Forensic Science International: Genetics

An empirical investigation into the effectiveness of genetic genealogy to identify individuals in the UK

--Manuscript Draft--

Manuscript Number:	FSIGEN_2019_449R1
Article Type:	Research Paper
Section/Category:	Regular Paper
Keywords:	Forensic Genealogy; Genetic Genealogy; SNP; GEDmatch
Corresponding Author:	Jim Thomson Eurofins Forensic Services Teddington, United Kingdom
First Author:	Jim Thomson
Order of Authors:	Jim Thomson
	Tim Clayton
	John Cleary
	Maurice Gleeson
	Debbie Kennett
	Michelle Leonard
	Donna Rutherford
Abstract:	The use of genetic genealogy techniques to identify Joseph James DeAngelo as the prime suspect in the Golden State Killer case in 2018 has opened up a new approach to investigation of cold cases. Since that breakthrough, genetic genealogy methods have been reported to be applied to around 100 cases. To date, all of these reports relate to investigations in the US, where the high uptake of “direct-to-consumer” (DTC) genetic testing by individuals conducting private ancestral research has provided the necessary publicly available data for successful forensic investigations. We have conducted a study to assess the likely effectiveness of genetic genealogy techniques if applied to investigations in the UK. Ten volunteers provided their own SNP array data, downloaded from a DTC provider of their choice. These data sets were anonymised and uploaded to the GEDmatch Genesis genealogy website, mimicking data sets from unsourced crime samples or unidentified human remains. A team of experienced genealogists then attempted to identify the donors of the anonymised data sets by working with matches on the database and identifying points where the matches’ trees intersect to determine their shared family lineages which were further investigated using traditional resources (such as birth, marriage, death and census records, social media and online family trees). Through these methods, four of the ten donors were identified, at least to the level of one of a set of siblings. This confirms that, despite the over-representation of US citizens on publicly accessible genealogy databases, there is still potential for effective use in investigations outside the US where legislation permits. One of our four identified individuals was of Indian heritage (via St Vincent and the Grenadines) highlighting that in the right circumstances individuals of non-European origin can be identified.
Suggested Reviewers:	
Opposed Reviewers:	
Response to Reviewers:	

Author responses to Reviewer Comments for FSIGEN_2019_449 An empirical investigation into the effectiveness of genetic genealogy to identify individuals in the UK

Reviewer #1: This paper details a useful investigation of the relative usefulness of genetic genealogy to identify unknown individuals from high quality samples. A similar study was also undertaken and published in Buzzfeed News which the authors may or may not have been aware of, but this has not been referenced. Although the focus of this study was about the technique's usefulness in the UK, the outcome was similar.

Response: Description and reference for this (and reference to another journalistic study by Kristen Brown article) added at lines 76-81

While the authors clearly recognised potential ethical and governance issues that would need to be addressed before such a process could be widely accepted, they have not mentioned other factors that can influence the use of the technique, such as the quality and quantity of available material for analysis, potentially limited to sexual assaults with ejaculation or unidentified deceased from which an appropriate amount of material can be recovered, along with individuals with the relevant genealogical skills

Response: Paragraph to comment on these practicalities added to Introduction at lines 44-55

The authors have utilised the GEDmatch Genesis tools and it would be helpful if there was a short summary of these features in the text and, given the recent website changes to require an opt-in, as well as the purchase of the site by Verogen, it would be helpful to add something about the current size of the database accessible by law enforcement and any potential changes that may have been revealed

Response: New paragraph describing relevant GEDmatch tools added to Materials and Methods section at lines 142-154

Response: Discussion updated to include updated figures for GEDmatch opt-in (line 464), details of GEDmatch acquisition (lines 468-474) and reference to Othram DNA Solves database for law enforcement investigations (lines 481-484)

More importantly, can the authors quantify the number of genealogy hours employed for each of the cases, or at least a median and range for the set?

Response: ***A paragraph detailing estimated time spent by the genealogy team on the investigations has been added (lines 229-240)***

Reviewer #2: The article provides a persuasive rationale for why research into the efficacy of genetic genealogy outside the US is needed, both in the introduction and concluding paragraph. The limitations of the study are generally well stated, and the results are both novel and useful. A few minor points to consider are detailed below:

· More detail related to the DTC tests needs to be given in the methods (I know it is mentioned at the bottom of Table 1) – from which companies were the tests purchased and how big were the SNP arrays,

Response: Additional section on Genetic Testing by DTC providers added to M&M. Lines 108-117.

… and a comment on whether there is any suggestion that the different types of DTC uploads affected the power of the GEDmatch search?

