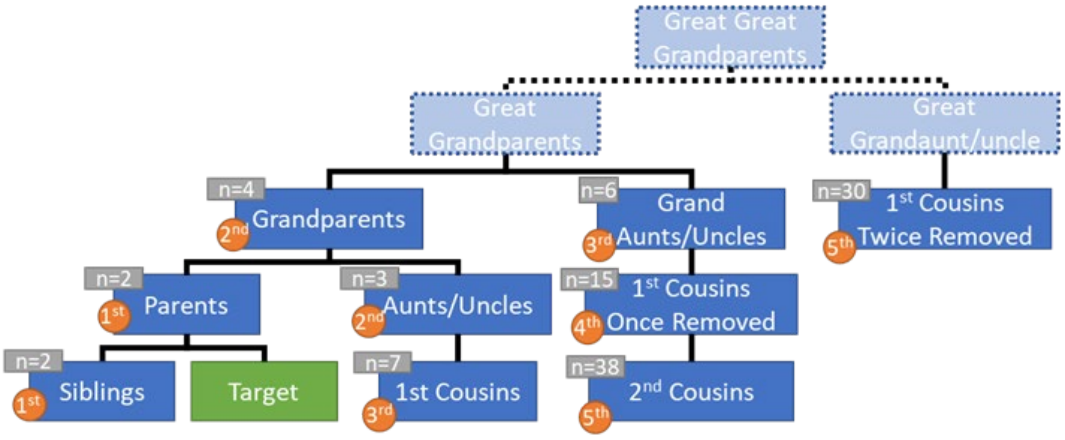


Figure A-1. Assumed Family Structure for Up to Fifth-Degree Relatives

We can use the same approach described in Chapter 2C to calculate the probability of identifying the target. The probability that the target cannot be identified, either because no relative is in the SNP dataset, matching is not successful, or assembling a family tree is not successful is calculated using the following formula:

$$P_B = (1 - P_{LD} P_{GG,B} \frac{N_{SNP}}{N})$$

Note: P_{LD} refers to the probability that LD can be used to match GS_A to SNP_B and P_{GG} refers to the probability that assembling a family tree linking person A and person B is successful.

To calculate the overall probability of a match with any relative, the P_B values for each relative type can be multiplied together and then subtracted from 1. We assume the number of fourth-degree relatives to be 14 and the number of fifth-degree relatives to be 68 individuals, which were calculated based on the assumption that each couple has 2 or 3 children.

As in Chapter 2C, there are probabilities associated with being able to match GS_A to SNP_B, which are noted in Table A-1. These probabilities were obtained in the same manner as those for up to third-degree relationships. We only have references for out to fifth-degree relatives for SNP data and not STR data. Higher degrees of relationships can be used, but we don't have the probability of the genetic matching for all degrees past fifth or the data for STR-SNP matching past third-degree.

Table A-1. P_{LD} and P_{GG} for Up to Fifth-Degree Relatives

Relationship	P_{LD} for SNP-SNP Matching	P_{GG} for Identification
Individual	> 99%*	100%
Parent	>99%*	100%
Sibling	99%*	100%
Second-degree	94%*	100%
Third-degree	91%*	100%
Fourth-degree	56.7**	66%
Fifth-degree	78%*	66%

* The values in this column represent the median match probabilities between those reported by de Vries and Morimoto, which tested at 652,027 loci and 174,254 loci respectively.

** The values in this column were from Morimoto et al. as de Vries et al. did not report these values for fourth-degree relatives. These values are lower than that of fifth-degree relatives likely because there were very few fifth-degree relatives in Morimoto's dataset.

Finally, we used the available literature data presented in Table 2 to calculate the percent of cases where a fourth- or fifth-degree match were made to determine the probability of successful identification. We did not consider cases where there was uncertainty in the degree of relationship. It is important to note that the amount of time it took to assemble family trees for fifth-degree relationships was approximately an order of magnitude higher than for first- through third-degree relationships. If time from sample collection to identification is of the essence for a particular capability, conducting this investigation out to fifth-degree relatives may be prohibitive for most cases.

Figure A-2 depicts the probability of identifying a target individual from GS_A analyzed using an STR or SNP assay (examining 20 or 40 STR loci or hundreds of thousands of SNP loci) with a SNP record of the target or one of their relatives out to fifth-degree as a function of SNP dataset size. Again, we are only showing SNP-SNP identification data because we lack matching data for STR-SNP in fourth- or fifth-degree relatives.

