Forensic Science International 299 (2019) 103–113

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Genetic genealogy for cold case and active investigations

Ellen M. Greytak*, CeCe Moore, Steven L. Armentrout

Parabon NanoLabs, Inc., 11260 Roger Bacon Dr. Suite 406, Reston, VA, 20190, USA

ARTICLE INFO

Article history:
Available online 27 March 2019

Keywords:
Genetic genealogy
Forensic genetics
DNA
SNPs
Cold cases
Human identification

ABSTRACT

Investigative genetic genealogy has rapidly emerged as a highly effective tool for using DNA to determine the identity of unknown individuals (unidentified remains or perpetrators), generating identifications in dozens of law enforcement cases, both cold and active. The amount of press coverage of these cases may have given the impression that the analysis is straightforward and the outcome guaranteed once a sample is uploaded to a database. However, the database query results serve only as clues from which in-depth genealogy and descendancy research must proceed to determine the possible identities of an unknown individual. While there certainly will be more announcements of cases solved using this new technique, there are many more cases where identification has not yet been possible due to the wide variety of complications present in these investigations. This paper lays out the fundamentals of genetic genealogy, along with the challenges that are encountered in many of these investigations, and concludes with a set of case studies that demonstrate the variety of cases encountered thus far.

1. Introduction

Traditional genealogy has been practiced for centuries, using documentary records and oral histories to trace families backwards in time. Until recently, these were the only ways to connect extended family members, but with the advent of direct-to-consumer (DTC) genetic testing, it is now possible to find relatives through shared DNA. This has enabled thousands of individuals who have lost their biological identity through adoption, abandonment, anonymous gamete donation, misattributed parentage, etc., to regain their genetic heritage. More recently, these same tools have been used to identify DNA from suspected perpetrators in more than thirty law enforcement cases, only some of which have been publicly announced (Table 1).

2. Generating data

Unlike traditional forensic DNA analysis, which uses autosomal short tandem repeats (STRs) to generate an identity profile from ~20 loci, genetic genealogy uses hundreds of thousands of single nucleotide polymorphisms (SNPs) spread across the autosome. Participants in genetic genealogy have had their DNA tested by a direct-to-consumer (DTC) genetic testing company, such as 23andMe or AncestryDNA, which use microarrays to genotype up to ~1 million SNPs. DTC companies obtain DNA from spit kits or cheek swabs and thus always have a large amount of high-quality single-source DNA to work with. Forensic DNA samples, on the other hand, often only have a small amount of degraded DNA, which may be mixed with DNA from one or more other individuals. Microarray genotyping has previously been shown to be effective and accurate with forensic samples [1], and Parabon has used it for casework since 2015, generating high genotyping call rates from forensic samples down to 1 ng of DNA (Table 2). Parabon has also found it is possible to accurately deconvolute microarray data from two-person mixtures, as long as the person-of-interest is at least 40% of the mixture and a single-source reference sample from the second contributor is available.

Parabon's casework currently uses the Illumina CytoSNP-850K array, an off-the-shelf chip that contains >98% of the SNPs on the OmniExpress chip used by Ancestry.com, FamilyTreeDNA, and MyHeritage. 23andMe previously also based their chip on the OmniExpress but has since moved to smaller custom chips that overlap less with the other DTC companies. For law enforcement cases, extracted DNA samples are processed at a CLIA-certified lab, and the data is uploaded securely to Parabon.

3. Determining relatedness from DNA

Given enough SNPs, it is possible to determine the degree of relatedness between two people, which is defined by the expected amount of shared DNA, not the number of meioses (Fig. 1).

While several relationship inference methods had previously been proposed [2,3], 23andMe was the first DTC company to introduce an accurate, scalable approach to inferring approximately

* Corresponding author.
E-mail address: ellen@parabon.com (E.M. Greytak).

https://doi.org/10.1016/j.forsciint.2019.03.039
0379-0738/

Table 1
Cases for which law enforcement agencies have announced identification of DNA from a suspected perpetrator with the aid of genetic genealogy (through 1/31/19).

	Location	Case	Year(s)	Identified as	Date announced	Genetic genealogist
1	California	Multiple homicides and sexual Assaults —"Golden State Killer"	1974–1986	Joseph James DeAngelo	April 24, 2018	Barbara Rae-Venter
2	Snohomish County, WA	Double homicide of Jay Cook (20) and Tanya Van Cuylenborg (18)	1987	William Earl Talbott II	May 21, 2018	Parabon
3	Tacoma, WA	Homicide of Michella Welch (12)	1986	Gary Charles Hartman	June 20, 2018	Parabon
4	Lancaster, PA	Homicide of Christy Mirack (25)	1992	Raymond Charles Rowe[b]	June 25, 2018	Parabon
5	Brazos County, TX	Homicide of Virginia Freeman (40)	1980	James Otto Earhart[a]	June 25, 2018	Parabon
6	Fort Wayne, IN	Homicide of April Tinsley (8)	1988	John Dale Miller[b]	July 15, 2018	Parabon
7	Woonsocket, RI	Homicide of constance Gauthier (81)	2016	Matthew Norman Dessault	July 18, 2018	Parabon
8	St. George, UT	Sexual Assault of Carla Brooks (79)	2018	Spencer Glen Monnett[b]	July 28, 2018	Parabon
9	Fayetteville, NC	Multiple Sexual Assaults —"Ramsey Street Rapist"	2006–2008	Darold Wayne Bowden	August 22, 2018	Parabon
10	Champaign County, IL	Homicide of Holly Cassano (22)	2009	Michael F. A. Henslick	August 29, 2018	Parabon
11	Montgomery County, MD	Multiple Sexual Assaults	2007–2011	Marlon Michael Alexander	September 14, 2018	Parabon
12	Sarasota, FL	Homicide of Deborah Dalzell (47)	1999	Luke Edward Fleming	September 19, 2018	Parabon and Barbara Rae-Venter
13	California	Multiple Sexual Assaults —"NorCal Rapist"	1991–2006	Roy Waller	September 21, 2018	Law Enforcement
14	Greenville, SC; Memphis, TN; Portageville, MO	Multiple Homicides and Sexual Assaults	1990–1998	Robert Eugene Brashers[a]	October 5, 2018	Parabon
15	Starkville, MS	Double homicide of Betty Jones (65) and Kathryn Crigler (81)	1990	Michael W. DeVaughn	October 8, 2018	Parabon
16	Greenbrier, AR	Homicide of Pam Felkins (32)	1990	Edward Keith Renegar[a]	October 29, 2018	Parabon
17	Fulton County, GA	Homicide of Lorrie Ann Smith (28)	1997	Jerry Lee	November 1, 2018	Parabon
18	Anne Arundel County, MD	Homicide of Michael Temple (29)	2010	Fred Lee Frampton, Jr.	November 2, 2018	Parabon
19	Orlando, FL	Homicide of Christine Franke (25)	2001	Benjamin L. Holmes	November 5, 2018	Parabon & Florida Dept. of Law Enforcement
20	Carlsbad, CA	Homicide of Jodine Serrin (39)	2007	David Mabrito[a]	November 13, 2018	Parabon and Barbara Rae-Venter
21	Santa Clara, CA	Homicide of Leslie Marie Perlov (21)	1973	John Arthur Getreu	November 21, 2018	Parabon
22	College Station, TX	Multiple Sexual Assaults	2018	Christopher Quinn Williams	December 12, 2018	Parabon
23	Cedar Rapids, IA	Homicide of Michelle Martinko (18)	1979	Jerry Lynn Burns	December 19, 2018	Parabon
24	Hernando County, FL	Sexual Assault of Unnamed Victim (12)	1983	William L. Nichols[a]	January 10, 2019	Parabon
25	Orange County, CA	Sexual Assaults of Two Unnamed victims (9 and 31)	1995 & 1998	Kevin Konther	January 11, 2019	Law Enforcement
26	La Mesa, CA	Homicide of Scott Martinez (47)	2006	Zachary Aaron Bunney	January 24, 2019	Parabon
27	Fremont, CA	Homicide of Jack Upton (30)	1990	Russell Guerrero	January 24, 2019	Parabon
28	Portland, OR	Homicide of Anna Marie Hlavka (20)	1979	Jerry Walter McFadden[a]	January 31, 2019	Parabon

[a] Deceased.
[b] Pled guilty.