Response: A sentence has been added to comment that the study was too small for any differences between outcomes based on DTC provider to be considered significant (lines 448-450).

In this respect, we do not think the different chips made any difference to the outcomes of the searches but we do not have any empirical evidence and it is not something we considered when doing the analyses. The critical factor in each case was the nature of the cousin matches at GEDmatch. GEDmatch do not reveal much about the way they do their matching but the thresholds are very permissive with matches potentially going right down to 7 cM. There might be some discrepancies with the low-level matches but these would not have been useful anyway for our purposes and we do not generally work with small segments under about 15 cM anyway because there is such a high false-positive rate.

There are far more Ancestry testers at GEDmatch than from any other company. That is mainly because Ancestry has the largest database (16 million testers). 23andMe have tested over 10 million people but these people are largely testing for health rather than ancestry and are far less likely to opt in to cousin matching. The three people in our study who tested at 23andMe all had lots of matches with people who had tested at Ancestry so we do not think they were missing out on any matches. We believe the fact that we did not solve any of the 23andMe uploads was just a chance finding in a very small dataset. Another factor might be that the volunteer subjects who used 23andMe were potentially less likely to be interested in genealogy, and hence less likely to have closer relatives in the databases.

· Table 1 – The legend here has to include what the ethnicity codes mean since it will not be intuitive to most people.

Response: Legend revised with additional information

· Since all volunteers had previously purchased a DTC test themselves, do you worry that this is self-selecting a segment of the population along, for example, socio-economic lines, where relatives are also more likely to have had a test? Similarly, we know close relatives often get tests together (either because they see the results from a close relative and want to do it, or because they are given as gifts from a satisfied 'customer'), so this again may bias the results in relation to subject 1. Neither of which invalidate the study, but it is something to consider

Response: Additional paragraph on the potential for bias in the sample set added to the discussion (lines 455-466)

· Line 203 – you need to make it clearer that AD is the name you have given to the individual who was a match since to start with I assumed it was some sort of acronym, especially as SA is always written in quotation marks

Response: Clarified by comment in the results section (lines 248-252) and modification of sentence at line 267

· It would be much easier to understand the familial links for subject 3 if a family tree was included.

Response: Family tree added (figure 1)

Highlights

- Ten volunteer UK residents provided high density SNP data downloaded from a direct-to-consumer testing company.
- Data were uploaded anonymously to GEDmatch and genetic genealogy techniques used to attempt to identify the volunteers by name
- Four out of a sample of ten UK residents were identified by name or as one of a group of named siblings.
- Demonstration that genetic genealogy techniques can be effective in investigation of individual identity for cases in the UK

An empirical investigation into the effectiveness of genetic genealogy to identify individuals in the UK.

Jim Thomson[a], Tim Clayton[a], John Cleary[b,c], Maurice Gleeson[b], Debbie Kennett[b,d], Michelle Leonard[b], Donna Rutherford[b].

[a] Eurofins Forensic Services, Teddington, TW11 0LY UK
[b] ISOGG UK and Ireland
[c] Heriot-Watt University, Boundary Road North, Edinburgh, Scotland, EH14 4AS, UK
[d] University College London, Gower Street, London WC1E 6BT, UK

Corresponding author email: jimthomson@eurofins.co.uk

An empirical investigation into the effectiveness of genetic genealogy to identify individuals in the UK.

Abstract

The use of genetic genealogy techniques to identify Joseph James DeAngelo as the prime suspect in the Golden State Killer case in 2018 has opened up a new approach to investigation of cold cases. Since that breakthrough, genetic genealogy methods have been reported to be applied to around 100 cases. To date, all of these reports relate to investigations in the US, where the high uptake of "direct-to-consumer" (DTC) genetic testing by individuals conducting private ancestral research has provided the necessary publicly available data for successful forensic investigations. We have conducted a study to assess the likely effectiveness of genetic genealogy techniques if applied to investigations in the UK. Ten volunteers provided their own SNP array data, downloaded from a DTC provider of their choice. These data sets were anonymised and uploaded to the GEDmatch Genesis genealogy website, mimicking data sets from unsourced crime samples or unidentified human remains. A team of experienced genealogists then attempted to identify the donors of the anonymised data sets by working with matches on the database and identifying points where the matches' trees intersect to determine their shared family lineages which were further investigated using traditional resources (such as birth, marriage, death and census records, social media and online family trees). Through these methods, four of the ten donors were identified, at least to the level of one of a set of siblings. This confirms that, despite the over-representation of US citizens on publicly accessible genealogy databases, there is still potential for effective use in investigations outside the US where legislation permits. One of our four identified individuals was of Indian heritage (via St Vincent and the Grenadines) highlighting that in the right circumstances individuals of non-European origin can be identified.