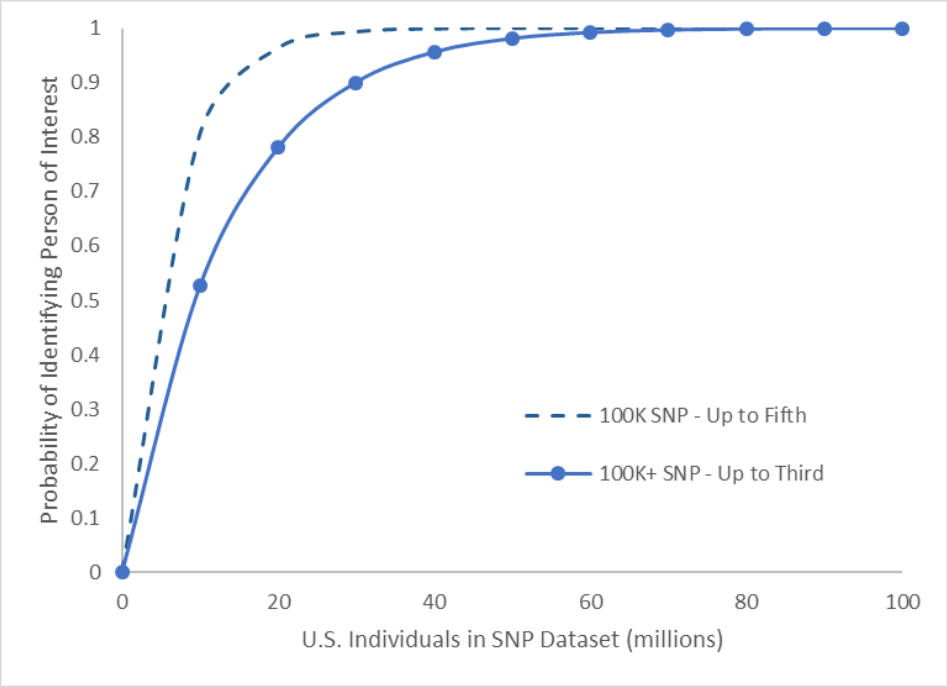


Figure A-2. Overall Probability of Identifying Target from Up to Third-Degree or Up to Fifth-Degree Relative Using an SNP Genetic Test as a Function of SNP Dataset Size

As shown in Figure A-2, the probability of identifying the person of interest increases when considering more distant relationships. However, it is important to consider that the amount of time it takes to conduct the genealogical investigation may be time and cost-prohibitive for genetic surveillance purposes. Additionally, more data that can be used to derive P_{LD} values for STR matching is needed to characterize how probability of identification changes by considering distant relatives.

Appendix B. Illustrations

Figures

Tables

This page is intentionally blank.

Appendix C. References

"GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines." GSK. July 25 2018, accessed November 2023. https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.

"Executive Order 14117 of February 28, 2024, Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern," *Code of Federal Regulations* (2024): 15421-15430, https://www.govinfo.gov/content/pkg/FR-2024-03-01/pdf/2024-04573.pdf.

Aldhous, Peter. "We Tried to Find 10 BuzzFeed Employees Just Like Cops Did for the Golden State Killer." BuzzFeed News, April 9, 2019. https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy.

Chie Morimoto, Sho Manabe, Takahisa Kawaguschi, Chihiro Kawai, Sjuntaro Fujimoto, Yuya Hamano, Ryo Yamada et al. "Pairwise Kinship Analysis by the Index of Chromosome Sharing Using High-Density Single Nucleotide Polymorphisms." *PLoS ONE* 11 7 (2016): e0160287, https://doi.org/10.1371/journal.pone.0160287.

Cubeta, Robert, Kristen Bishop, Ashley Farris, Joseph Hamill, Janet Marroquin Pineda, and Jay Shah. *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data with Case Studies*. IDA Paper P-33456 (Alexandria, VA: Institute for Defense Analyses, July 2023), TOP SECRET//NO FORN//SI.

Cubeta, Robert, Kristen Bishop, Janet Marroquin Pineda, Ashley Farris, Clay Hamill, and Jay Shah. *Methodology to Assess Risk from Strategic Competitor Acquisition of U.S. Biological Data and Application to an Agricultural Bioprocessing Case Study*. IDA Paper P-33619 (Alexandria, VA: Institute for Defense Analyses, August 2023).

de Vries, Jard H, Daniel Kling, Athina Vidaki, Pascal Arp, Vivian Kalamara, Michael M.P.J. Verbiest, Danuta Piniewska-Róg, et al. "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy." *Forensic Science International: Genetics* 56 (2022), https://doi.org/10.1016/j.fsigen.2021.102625.

Edge, Michael D., Bridget F. B. Algee-Hewitt, Trevor J. Pemberton, Jun Z. Li, and Noah A. Rosenberg. "Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets." *PNAS* 144 no. 22 (2017): 5671-5676, https://doi.org/10.1073/pnas.1619944114.

Ertürk, Mine Su, Colleen Fitzpatrick, Margaret Press, and Lawrence M. Wein. "Analysis of the Genealogy Process in Forensic Genetic Genealogy." *Journal of Forensic Sciences* 67 (2022): 2218-2229.

Greytak, Ellen M., CeCe Moore, and Steven L. Armentrout, "Genetic Genealogy for Cold Case and Active Investigations." *Forensic Science International* 299 (2019): 108.

Kim, Jaehee, Michael D. Edge, Bridget F.B. Algee-Hewitt, Jun Z. Li, and Noah A. Rosenberg. "Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci." *Cell* 175 (2018): 848-858, https://doi.org/10.1016/j.cell.2018.09.008.