Table 2
Summary of Parabon's >250 forensic DNA samples used in genetic genealogy casework and the resulting microarray genotyping call rates.

Source		Type		Quantity		Call rate	
Semen	48.0%	Single source	79.4%	≤2.5 ng	22.7%	>95%	47.5%
Blood	24.6%	Low mixture	16.4%	2.5–5 ng	12.6%	90–95%	12.2%
Tissue	10.1%	High mixture (Deconvoluted)	4.2%	5–10 ng	13.0%	80–90%	17.5%
Saliva	7.7%			10–20 ng	17.8%	70–80%	6.1%
Bone	4.8%			20–40 ng	27.1%	60–70%	12.2%
Touch	4.8%			40–80 ng	3.2%	<60%	4.6%
				>80 ng	3.6%		

how closely related two DNA samples are from autosomal SNPs [4]. Each person has two copies of each of the 22 autosomal chromosomes ("autosomes"), one inherited from their mother and one inherited from their father. Autosomes are not inherited intact from each parent; rather, each parent's own pair of chromosomes is randomly recombined into a new chromosome that is passed onto the child. While recombination occurs randomly, nucleotides that are closer to one another on a chromosome are more

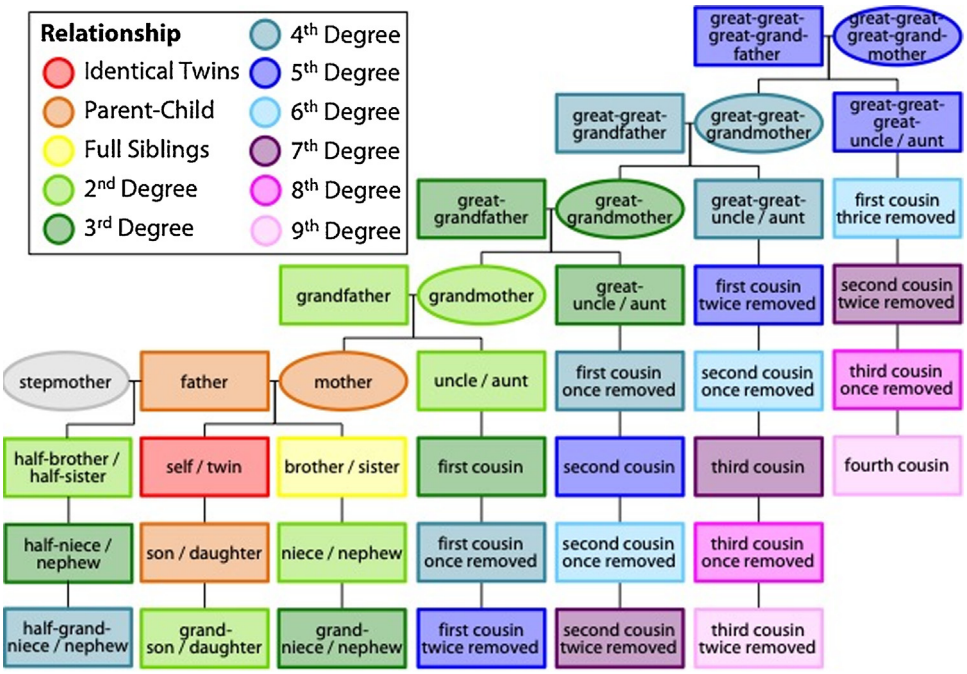

Fig. 1. Pedigree showing the degrees of relatedness, as defined by the expected amount of shared DNA. Each relationship is defined with respect to the red "self/twin" box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

likely to be inherited together, while nucleotides that are far apart are more likely to be separated by recombination. The probability of recombination between two nucleotides is quantified as their genetic distance, which is measured in centimorgans (cM), such that 1 cM equates to a 1% probability of recombination.

Rather than simply looking at the total number of shared SNPs, genetic genealogy takes advantage of the fact that recombination will break up long stretches of shared DNA over the generations, such that more closely related people will share longer stretches of DNA ("segments") that are identical-by-descent (IBD) (Fig. 2). The more recombination events that have occurred, the shorter the shared IBD segments will be, so the number and length of IBD segments in cM can be used to approximate the degree of relatedness.

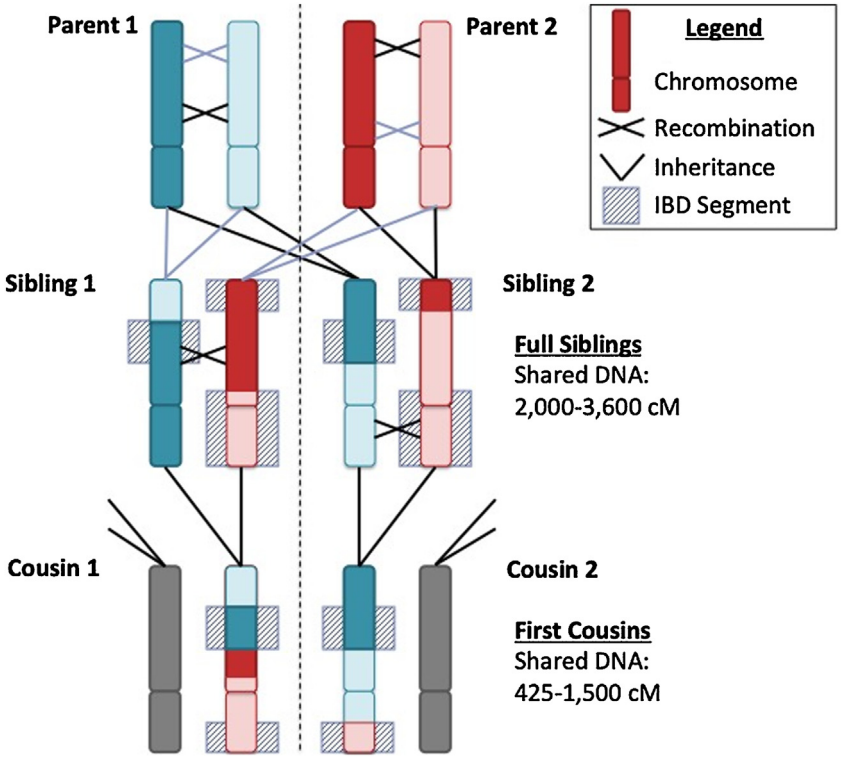

Fig. 2. Inheritance of DNA segments on a single chromosome. The lengths of the shared segments (shaded boxes) are summed across all 22 autosomes to give the total amount of shared DNA.