Keywords

Genetic Genealogy, Forensic Genealogy, GEDmatch, SNP

Introduction

The rapid rise in prominence of genetic genealogy methods applied to forensic investigations in 2018 and 2019 has had a transformative impact on investigation of cold cases for which unsourced DNA evidence is available [1]. This investigative approach, known in various circles as investigative genetic genealogy or forensic genealogy, takes advantage of the rapidly expanding resources available to amateur genealogical researchers investigating their bio-ancestral origins and family histories. Direct-to-consumer (DTC) genetic testing using high density SNP arrays has become cheap and widely available, fuelling a rapid increase in the size of searchable SNP databases held by the DTC companies and by other independent organisations, such as GEDmatch.

To utilise these resources for forensic investigations, sufficient good quality DNA assumed to originate from the perpetrator of a crime (or from unidentified remains) is needed. This requirement may limit this approach to cases where semen or relatively large bloodstains attributable to the perpetrator are available, or to bodies or body parts from which sufficient DNA can be recovered. From such samples, SNP genotype data compatible with those generated by the DTC providers must be generated. Typically, the

high density SNP arrays used for these analyses include approximately 600,000 – 730,000 SNPs, dependent on the supplier, most of which are autosomal but which often also include X and Y chromosome and mitochondrial SNPs [2]. These data can then be uploaded to a permissible genealogy database such as GEDmatch, whereupon individuals with the relevant genealogical skills are needed to conduct the required searches and pursue the genealogical investigations.

Although publicly available genetic data had previously been used to assist in cold case identifications [3–5] it was the identification of Joseph James DeAngelo as the prime suspect in the Golden State Killer case which alerted many to the potential of such investigations. DeAngelo was identified following SNP genotyping of a semen sample taken from a victim of the killer in the 1980s. The data were uploaded to the GEDmatch genetic genealogy database, and investigative work by Barbara Rae-Venter, a genetic genealogist working with the FBI, finally led to the identification of DeAngelo who was arrested and charged with first degree murder in April 2018. [6–8]

Between April 2018 and January 2019 a total of 28 cases were publicised wherein the law enforcement agencies announced identification of DNA from a suspected perpetrator with the aid of genetic genealogy [9].
These cases and a number of subsequent reports all related to crimes committed in the USA where the uptake of genetic genealogy tests by the public has been highest. Erlich *et al* [10] determined that nearly 60% of genealogy searches in a large database of 1.28m individuals would return a match to a relative with at least 100 cM of shared DNA, a level usually indicative of a 3rd cousin (3C) relationship or closer. However, this empirical study utilised a test set comprised of subjects already on the database and so is subject to any bias inherent in the database constitution (such as over-representation of US residents of European ancestry). In two other journalistic studies, also focussed on identification of US residents, Aldhous [11] reported that 6 out of 10 Buzzfeed employees who provided DNA results from an unspecified DTC genetic testing provider were identified following genetic genealogy investigations, and Brown [12] reported the successful identification of herself by an independent genealogist. Despite these useful investigations, the question of whether such investigative methods would yield successful results if applied to searches outside the USA (relating to (a) unidentified human remains or (b) the victims or perpetrators of crimes) remains untested. This study seeks to provide preliminary information to address this question based on a small convenience sample of ten individuals to assess the potential effectiveness of genetic genealogy methods in identifying UK residents.

Materials and Methods

Selection of subjects

Ten volunteer subjects were recruited from <*information removed*> staff. The primary criterion was that the individuals had privately purchased and completed a DNA ancestry/genealogy test from one of the major direct-to-consumer (DTC) providers. All volunteers were fully briefed on the project aims and potential implications of third-party scrutiny of their DNA records and associated genealogy and provided full consent. The project was also approved by our in-house ethical approval procedures. The only other criterion applied was that all subjects were UK residents with at least three years of continuous residency. The recruitment process was not intended to provide a

representative sample of the UK population, but represented a convenience sample suitable for an indicative feasibility assessment.

For each subject, sex and date of birth were recorded. Subjects also provided information on their country of birth; their own self-defined ethnicity (using United Kingdom Home Office self-defined ethnicity codes) [13]; and the country of birth and ethnicity (if known) of both of their parents and all of their grandparents.