Livescience and Laura Geggel. "23andMe Is Sharing Genetic Data with Drug Giant." *The Scientific American*, July 18 2018, accessed November 2023, https://www.scientificamerican.com/article/23andme-is-sharing-genetic-data-with-drug-giant/.

Living DNA. "What does my relationship prediction mean?" Last updated in 2020. https://support.livingdna.com/hc/en-us/articles/360013536560-What-does-my-relationship-prediction-mean-.

Livingston, Gretchen. *Childlessness Falls, Family Size Grows Among Highly Educated Women* (Washington D.C.: Pew Research Center, 2015), 11.

Molla, Rani. "Why DNA tests are suddenly unpopular." *Vox*, February 13, 2020, accessed August 2, 2023. https://www.vox.com/recode/2020/2/13/21129177/consumer-dna-tests-23andme-ancestry-sales-decline.

National Counterintelligence and Security Center. *Safeguarding Our Future: Protecting Personal Health Data from Foreign Exploitation* (Washington, D.C., 2021): 1.

Regalado, Antonio. "More than 26 million people have taken an at-home genetic ancestry test." *MIT Technology Review,* Last Updated February 11, 2019, accessed August 2, 2023. https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

Thiebes, Scott, Phillipp A. Toussaint, Jaehyeon Ju, Jae-Hyeon Ahn, Kalle Lyytinen, Ali Sunyaev. "Valuable Genomes: Taxonomy and Archetypes of Business Models in Direct-to-Consumer Genetic Testing." *J Med Internet Res* 22, no. 1 (Jan 21 2020). https://doi.org/10.2196/14890, https://www.ncbi.nlm.nih.gov/pubmed/31961329

Thompson, Jim, Tim Clayton, John Cleary, Maurice Gleeson, Debbie Kennett, Michelle Leonard, and Donna Rutherford. "An Empirical Investigation into the Effectiveness of Genetic Genealogy to Identify Individuals in the UK." *Forensic Science International: Genetics* 26 (2020): 102263.

Tripp, Simon, and Martin Grueber. "The Economic Impact and Functional Applications of Human Genetics and Genomics." TEConomy Partners, LLC, May 2021. https://www.ashg.org/wp-content/uploads/2021/05/ASHG-TEConomy-Impact-Report-Final.pdf.

U.S. Department of Defense, Office of the Secretary of Defense. *Direct-to-Consumer Genetic Testing Advisory for Military Members* (Washington, D.C., 2019): 1.

United Nations, Department of Economic and Social Affairs. *World Population Prospects 2022 Summary of Results* (New York, 2022): 5. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

Wickenheiser, Ray. "Investigative Genetic Genealogy: Current Status and Future Potential." *Forensic Science International: Synergy 3* (2021).

Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi. "Identity inference of genomic data using long-range familial searches." *Science* 362 (2018): 2, https://doi.org/10.1126/science.aau4832.

Appendix D. Abbreviations

cM	centimorgans
CODIS	Combined DNA Index System
DNA	deoxyribonucleic acid
DOD	Department of Defense
DTC	direct-to-consumer
FBI	Federal Bureau of Investigation
GSK	GlaxoSmithKline
IDA	Institute for Defense Analyses
LD	Linkage disequilibrium
MTA	Science & Technology Exploitation and Analytics, Maintaining Technology Advantage
NSIB	National Security Innovation Base
SNP	single nucleotide polymorphism
STR	short tandem repeat
US	United States

This page is intentionally blank.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
XX-06-2024	Final	May 2023 - May 2024

4. TITLE AND SUBTITLE

Probability of Identifying a Target from Human Genetic Datasets

5a. CONTRACT NO.

HQ0034-19-D-0001

5b. GRANT NO.

5c. PROGRAM ELEMENT NO(S).

6. AUTHOR(S)

Ashley Farris
Robert Cubeta

5d. PROJECT NO.

5e. TASK NO.

AI-6-5283

5f. WORK UNIT NO.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 E. Glebe Rd Alexandria,
VA 22305

8. PERFORMING ORGANIZATION REPORT NO.

IDA Product 3000645

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OUSD(R&E)
Pentagon, Arlington, VA

10. SPONSOR'S / MONITOR'S ACRONYM(S)

OUSD(R&E)

11. SPONSOR'S / MONITOR'S REPORT NO(S).

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A. Approved for pubic release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Director, Science & Technology Exploitation and Analytics, Maintaining Technology Advantage (MTA), Office of the Under Secretary of Defense, Research & Engineering asked IDA to develop a brief paper describing a characterization of how U.S. genetic data can be used by strategic competitors to re-identify individuals. The executive summary portion of this paper is written to be used by the sponsor and IDA to explain the scope of the privacy and national security risks associated with strategic competitors gaining access to human genetic data on the U.S. population to stakeholders. A detailed methodology section is also included to explain the approach, assumptions, and limitations to the sponsor and stakeholder community.

15. SUBJECT TERMS

genetic data; biological data; genomics; genetic privacy; biometrics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NO. OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Patrick Lee
U	U	U	U	42	19b. TELEPHONE NUMBER (Include Area Code) (703) 571-4028

This page is intentionally blank.