Table 3
The range of DNA shared by pairs of people with each relationship. While most pairs from a given relationship fall within a narrower range, these values represent the full ranges that have been observed [19].

cM Range	Degree	Relationship
3600	1	Parent–child
2000–3600	1	Full sibling
1060–2500	2	Half-sibling, avuncular, double first cousin, grandparent/grandchild
425–1500	3	First cousin (1C), half-avuncular, great-grandparent/great-grandchild, great-avuncular
160–950	4	First cousin once-removed (1C1R), half-first cousin ($^1/_2$ 1C), half-great-aunt/uncle/half-great-niece/nephew
65–650	5	Second cousin (2C), first cousin twice-removed (1C2R), half-first cousin once-removed ($^1/_2$ 1C1R)
0–375	6	Second cousin once-removed (2C1R), half-second cousin ($^1/_2$ 2C), first cousin thrice-removed (1C3R), half-first cousin twice-removed ($^1/_2$ 1C2R)
0–245	7	Third cousin (3C), second cousin twice-removed (2C2R)
0–185	>7	Third cousin once-removed (3C1R), distant cousins

To detect IBD segments, genetic genealogy algorithms search for regions of the genome where two individuals share at least one allele at every SNP. To be counted, these segments must contain a minimum number of SNPs (typically ~500) and be over a certain length (typically 5–7 cM), which screens out most segments that are shared by chance rather than due to common descent. When summed across all autosomes, the amount of DNA shared IBD strongly correlates with the degree of relatedness between two individuals, such that more distant relatives tend to share less DNA (Table 3). However, due to the random nature of recombination, the amount of shared DNA can vary greatly for relatives of the same degree, and this variation increases with more recombination events, such that ~10% of third cousins and ~50% of fourth cousins share no detectable IBD segments.

4. Genetic genealogy databases and genetic privacy

DTC genetic testing companies' private databases have exploded in size, with AncestryDNA currently containing nearly 15 million individuals, 23andMe containing nearly 10 million, and MyHeritage and FamilyTreeDNA (FTDNA) together containing roughly 3.5 million [5]. AncestryDNA and 23andMe maintain their databases separately and are not accessible to law enforcement, as the only way to submit a sample is via a cheek swab or spit kit. MyHeritage and FTDNA both allow uploads of data generated from other sources, but law enforcement usage of either requires written permission from the company, as well as a court order for MyHeritage or "the required legal documentation" for FTDNA.

GEDmatch, on the other hand, is not a DTC company. It was created by Curtis Rogers and John Olson in 2010 as a public database where individuals from different testing companies could compare their DNA by downloading their raw data from a DTC company's site and uploading it to a common database. After the Golden State Killer suspect was identified through surreptitious use of GEDmatch, the site's administrators decided to explicitly allow law enforcement usage. They posted a notice on the front page of the site (Fig. 3) and also updated their Terms of Service to state that law enforcement can and is using GEDmatch to identify remains and perpetrators of violent crimes, defined as homicides or sexual assaults [6]. Both new and existing users were required to view these new Terms and decide whether to accept them before using the site. Critics of genetic genealogy argue that many people who joined the site prior to this update may not have considered the possibility that their desire to locate relatives could lead to the discovery that they are related to someone whose DNA is associated with a crime and to the apprehension of that relative. Indeed, it is possible some of them still may be unaware of the new warning, and individuals who had their data uploaded by another individual or have been inactive on the site may not have reviewed the new Terms to decide whether to consent. However, even prior to implementing these new Terms, GEDmatch's Terms clearly stated that any data set to "public" would be searchable by anyone. The law has generally allowed information made available to the public to be used in criminal investigations. Users can easily have their data set to "private," hiding it from all search queries, or removed entirely. Thus, the DNA data files in a public database like GEDmatch come from individuals who have proactively downloaded their data from a private DNA testing company's website, uploaded the information to a public website, reviewed the Terms of Service that permits law enforcement usage, and opted in to public comparisons against their data.

Additionally, no sensitive genetic information is disclosed to law enforcement during a genetic genealogy search, as the raw genetic data from GEDmatch users is not accessible. Raw genetic data can contain sensitive health-related information, and this type of private genetic information should be protected. In keeping with this precept, no raw genotypes are displayed or made available for download by GEDmatch. GEDmatch simply performs comparisons among samples, returning the lengths and chromosomal locations of shared DNA segments, which are used to determine the approximate relationship between individuals. Similarly, data obtained from abandoned DNA at a crime scene and

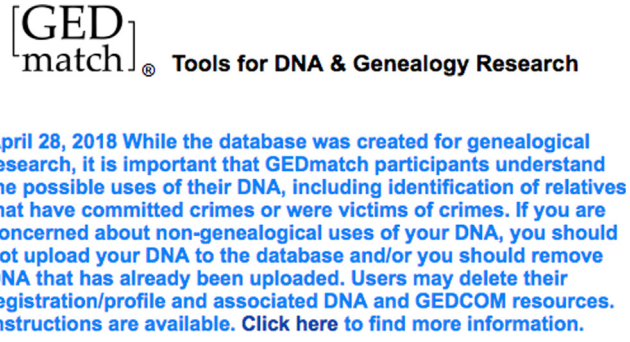

Fig. 3. Notice posted on GEDmatch's homepage after the site's use in the Golden State Killer investigation was made public.

used for genetic genealogy are not exposed to other users and can be prevented from appearing in search results (an option available to all users). At Parabon, genetic data is kept on an encrypted server only accessible to authorized employees, and the company's GEDmatch accounts can only be accessed by the bioinformatics team and the lead genetic genealogist, CeCe Moore. These facts mitigate many of the privacy concerns surrounding genetic genealogy, as individuals have control over whether their data is used as part of law enforcement investigations, and sensitive raw data is not accessed [7,8].

Unlike with familial searching of law enforcement databases, no one is legally required to contribute to a genetic genealogy database, and the samples are not in the possession of government agencies. The persons contributing to GEDmatch are warned explicitly that criminal investigators as well as fellow genealogy enthusiasts are able to perform comparisons against their data. If they choose to participate anyway, there is no reason why law enforcement should not be able to use this information. These significant differences from familial searching argue against automatically applying familial search policies, such as restricting analysis to the end of an investigation, to genetic genealogy. The two techniques are entirely independent; familial searching has previously been performed in some genetic genealogy cases and not in others. The public is strongly in favor of the use of genetic genealogy to investigate violent crimes: GEDmatch saw a significant increase in the number of participants after the Golden State Killer arrest [9], and a recent survey showed overwhelming public support [10].

5. Database searching

A GEDmatch one-to-many query compares the DNA of interest to all public data in the database, returning a list of individuals who share the most autosomal DNA. Each "match" includes the individual's name or alias, the email address associated with their GEDmatch account, and any haplogroup or family tree information they have chosen to share (Fig. 4).

A one-to-one comparison can then be run on each match using a more precise algorithm to see the lengths and chromosomal locations of the shared segments. Comparing the amount of shared DNA to reference data (e.g., Ref. [11]) gives the probability that the relationship between the unknown individual and the match falls into each degree of relatedness. For example, a match sharing 100 cM could be anywhere from 5th degree to >8th degree, with 6th degree being most likely.