Genetic testing by DTC providers

Genetic tests were procured from either Ancestry (Lehi, UT, USA) or 23andMe (Sunnyvale, CA, USA) by the individual volunteers using the standard UK web-based order processes. Tests were conducted between April 2016 and January 2019 using the then current SNP panels. For all seven Ancestry users, this was the Ancestry v2 chip based on the Illumina OmniExpress Plus chip with ~669,000 SNPs [14]; for two 23andMe users, this was the 23andMe v4 chip, a bespoke customised Illumina chip with ~602,000 SNPs[15] ; and for one 23andMe user, this was the 23andMe v5 chip, the Illumina GSA chip with~640,000 SNPs [16].

Upload of data to GEDmatch Genesis

All subjects downloaded their own raw DNA data files from their DTC provider. These data files were transferred to the project manager. For each subject, a new account was created on the GEDmatch website using a random user name and a generic email address unattributable to the subject. DNA SNP data from each subject was uploaded to the GEDmatch Genesis database and all uploads were changed to a privacy status of "*Research*" which is an option "*provided primarily for artificially created research kits*" [17] and which ensures that the subject's DNA will not be included in match results of other users. Following successful upload, the kit numbers generated by GEDmatch for each subject were provided to the genealogy team, enabling them to conduct the required searches. All searches were conducted using the GEDmatch Genesis beta website (which was launched on 20 December 2018). GEDmatch Genesis and GEDmatch were merged on 1st June 2019 and the website has reverted to the original GEDmatch domain name but the new Genesis features are now integrated into the website [18].

GEDmatch searching

All GEDmatch searches and subsequent genealogical investigations were conducted on a *pro bono* basis by a team of five highly-experienced genetic genealogists.

GEDmatch is a platform for testers who have used one DTC provider to compare their data with testers from all others, offering a number of online tools for its users, some of which are free to use while others require a small monthly subscription. The base function is a search of the kit of interest for matches across the whole available database ("One-to-many" comparison), which can be followed by detailed comparisons between a kit and each of its matches ("One-to-one" comparison, for autosomal DNA plus a specific tool for the X chromosome). Another free to use tool allows the user to see whether some of a kit's matches also match each other ("People who match both or 1 of 2 kits"), which enables networks of shared relationships to be built around the subject, and a

related subscription tool maps “Clusters” of mutually related kits. These network and cluster building tools were the most important tools for the analysis presented here, along with the shared cM data, as they indicated where intersections between the family trees of the subject and matches may be found.

Following upload, “one-to-many” searches were carried out for each subject to provide the list of 3000 closest matches in the GEDmatch database (estimated size at time of study: 1.2 million) [19]. Population admixture proportions were estimated using the Eurogenes K13 model [20], one of a number of admixture calculators included in the GEDmatch toolkit. Each kit was assessed for runs of homozygosity by using the “Are your parents related?” tool.

Genetic genealogy investigation strategy

Investigation of the identity of each subject progressed in several distinct stages:

1. Assessment of the amount of shared DNA with the top matches and the degrees of relationship predicted from them using the Shared cM Project tool that supplies ranges of known relationships observed within crowd-sourced data. The probabilities of predicted relationships quoted below are based on this tool. [21]
2. Identification of clusters of shared matches that might indicate a common ancestor (cluster analysis).
3. Examination of family trees of the top matches to identify points of intersection, likely to indicate common ancestors. Some matches were linked to online family trees but these varied in completeness and accuracy. Genealogical work using standard genealogical records (e.g. birth, marriage and death records, census records, etc.) or public domain information (e.g. online obituaries, electoral registers, social media etc.) followed to verify, correct and extend existing trees, and to build new trees from scratch where none existed.
4. Following identification of possible common ancestors, trees were built forward in time from them to identify candidates for the person of interest fitting the profiles of gender, age, circumstantial evidence, and other information.

The pilot study ended at the beginning of June 2019 and all the profiles were then deleted from GEDmatch.

Results

Subjects

The ten volunteer subjects included six males and four females aged between 22 and 56. Seven were born in England or Wales with parents and grandparents also declared as White British from the UK. One was born in England to a Chilean father and English mother, although declared her father’s and paternal grandfather’s ethnicity as White British (reported as members of a British expatriate community in Chile). One was born in England to parents and grandparents of Indian ethnicity who had been resident in St Vincent and the Grenadines. One was born in Romania to Romanian parents and grandparents.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Assigned Name Initials	"AW"	"JS"	"SA"	"SE"	"MM"
Sex	M	F	M	F	M
Date of birth	1977	1997	1964	1991	1977
Country of birth	England	England	England	Romania	Wales
Self-declared ethnicity	W1	W1	W1	W9	W1
Father	England, W1	England, W1	England, W1	Romania, W9	Wales, W1
Mother	England, W1	England, W1	England, W1	Romania, W9	Scotland, W1
Paternal GF	England, W1	England, W1	Wales, W1	Romania, W9	Wales, W1
Paternal GM	England, W1	England, W1	England, W1	Romania, W9	Wales, W1
Maternal GF	England, W1	England, W1	England, W1	Romania, W9	Scotland, W1
Maternal GM	England, W1	England, W1	England, W1	Romania, W9	Scotland, W1
DTC Provider	Ancestry	Ancestry	Ancestry	23andMe	Ancestry