However, there are additional complications. First, in addition to multiple possible degrees of relatedness, each degree contains many relationship types that must be considered (e.g., 5th degree relatives around the same age could be second cousins, first cousins twice-removed, or half-first cousins once-removed). Second, the amount of DNA shared by each relationship varies among populations. Populations founded by a small number of individuals can have low genetic diversity and high background relatedness, or *endogamy*. In such populations, individuals with a given relationship will share significantly more DNA than in other populations, such that even very distant cousins can share significant amounts of DNA. Endogamy manifests as a large number of matches, each sharing many small segments, indicating that the segments were actually inherited from distant ancestors [12]. Another challenge is *pedigree collapse*, in which the same families intermarry multiple times throughout history, which can inflate the amount of shared DNA between their descendants.

6. Casework match results

More than 80% of samples from Parabon's >250 law enforcement cases have resulted in a match at the third cousin level or closer (>60 cM), with subjects of European descent having a higher probability of success due to their overrepresentation in genetic genealogy databases [13] (Fig. 5A). European descent was assessed by Snapshot DNA Phenotyping, which infers an individual's genetic admixture from seven continental populations (African, Middle Eastern, European, Central/South Asian, East Asian, Oceanian, and Native American). In this analysis, samples were considered "European" if they had at least 80% European ancestry. Note that the law enforcement cases submitted to Parabon are primarily from North American agencies, and samples from other regions will likely have lower match probabilities due to lower participation in DTC genetic testing and use of GEDmatch.

The closeness of the top match is not the sole variable in determining viability for genetic genealogy. A comprehensive assessment must include consideration not only of the closest match, but of the quality of the supporting matches and the amount of information available about each match. For example, progress may be difficult if the top match has unknown parentage and/or is from a country where records are not available. Parabon assesses each sample on a subjective scale: (1) very high probability of identification (e.g., parent–child match), (2) high probability of identification, (3) medium probability of identification, (4) low probability of identification but likely to generate actionable information, and (5) unlikely to generate actionable information. An assessment does not guarantee a particular outcome but is intended to help agencies to decide how to proceed. Thus far, ~80% of European samples and ~60% of non-European samples have been assessed as workable (assessments 1–4) (Fig. 5B).

Importantly, just because a sample does not have sufficient promising match data today does not mean it never will. Hundreds of new individuals upload their data to GEDmatch every day [9], and as the database grows, the proportion of samples with close matches will increase. Thus, Parabon monitors all unsolved cases for new matches on a weekly basis.

7. Genealogy research

While most of the discussion surrounding genetic genealogy focuses on the database matches, the vast majority of genetic genealogy work happens after the match list is generated. Many US

					Haplogroup		Autosomal				X-DNA		
Type	List	Select	Sex	GED/WikiTree	Mt	Y	Details	Total cM	largest cM	Gen	Details	Total cM	largest cM
					▼ ▲	▼ ▲		▼	▼	▼ ▲		▼	▼
F2	L	☐	F				A	85.9	44.4	3.7	X	0	0
V4	L	☐	F		T2b4		A	72	44.4	3.8	X	0	0
F2	L	☐	F				A	66.3	58.2	3.9	X	10.5	5.3
F2	L	☐	F	GED			A	62.9	14.5	3.9	X	7.1	7.1
F2	L	☐	U			I1	A	59	51.3	4.0	X	0	0

Fig. 4. Top five results from a GEDmatch one-to-many comparison, with potentially identifying information (kit numbers, names, and email addresses) removed.

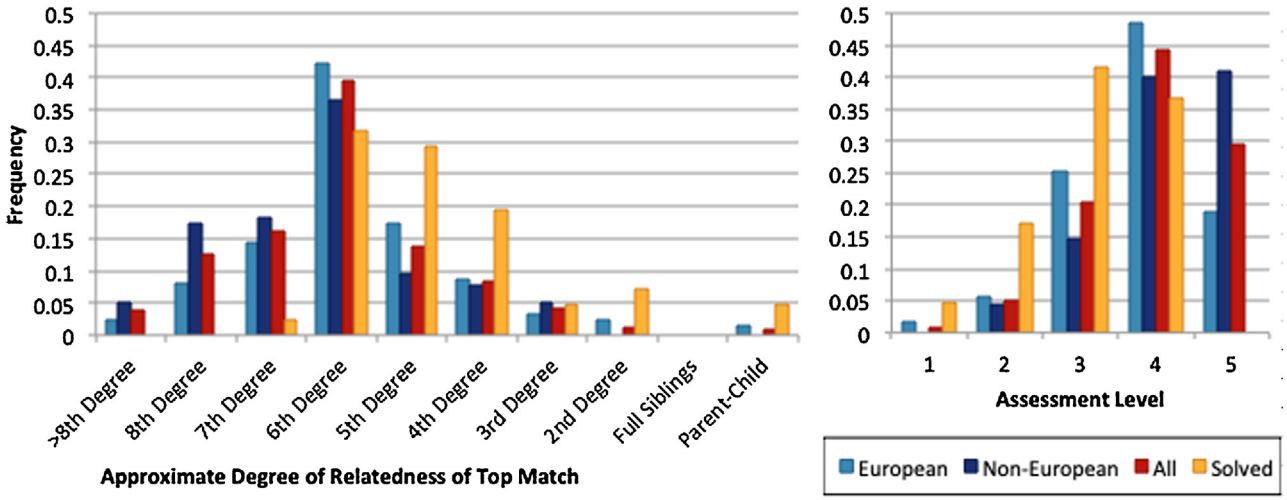

Fig. 5. For Parabon's >250 law enforcement samples, the frequency of (A) the top GEDmatch one-to-many match being in each degree of relatedness and (B) samples receiving each assessment level. Results are reported for European, non-European, and all samples, as well as for those cases that have been solved (i.e., resulted in an identification) thus far. Degree of relatedness is based solely on the amount of shared DNA, not the true relationship determined through genealogy: parent–child (>3300 cM), full siblings (2200–3300), 2nd degree (1300–2200), 3rd degree (650–1300), 4th degree (340–650), 5th degree (200–340), 6th degree (90–200), 7th degree (60–90), 8th degree (30–60), >8th degree (<30).

records are available to the public and have been compiled into searchable databases accessible via subscription. For example, Ancestry.com provides a mechanism for accessing a large collection of records, such as the census through 1940, vital records (birth, marriage, death) from many states, the Social Security Death Index, and Newspapers.com. Some Ancestry.com users also create and share public family trees, although these can contain errors, so they must be examined critically. People search databases and public social media can also be used to help determine family structures. In some cases, law enforcement may be asked to assist with this research using their greater access to records.

A previous analysis of the MyHeritage DTC database showed that ~60% of individuals of Northern European descent will have a match at 100 cM or closer [14]. Using simulation, the authors showed that it is often possible to identify an unknown individual from a single third cousin level match given knowledge of his or her sex, location within 100 miles, and age within 5 years. However, in addition to the fact that such detailed demographic information is often not available in law enforcement cases, this assumes that, given a third cousin match, it is straightforward to obtain a complete list of the match's relatives at that distance (the authors determined this number to be ~850, not including half relatives). In reality, a massive amount of work is required to expand a match into a list of relatives [7,8].