	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10
Assigned Name Initials	"GP"	"AB"	"BM"	"AJ"	"JB"
Sex	M	M	F	F	M
Date of birth	1983	1982	1992	1984	1963
Country of birth	England	England	England	England	England
Self-declared ethnicity	W1	W1	W1	A1	W1
Father	England, W1	England, W1	Chile, W1	St.Vincent, A1	England, W1
Mother	Guernsey, W1	England, W1	England, W1	St.Vincent, A1	England, W1
Paternal GF	Unknown	England, W1	Chile, W1	St.Vincent, A1	Shetland, W1
Paternal GM	Unknown	England, W1	Chile, W1	St.Vincent, A1	Shetland, W1
Maternal GF	Guernsey, W1	England, W1	Ireland, W2	St.Vincent, A1	England, W1
Maternal GM	Guernsey, W1	England, W1	Ireland, W2	St.Vincent, A1	England, W1
DTC Provider	23andMe	Ancestry	23andMe	Ancestry	Ancestry

Table 1. Details of the ten volunteer subjects. Ethnicity is defined using the United Kingdom Home Office self-defined ethnicity codes [13]; W1=British; W2=Irish; W9=Any other white background (other than British, Irish or Gypsy/Irish traveller); A1=Indian

One-to-Many matches

One-to-many comparison searches were conducted for each subject. Outcomes are summarised in table two, with matches categorised by the total shared cM of DNA between the subject and the nearest matches. A greater amount of shared DNA is indicative of a closer relationship between the subject and the matched sample. One subject (“AW”) was identified as having two close relatives (likely to be a parent or child, and a sibling) on the GEDmatch database. Two subjects had closest matches in the 80-500 cM range, suggestive of second or third cousin (2C-3C) relationships and a further three had closest matches with 50-80 cM shared, suggestive of 3C-4C. The remaining four subjects had no matches sharing more than 50 cM although, along with all other subjects, had several or many (up to 859) more distant matches in the 20-50 cM range suggestive of 4C or more distant relationships.

Range cM	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10	Likely relationship
~3570	1	0	0	0	0	0	0	0	0	0	Parent or child
2200-3300	1	0	0	0	0	0	0	0	0	0	Sibling (full)
1200-2200	0	0	0	0	0	0	0	0	0	0	Half siblings, aunts, uncles, grand-parents/children
550-1200	0	0	0	0	0	0	0	0	0	0	1C ~ 1C removed
80-500	0	0	1	0	0	0	0	0	2	0	approx 2C~3C or 1C removed
50-80	2	0	0	0	3	0	0	2	4	1	approx 3C~4C
30-50	6	1	7	8	7	16	6	36	2	5	approx 4C or more distant (could be closer)
20-30	108	99	81	74	71	199	67	823	2	41	approx 5C or more distant (could be closer)

Table 2. Summary of one-to-many searches. For each subject, the numbers of close matches in each cM range is shown.

Time spent on identification investigations

The genealogical team spent 2-3 hours on a triage analysis of each kit to judge its likely chances of success. 5 kits with low matches were judged as requiring potentially over 100 hours of work to identify distant cousins of the matches and these were not pursued, although a start was made in some cases. In total some 200-300 hours were put in by the genealogical team on pursuing the 5 cases judged most promising, including reporting and team meetings. The quickest case was solved in around 3 hours, while more complex ones needed between 50-100 hours to solve. The genealogical team were aware that the targets were associated with <*information removed*>, but had no staff list to work from so that the identifications of the targets' names had to arise out of the genealogical tracing.

Outcomes of identification investigations

Following the initial one-to-many matching, a degree of triage was applied with further investigations focussed mainly on those individuals with closer nearest matches. The results for all subjects are discussed below starting with the four subjects who were identified by the investigation either by name, or as one of a set of named siblings. In the following descriptions, initials in quotation marks (e.g. "AW") refer to the false identities assigned to the ten subjects under investigation, and other initials (e.g. AD) refer to real individuals identified as matches or likely relatives of those subjects during the investigations.