The first task is to definitively identify each match, which itself can be quite difficult. Although GEDmatch displays the name and email address associated with each matching kit, users can choose to use an alias or an anonymous email address, and kits are sometimes managed by someone other than the match themselves. Moreover, even if a user associates their actual name, it may be common (e.g., John Smith), which can complicate identification. Consequently, the initial identification of matches is both critical and challenging, and often requires considerable genetic genealogical skill and creative problem solving, e.g., deciphering initials, inferring identities from other identifiable matches, and figuring out who DNA is from when the kit is managed by someone else. Even though contacting matches via the given email address might enable identification and even produce family tree information, Parabon seldom contacts matches directly so as to minimize the number of people involved in an investigation and reduce the risk of tipping off a suspect. Matches closer than third cousins are only contacted with the permission of the investigating agency, and the agency can choose to make the contact instead. Any contact includes the fact that the questions are in regard to a law enforcement investigation (no specifics of the case are given), and the individual is informed they are free to participate or not. If the individual asks not to be involved, they are not contacted again.

Once the matches are identified, their family trees must be constructed back to the set of possible common ancestors with the unknown individual. The number of generations back in time to the common ancestors of interest is determined by the distance of the matches' relationships, although since the estimates are not usually specific to a single relationship, often the family trees must be built even further back than these levels would imply. Building family trees back in time requires traditional genealogy research: combing through public records to determine the identities of each generation's parents.

However, records are not always available – not all US states maintain an accurate and public birth index, many families trace back to immigrants from other countries where records are not readily available, etc. In addition, biological family trees often do not match documented family trees due to misattributed paternity, unrecorded adoption, unknown parentage, etc., and individuals in these situations are overrepresented in genetic genealogy databases. Surnames and spellings also often change through the generations, further complicating the analysis.

8. Descendancy research

Once possible common ancestors have been identified, the family trees must then be built forward in time ("descendancy research" or "reverse genealogy") to elucidate the possible identities of the unknown individual (Fig. 6).

The possible ancestors from which the unknown individual descends can sometimes be narrowed using genomic ancestry (e.g., if the family tree is Northern European, but the unknown individual has 25% ancestry from another population, the genetic genealogist can search among the possible grandparents for one who married someone from that ancestral group). Shared DNA on the X-chromosome can also narrow down the possible paths

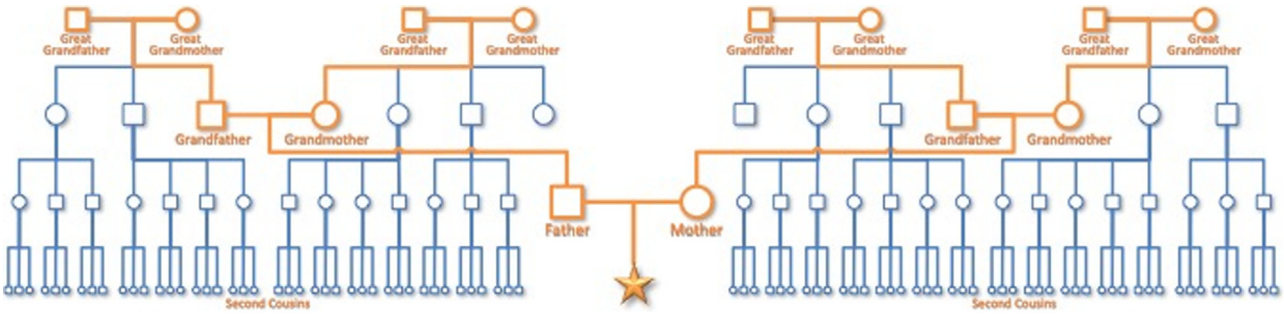

Fig. 6. A hypothetical family tree resulting from genetic genealogy research. Given a match in GEDmatch (orange star), the family tree is built backward in time to the possible common ancestors (orange) and then forward in time (blue) to determine the possible identities of the unknown individual (in this case, from among the "second cousins"). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

between matches, as males only inherit X-DNA from their mothers. Thus, if an unknown male shares X-DNA with a match, they must be related through his mother, and the path between them cannot pass through two males in a row. When available, Y-chromosome and mitochondrial (mtDNA) haplogroups can also narrow down the possibilities, as these are passed directly from father to son and from mother to child, respectively. Thus, individuals share a mtDNA haplogroup with their maternal lineage, and males share a Y haplogroup with their paternal lineage.

DNA sharing among matches can also be used to narrow down where the unknown individual falls in the tree. If matches do not share any DNA with one another, they are likely related to the individual on different branches of his or her family tree, and the genetic genealogist can then search for an intersection ("triangulation") between the two matches' families in the form of a marriage that produced children or an out-of-wedlock birth (Fig. 7). While there could be hundreds or thousands of individuals who are second or third cousins to a single match, there are typically only a few individuals who are cousins at the right distance to multiple matches.

9. Narrowing down the possible identities

Once candidate individuals have been identified, the genetic genealogist can use a variety of factors to include or exclude them, in addition to traditional investigative information, such as a connection to the crime scene or the victim. Sex is known from the DNA, and some age information may be available – for unidentified remains, age can be estimated; for perpetrators, at minimum, they had to be alive and physically capable of committing the crime. The individual also had to be in a given location at a given time, which may mean he or she lived nearby. While the GEDmatch matches may be spread across the US or even the world, it is sometimes possible to focus on a particular branch of the family that moved close to the location of interest.

Parabon's genetic genealogists also use Snapshot DNA Phenotyping [15] to prioritize among individuals and confirm or exclude hypotheses. An individual's eye color, hair color, and skin color can often be determined from mugshots, yearbook photos, or social media and compared to the predictions. Full siblings cannot be distinguished using genetic genealogy, as they share all the same genealogical relationships with the matches. However, if they differ in phenotype, this can be used to prioritize among them. Similarly, if genealogy research leads to an individual whose phenotypes are at odds with the predictions, this can spur continued research, while a close similarity can help corroborate an identification.

The degree to which the identity of the unknown individual can be narrowed down varies from case to case. In the best-case scenario, a single individual or a set of siblings can confidently be identified through matches to multiple branches of their family tree. More often, there are multiple cousins (descendants of a particular set of common ancestors) who are consistent with the available information. These leads can then be followed up through additional research, traditional investigation, and/or targeted kinship testing of family members to more precisely place the unknown individual in the family tree. Parabon's Snapshot Kinship Inference tool uses genome-wide SNP data to predict the precise degree of relatedness between individuals, out to 6th-degree relatives [16]. Using a machine learning model built on thousands of reference subjects with known relationships, Snapshot predicts the probability that a pair belongs to each degree of relatedness. Confidence is calculated using the probability of the most likely degree and the precision calculated for that degree in cross-validation.

10. Law enforcement leads

During decades-long cold case investigations, hundreds or thousands of individuals may be investigated before the perpetrator is found. Genetic genealogy offers an efficient means of narrowing an investigation, often to only a few individuals. The number of possible relatives included in a genetic genealogy analysis varies depending on the number and distance of the matches. Even when the only matches are distant and large family trees must be constructed because common ancestors are many generations in the past, experienced genetic genealogists can triangulate among the matches to determine the most promising

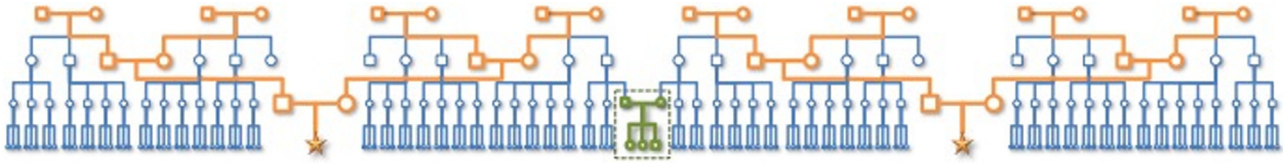

Fig. 7. Triangulation between two hypothetical family trees. Given two matches in GEDmatch who are unrelated to one another (orange stars), family trees are built for each and then searched for an intersection (green) in the form of a marriage or out-of-wedlock birth. Children of this intersection are related to both matches, while all other individuals in the tree are only related to one match. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

branches of the family tree and limit the amount of unnecessary tree building. Given sufficient triangulation and time, the number of leads can be reduced to the offspring of a single couple.

No matter how confident the identification, however, genetic genealogy alone cannot prove identity with 100% certainty. There is always a remote possibility that the unknown individual could have been adopted or abandoned, and his or her existence could be unknown to family and not revealed through official records. Therefore, genetic genealogy leads must be verified through a direct DNA comparison between the person-of-interest's STR profile and that of the crime scene sample. It is this traditional forensic DNA match that is used for prosecution.

11. Case studies

The following case studies demonstrate how genetic genealogy has been used to assist investigators with identifying a suspect in cold case investigations. Only information approved for public release by the investigating agencies is included, so some case details (e.g., DNA sample source, exact GEDmatch match information) have been obfuscated.

11.1. Case study #1: Snohomish County, WA; 31-year-old cold case (double homicide)

This case study demonstrates the ideal genetic genealogy case, where there are close matches and clear familial connections that point to only a single conclusion. However, even seemingly straightforward cases require a large amount of research and the expertise to recognize and cope with confounding factors such as unknown and misattributed parentage.

11.1.1. The crime

In 1987, a young Canadian couple, Jay Cook (20) and Tanya Van Cuylenborg (18), traveled from British Columbia to Washington State in a van. After purchasing a ferry ticket to Seattle, they were never heard from again. Days later, Tanya's body was found in a ditch in the woods, and a few days after that, Jay's body and the van were found in two separate locations. DNA evidence was obtained for an unknown suspect ("Subject").

11.1.2. GEDmatch

There were two matches at approximately the 5th degree relative level, plus additional more distant matches. The top two matches had no shared DNA between them, meaning they were most likely related to the Subject on different branches of his family tree.

11.1.3. Family trees

Family trees were constructed for both key matches back to their great-grandparents and beyond using census records, vital records, newspaper archives, public "people search" databases, public social media data, and public family trees. Next, descendancy research was performed to trace the descendants of each set of ancestors to determine if an intersection between them could be found.

A triangulating marriage was found between a granddaughter of Match #2's great-grandparents and a son of Match #1's great-grandmother. Extensive research revealed that this son had taken his stepfather's surname, initially obscuring his true relationship to Match #1. Thus, the children of this marriage were half first cousins once-removed to Match #1, as well as second cousins to Match #2. While both of these relationships are 5th degree, it is critical to consider all possible relationship types, as half relationships are quite common. No other marriages were found between the descendants of these ancestors. There was only one son from this marriage, William Earl Talbott II, and he was therefore the only known male who could be carrying this mix of DNA from both matches' families (Fig. 8).

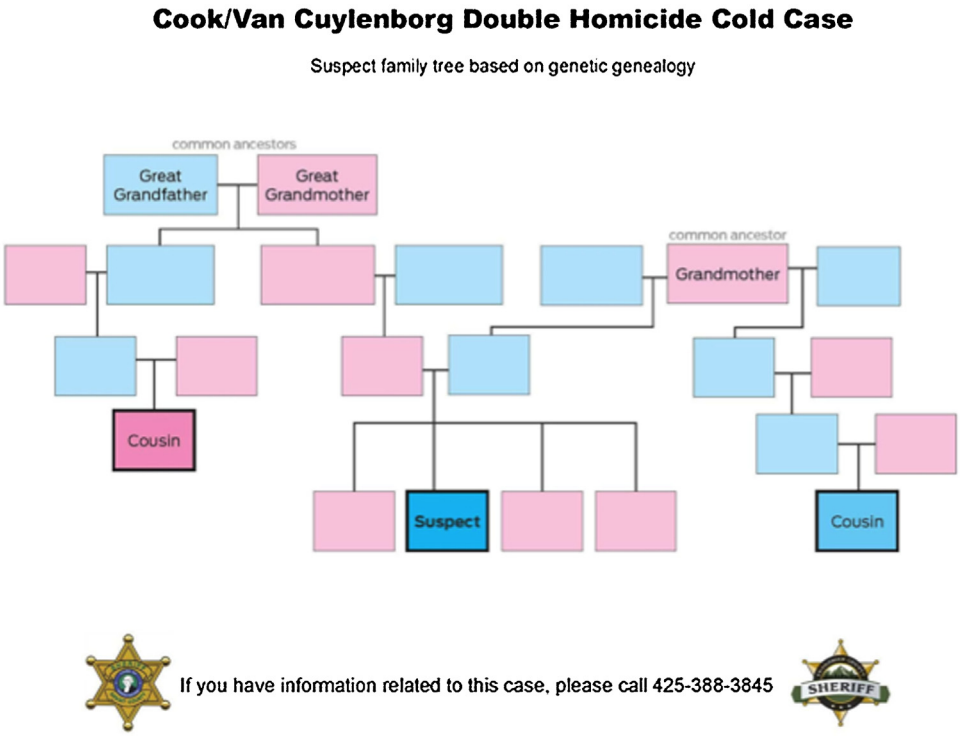

Fig. 8. Anonymized family tree released by the Snohomish County Sheriff's Department as part of their announcement of the arrest of William Earl Talbott II. The tree shows the position of Mr. Talbott (Suspect) and two GEDmatch matches (Cousins) used to determine his identity.

Mr. Talbott had never been arrested for a crime that would require submitting DNA to a database. He had no known connection to the victims and no reason to have been on the investigators' radar. His phenotypes matched those predicted by Snapshot, but without other information to tie him to the crime, this had not been enough to identify him as a suspect.

11.1.4. Resolution

Based on the lead provided by genetic genealogy, the detectives were able to collect DNA from a cup discarded by Mr. Talbott, which, using traditional STR analysis, was shown to match the DNA from the crime scene. He was arrested and is currently awaiting trial.

11.2. Case study #2: Tacoma, WA; 32-year-old cold case (homicide)

Triangulation between matches using documentary sources is sometimes not possible. In addition to being able to tenaciously research records and meticulously build family trees, this case study shows how genetic genealogists must be able to think creatively about possible hypotheses to explain the available data.

11.2.1. The crime

12-year old Michella Welch went missing on 26 March 1986. She had taken her two younger sisters to Puget Park in Tacoma, Washington and then ridden her bicycle home to make lunch while her sisters played nearby. When the sisters returned to the park, they found a brown paper bag with their lunches but no Michella. By 3:10 p.m., officers arrived at the park and started searching for the missing girl. A tracking dog found her body around 11:30 p.m. She had been beaten and sexually assaulted and died from a cut to the neck.

11.2.2. The DNA

Another young Tacoma girl, Jennifer Bastian, was also killed around the same time, and investigators had long believed one person committed both crimes. More than 10,000 investigative hours went into the cases in 1986 alone. Recent DNA testing showed that the crimes were committed by different men, but neither DNA profile resulted in a CODIS match.