Subject 1 (assigned initials "AW")

One parent/child match and one sibling match were identified on GEDmatch making this the most straightforward identification case. Cross-referencing of recent electoral registers and birth, marriage and death indexes enabled a candidate family group to be identified. The person of interest "AW" was identified as one of two twin brothers. The investigation did not identify whether these twins were monozygotic or dizygotic which would have implications for any further DNA identification testing were this a real investigative case. There was a third brother who could be eliminated as being "AW", as

he also had a genetic profile in GEDmatch matching in the expected sibling range to "AW".

Subject 3 (assigned initials "SA")

There was a good top match at 125 cM to an individual called AD, predicting approximately 85% probability of being 3C or closer to "SA", allowing the descendants of common great-great-grandparents to be targeted in the search. AD's paternal side was eliminated by an identified close match not shared with "SA", supported by a small but likely significant 13 cM X chromosome match between AD and "SA", suggesting that the connection may be found on both persons' maternal trees (since, both being male, they had received X chromosomes only from their mothers). The shared matches for the two gave no further clear leads: for example, one promising shared match, JB, appeared likely to connect to AD through an unknown parent, meaning this parental connection would need to be solved before the connection to AD and "SA" could be solved in turn.

A genealogical search strategy was adopted constructing a tree forwards in time to identify all descendants of AD's maternal great-great-grandparent couples. The search was limited to living male descendants whose immediate ancestor in the tree was their mother, and so could have shared X chromosome segments with AD (ruling out any male to male descents) to identify candidates to be "SA" or candidates for targeted testing to find closer relatives. This led to candidate siblings, one of whom was Subject 3, whose tree intersected with AD's tree at two points, sharing two of AD's great-great-grandparent couples. The candidate siblings were younger than AD, making them 2C1R (second cousins one removed) to AD along two routes, therefore double 2C1R to AD (Figure 1).

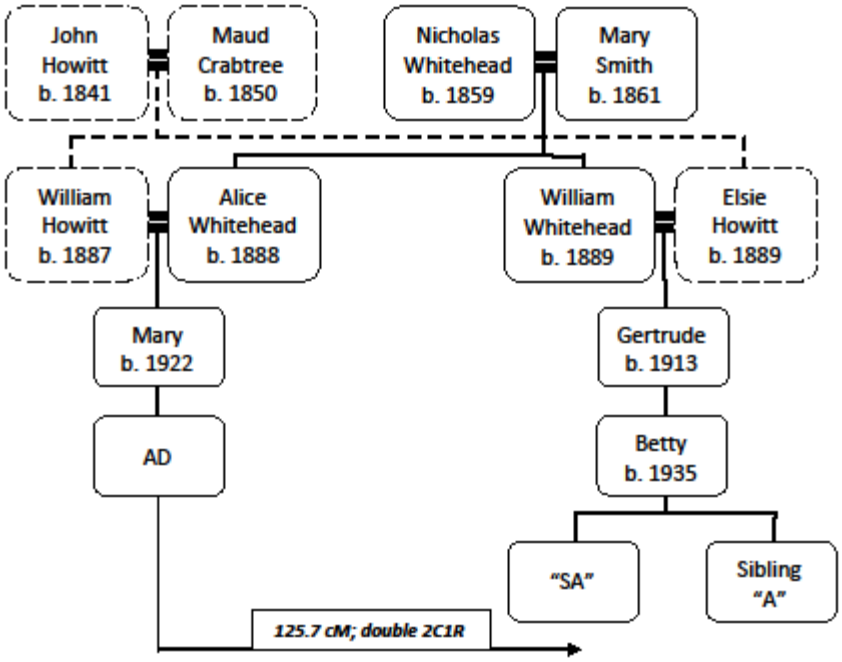

Figure 1: Tree of recent ancestry of “SA”. "SA" and his closest match AD are second cousins once removed (2C1R) twice, because each descends from two ancestral couples due to marriages between two pairs of siblings. Names have been changed.

Subject 9 (assigned initials “AJ”)

The top match for “AJ” (364 cM – likely between 2C-1C1R) was identifiable from GEDmatch as KW, resident in High Wycombe, UK and related to members of a community of Indo-Vincentians (people from St Vincent and the Grenadines of Indian origin). “AJ” was correctly identified as Subject 9, and was closely related to a number of people in the St Vincent and Grenadines Indian community in the UK, who are descended from 19th century Indian indentured settlers in St Vincent, named B and K, and whose network of family and social relationships and online family trees gave the clues that led to identifying the female subject. Although the identification was successful, the genealogists had cautioned that they could not guarantee they had identified all the descendants of B and K who might fit the profile because the relevant birth, marriage and death records in St Vincent are not accessible from the UK. They suggested that identifying Subject 9’s mother, who is also likely to be descended from the founding couple, would help to corroborate the identification. Subject 9 was more distantly related to her matches through her identified paternal line than predicted from the amount of shared DNA, likely due to pedigree collapse through endogamy, increasing observed levels of shared DNA across the family network, causing the search to be broadened to include more distant relatives of the matches who were also descended from the founder couple.