11.2.3. Genetic ancestry

The Subject was predicted to be predominantly Northern European with a small but notable amount of Northern Native American admixture (10%).

11.2.4. GEDmatch

The two top matches did not share DNA, suggesting they were most likely related to the Subject on different branches of his family tree.

11.2.5. Family trees

Trees were built for the two top matches back to their great-great-grandparents and beyond, and extensive descendancy research was performed, but no documented intersection was found between the two families. The analyst identified a pair of brothers who were cousins of Match #1, lived within a few miles of the crime scene in 1986, and had two Native American great-great-grandparents on different branches of their family trees, which was consistent with the predicted ancestry of the Subject. However, the Subject only shared about half as much DNA with Match #1 as would be expected for a cousin, and there should have been an intersection between the families that would connect these cousins to both matches.

When families are connected through DNA but do not intersect on paper (e.g., through a marriage license or a birth certificate), the explanation may be misattributed paternity: a pair of individuals from each family had a child together, but the true biological father was not recorded. Through census record research, it was discovered that relatives of the two matches had lived in the same small town when one of the cousins' ancestors was conceived. This was the only discovered geographical intersection between these families. Based on the amount of shared DNA, it was postulated that Match #2's relative was the unrecorded biological father of the cousins' ancestor (Fig. 9). Under this hypothesis, the cousins would actually be half cousins to Match #1, which matched the amount of shared DNA. They would also be related to Match #2 at the appropriate genetic distance.

11.2.6. Resolution

The genetic genealogy analysis identified a pair of brothers who could be the Subject, neither of whom had ever been arrested for a

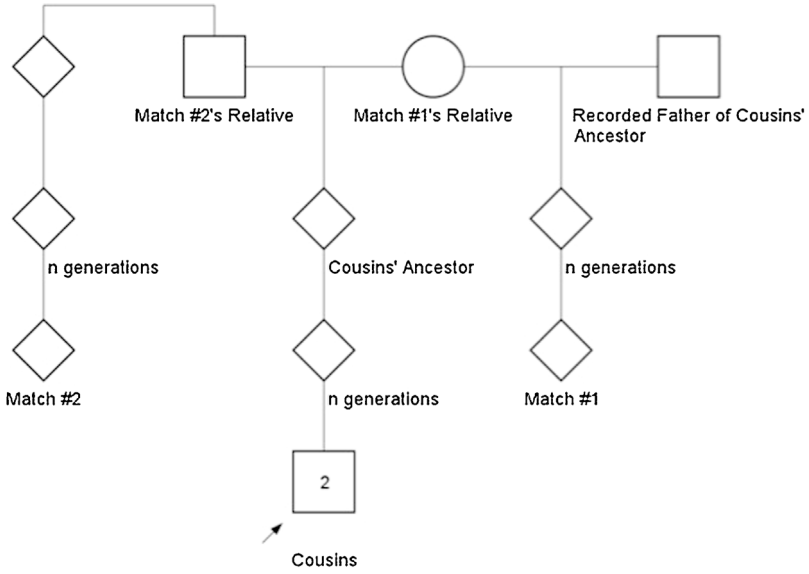

Fig. 9. Pedigree for two cousins of Match #1 who were identified as persons-of-interest in the Tacoma case, showing the apparent misattributed paternity between Match #1's relative and Match #2's relative.

crime that would have required submission of DNA to a database. Officers were eventually able to follow one of the brothers, Gary Charles Hartman, into a restaurant, where they obtained a napkin he had used and discarded. Traditional STR analysis showed that the DNA on the napkin matched the DNA found at the crime scene. More than thirty years after Michella Welch was found murdered in a Washington park, investigators announced that they had arrested a suspect in her murder. Hartman is currently awaiting trial.

11.3. Case study #3: nearly 40-year-old cold case (homicide)

When there are not enough strong matches in GEDmatch to fully narrow down the possible branches of a large family tree, cases cannot always be resolved efficiently through genetic genealogy alone. If an intersection between the matches' families cannot be found, the number of possible identities for the Subject can be very large. However, as this case study shows, if family members of the matches are willing to cooperate, targeted kinship testing can quickly include or exclude various branches of the family tree and thus arrive at a small number of included individuals. Due to the close relatives of the suspect who were eventually found in this investigation, the details of this case are not included to protect their privacy.

11.3.1. GEDmatch

The Subject's top two matches were both in the 6th–8th degree relative range and had no shared DNA between them, meaning they were most likely related to the Subject on different branches of his family tree. There were also additional, more distant matches.

11.3.2. Family trees

Trees were built for the two top matches back to their great-great-grandparents, but no intersection was found between the two families. The Subject was most likely a great-grandson or great-great-grandson of one of Match #1's great-great-grandparent couples, but without triangulation, it was not possible to narrow his identity down further. Parabon recommended more research to identify branches of the family that might have moved to the area of the crime, as well as targeted kinship testing of members of the top match's family.

11.3.3. Kinship testing

The investigating agency obtained a voluntary buccal swab from a cousin on Match #1's paternal side, from which DNA was extracted, genotyped, and compared to the Subject. Snapshot Kinship Inference predicted this individual was unrelated to the Subject, and Match #1's paternal family could therefore likely be excluded (assuming the familial relationships on paper were correct). The agency then obtained a voluntary buccal swab from a cousin on Match #1's maternal side, who was predicted with 94.2% confidence to be a 3rd degree relative (first cousin or genetic equivalent) to the Subject.

11.3.4. Targeted family trees

The analyst built family trees for the spouses of each of the kinship tester's maternal aunts and uncles back to their great-great-great-grandparents. One uncle's wife was determined to be a distant cousin to many of the Subject's more distant matches. This triangulation meant that one of the male children of this couple was most likely the Subject, as he would be related to the GEDmatch matches on both sides of his family tree – second cousins once-removed (6th degree relatives) to Match #1 and distant cousins (ranging from third cousins once-removed to fifth cousins once-removed) to Distant Matches #1–7 (Fig. 10).

Importantly, barring additional independent intersections between these family trees, the identified Persons of Interest were the only individuals who were related to both of these families. These children were also the right age at the time of the crime, lived nearby, and all appeared to have phenotypes consistent with the Snapshot predictions.