Subject 10 (assigned initials “JB”)

The top match at 53 cM was not especially close, with an approximately 50-60% possibility of being related to “JB” as 4C or closer, and the other matches were likely to be more distant. The closest match, AR, was identifiable as a resident of Aberdeenshire, Scotland, with her paternal ancestry from the Shetland Islands and maternal from north-east Scotland.

Several other close matches with “JB” were also Shetlanders or had recent ancestors (up to grandparents) from Shetland. A 5000-person tree was constructed for the Shetland cluster, drawing upon some very well constructed and referenced trees made available publicly by family historians with a deep knowledge of Shetland communities.

”JB” had a large cluster of matches indicating an ancestor from the Shetland Islands, who could be a parent, grandparent or great-grandparent. This community is endogamous so the matches within the cluster share several sets of ancestors in common. Mapping the intersections of the shared relationships between clusters of matches allowed modelling of the places on the trees under construction where “JB” could be. “JB” had more distant matches with non-Shetlanders, which were assumed to represent one side of his ancestry (paternal or maternal), and marriages on the tree between Shetlanders and non-Shetlanders were identified and profiled. The identification was made when one of the Shetland match’s trees intersected with a non-Shetland match’s tree, which turned out to be on Subject 10’s mother’s side and from Yorkshire, the father’s side being the Shetland side of the family. Through this route, “JB” was identified as probably being one of three brothers. One of these siblings was

Subject 10. There was also a sister who can be eliminated on the assumption that the target person is male.

Two further subjects (5 and 8) had promising matches likely to be at the 3C-4C level but in neither case could informative trees be developed and neither subject was identified.

Subject 5 (assigned initials "MM")

Three top matches shared 50-80 cM which suggested that identification might be possible. However, "MM" was not identified, but the 69 cM shared with the top match, WB, indicated an approximately 75% chance that the ancestors in common with that top match were 3C or closer, and just a 5% chance that the relationship was 4C or more distant. Therefore, the ancestors in common were likely to be within WB's 8 pairs of great-great-grandparents. This considerably limited the pool of potential persons to be "MM". There seemed a strong likelihood that "MM" was Welsh, or might be the child of Welsh parents. However, there was also a possibility of endogamy in these Welsh family trees which would make the relationship more distant than predicted.

Subject 8 (assigned initials "BM")

"BM" had two matches with just over 50 cM shared DNA and who could be identified. These matches were likely to be 3C-4C and there were good chances of capturing 100% of the potential ancestors in common. However, neither match had available trees and neither match shared X-DNA with "BM". Most of the other matches were too distant to make successful searching likely.

The remaining four subjects (Subject 2, 4, 6 and 7) were the least promising with no matches sharing more than 50 cM. For all four of these individuals, cluster analysis and, where appropriate, some preliminary investigations of ancestral trees were conducted but in no cases could substantial headway be made to indicate the identity of the subjects. Relationships between 4th cousins (4C) imply common ancestors among 16 great-great-great-grandparental couples and a typical person may not be able to identify all of their ancestors at this level, leaving blocked lines on their trees and the possibility that the common ancestors of interest can be on those blocked lines. Consequently, match lists with the closest matches at likely 4C distance present poor chances for successful identification. Summaries of the reported results for these four cases are shown below.

Subject 2 (assigned initials "JS")

The top genetic matches to "JS" were predicted to be no closer than 4C, and were likely in many cases to be more distant than that. None of the top matches were matches to each other so they were likely to converge with "JS" on different lines several generations back, making identifying a small pool of candidates to profile challenging.

Subject 4 (assigned initials "SE")

The top matches were all more than 80% likely to be no closer than 4C, with no clustering or trees to go on for further analysis.

Subject 6 (assigned initials "GP")

The top genetic matches to "GP" were predicted to be no closer than 4C, and were likely in many cases to be more distant than that. "GP" may have a recent ancestor with Ukrainian Jewish ancestry, but it was not possible to quantify how recently that ancestor lived.

Subject 7 (assigned initials "AB")

The top matches to AB were particularly low, with only a 12% or less chance of any of them being at the 3C level, and most were likely to be more distant than 4C. This rendered the possibility of identifying ancestors in common between AB and his matches remote. Some clusters of matches offered prospects for research as they have closely related family clusters of their own with trees, but as their relationships to AB were likely very distant, any reconstructed trees were likely to be partial up to the necessary degree of ancestry and may not have included the ancestors in common.