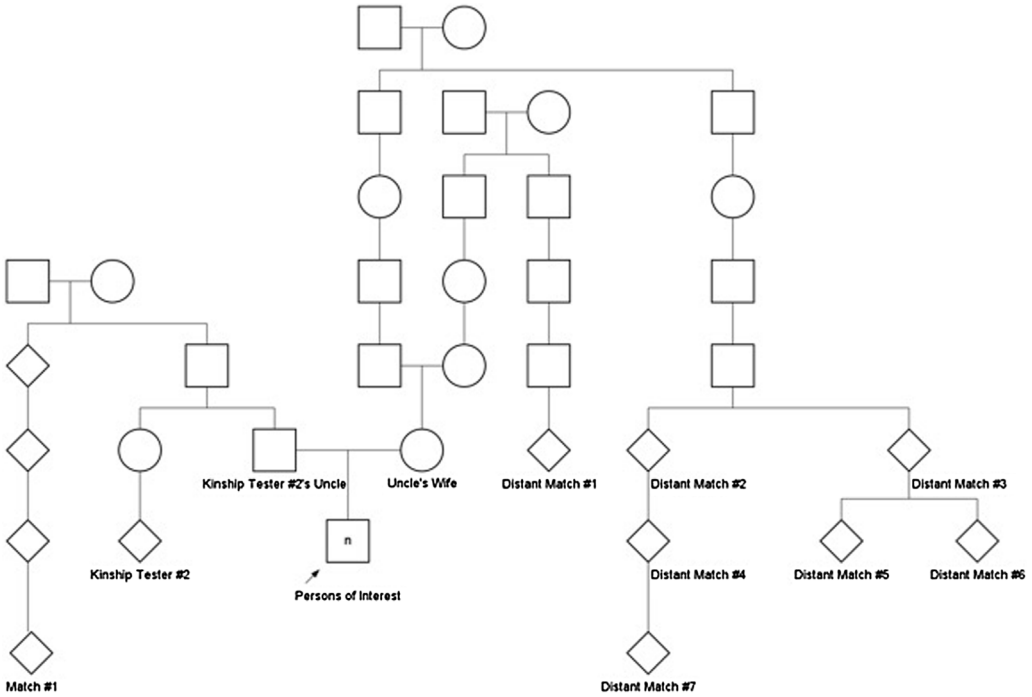

Fig. 10. Pedigree built for Match #1's family after the possible branches leading to the Subject were narrowed down through targeted kinship testing and subsequent triangulation with distant matches.

11.3.5. Resolution

The genetic genealogy analysis identified a set of brothers who could be the Subject, none of whom had ever been arrested for a crime that would have required submission of DNA to a database. Officers were eventually able to narrow the investigation down to a single brother and match his DNA to the crime scene DNA using traditional STR analysis. He has been arrested and is awaiting trial.

12. Conclusions

Genetic genealogy has been called "2018's biggest contribution to crime science" [17] and is rapidly changing the face of cold case investigations. Even for perpetrators who are completely under the radar or long dead, given DNA from a crime scene, it may be possible to identify them with genetic genealogy. Importantly, genetic genealogy has just as much power to generate leads in active cases as in cold cases. In fact, it was recently used to identify a perpetrator in a sexual assault case that had occurred only three months earlier [18], and he has since pled guilty. Rather than wait until years have passed and all other leads have been exhausted, investigators now have access to innovative forensic DNA technologies that can generate significant new leads and prevent cases from going cold. Looking to the future, genetic genealogy has the potential to significantly reduce the number of unsolved cold cases in North America while also reducing the rate at which cases go cold.

Conflict of interest

The authors are employees of Parabon NanoLabs, Inc., which provides genetic genealogy services to law enforcement.

CRediT authorship contribution statement

Ellen M. Greytak: Conceptualization, Software, Validation, Formal analysis, Writing - original draft, Visualization. **CeCe Moore**: Conceptualization, Methodology, Investigation, Writing - review & editing. **Steven L. Armentrout**: Conceptualization, Writing - review & editing, Supervision, Project administration.

References

[1] B. Keating, A.T. Bansal, S. Walsh, J. Millman, J. Newman, K. Kidd, M. Kayser, First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip, Int. J. Leg. Med. 127 (2013) 559–572, doi:http://dx.doi.org/10.1007/s00414-012-0788-1.

[2] C.D. Huff, D.J. Witherspoon, T.S. Simonson, J. Xing, W.S. Watkins, Y. Zhang, T.M. Tuohy, D.W. Neklason, R.W. Burt, S.L. Guthery, S.R. Woodward, L.B. Jorde, Maximum-likelihood estimation of recent shared ancestry (ERSA), Genome Res. 21 (2011) 768–774, doi:http://dx.doi.org/10.1101/gr.115972.110.

[3] A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies, Bioinformatics 26 (2010) 2867–2873, doi:http://dx.doi.org/10.1093/bioinformatics/btq559.

[4] B.M. Henn, L. Hon, J.M. Macpherson, N. Eriksson, S. Saxonov, I. Pe'er, J.L. Mountain, Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples, PLoS One 7 (2012), doi:http://dx.doi.org/10.1371/journal.pone.0034267.

[5] A. Regalado, More than 26 Million People Have Taken an at-home Ancestry Test, MIT Technology Review, 2019 Retrieved from https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

[6] GEDmatch.com. Terms of Service and Privacy Policy.

[7] E.M. Greytak, D.H. Kaye, B. Budowle, C. Moore, S.L. Armentrout, Privacy and genetic genealogy data, Science 361 (6405) (2018) 857, doi:http://dx.doi.org/10.1126/science.aav0330.

[8] E.M. Greytak, C. Moore, S.L. Armentrout, RE: identity inference of genomic data using long-range familial searches, Erlich et al, Science 362 (6415) (2018) 690–694 (eLetter, 10-29-18).

[9] J. Milian, Cold-case Murders, Rapes Cracked by Lake Worth Genealogy Website, The Palm Beach Post, 2018 Retrieved from https://www.palmbeachpost.com/news/20181129/cold-case-murders-rapes-cracked-by-lake-worth-genealogy-website.

[10] C.J. Guerrini, J.O. Robinson, D. Petersen, A.L. McGuire, Should police have access to genetic genealogy databases? Capturing the Golden State Killer and other criminals using a controversial new forensic technique, PLoS Biol. 16 (10) (2018)e2006906, doi:http://dx.doi.org/10.1371/journal.pbio.2006906.

[11] B.T. Bettinger, J. Perl, The Shared cM Project 3.0 Tool v4, (2018) Retrieved from https://dnapainter.com/tools/sharedcmv4.

[12] International Society of Genetic Genealogy, Endogamy Retrieved from https://isogg.org/wiki/Endogamy. (Accessed 30 January 2019), (2019) .

[13] E. Greytak, C. Moore, Closing cases with a single SNP array: integrated genetic genealogy, DNA phenotyping, and kinship analyses, Proceedings of the 29th International Symposium on Human Identification, (2018) .

[14] Y. Erlich, T. Shor, I. Pe, S. Carmi, Identity inference of genomic data using long-range familial searches, Science 362 (6415) (2018) 690–694, doi:http://dx.doi.org/10.1126/science.aau4832.

[15] E.M. Greytak, S. Armentrout, DNA phenotyping: predicting ancestry and physical appearance from forensic DNA, Proceedings of the 26th International Symposium on Human Identification, (2015) .

[16] E.M. Greytak, E.M. Gorden, C.K. Marshall, K. Sturk-Andreaggi, T.P. McMahon, S. L. Armentrout, SNP Recovery from Degraded Samples for Kinship Assessment, (2017) .

[17] S. Augenstein, Working Backward from Genealogy: Tracking a Dead Killer's Trail, Forensic Magazine, 2018.

[18] E. Havens, Elderly Woman in Home Invasion Rape Case: I Forgive My Attacker, St. George Spectrum & Daily News, 2019 Retrieved from https://www.thespectrum.com/story/news/2019/02/26/elderly-woman-home-invasion-rape-case-forgive-my-attacker/2995143002/.

[19] C. Ball, M. Barber, J. Byrnes, P. Carbonetto, K. Chahine, R. Curtis, J. Granka, E. Han, E. Hong, A. Kermany, N. Myres, K. Noto, J. Qi, K. Rand, Y. Wang, L. Willmore, Ancestry DNA Matching White Paper, (2016) Retrieved from https://www.ancestry.com/corporate/sites/default/files/AncestryDNA-Matching-White-Paper.pdf.