Discussion

The success of genetic genealogy methods in identifying victims and perpetrators of crime in the United States has been widely reported and is of significant interest to other jurisdictions. Whilst there are undoubtedly significant ethical and legislative hurdles and considerations to address, which in some jurisdictions may prevent such approaches being considered, there are also practical questions regarding the applicability of the method outside of the US.

We have provided some preliminary data to demonstrate the efficacy of the method in identifying UK residents, The demonstration that four out of a convenience sample of ten volunteers could be identified either by name, or as one of a set of siblings, is a useful indicator that the predominance of US individuals on GEDmatch and other genealogical databases does not preclude successful identification of individuals from outside the US.

Although GEDmatch does not publish the composition of their database by country of origin, co-founder Curtis Rogers has confirmed that "our number one users are in the United States, then Canada, then England, and then Australia" [22]. Although it is difficult to source reliable data on the proportion of US, Canadian and Australian residents with British ancestral origins it is evident that significant numbers of all of these nationals will have British ancestry and are likely to have living relatives in the UK. Of all countries other than the US, it seems likely that residents of the UK will be amongst those with the highest chance of having relatives present on GEDmatch or other permissible databases whether those relatives are currently resident in the UK or elsewhere. In this context, it is notable, although not statistically significant, that the subject in our study with Romanian ancestry generated no matches sharing more than 50 cM and no clustering or existing trees to give grounds for a successful identification. Without more detailed knowledge of the demographic composition of GEDmatch or other databases it is hard to predict the effectiveness of such searching methods if applied to residents of other countries although it seems likely that residents of European countries with higher representation in the US population, such as Germany and Ireland, may be most likely to lead to successful identifications.

The identification of one individual of Indian origin in our sample set is noteworthy in that it demonstrates that, in the right circumstances, searches for individuals of non-European origin can be successful. In this example (subject 9, from a small Indo-Vincentian population), and in the case of subject 10, whose family originated from Shetland (an island group 170km north of the Scottish mainland), the investigations were significantly assisted by well-documented genealogical records and family trees developed for these localised populations by interested historians and genealogists. Both of these individuals are now resident in the south of England, far from the geographical origins of the population groups which led to their identification, again highlighting that in a forensic investigation it is hard to predict any likelihood of successful identification without at least an initial triage and review of matches and cluster information.

The small size of this sample, and the possibility of bias introduced by selecting subjects from employees of a scientific services company who may fall into socio-economic groups more likely to have relatives who have also carried out such tests, precludes any extrapolation of the success rates obtained here to predict likely success in the UK population at large. In this respect we also acknowledge that subject 1 was aware that his parent and brother had already carried out such testing prior to this study, but his inclusion in the volunteer sample set was not influenced by this prior knowledge as we accepted all volunteers meeting our three year residency criteria.

The small sample set also precludes any further interpretation of the observation that all four of the identified subjects were tested by Ancestry whereas the three subjects tested by 23andMe remained unidentified.

The study reported here was completed prior to the change to GEDmatch's Terms of Service and Privacy Policy which has significantly altered the landscape for genetic genealogy searching, at least in the near future. In May 2018, following the initial reports of successful searches in the Golden State Killer case, GEDmatch amended their previously non-specific Terms of Service and Privacy Policy conditions to specifically allow law enforcement usage but only to "(1) identify a perpetrator of a violent crime against another individual; or (2) identify remains of a deceased individual" [23]. "Violent crime" is defined here as homicide or sexual assault. However, in May 2019 the policy was again changed, this time requiring that users "opt-in" if they consent to their DNA data to be used for Law Enforcement purposes [24]. All current users at that time were therefore automatically set as "opted-out" of law enforcement searches effectively reducing the sample set available for such searching from ~1.2 million to zero at that point. Since then, it is reported that about 200,000 users have opted in to the LE option (by December 2019) [25], but clearly the overall efficacy of investigative genetic genealogy is highly dependent on the available data set and at present the results described in this study will not be representative of LE searches of the much smaller dataset now available through GEDmatch. In December 2019 GEDmatch was purchased by the forensic genomics company Verogen, who have announced plans to add new features to improve the functionality of the database and to make the database more secure. Following the purchase, all EU users were opted out of LE matching and had to re-consent to comply with GDPR (General Data Protection Regulation). In the short term this will further reduce the potential utility of the database for law enforcement agencies in the EU [26].

Although to date the majority of reported successful searches have utilised GEDmatch, other options are available to counterbalance this reduced access to GEDmatch data. In January 2019 another DTC company FamilyTreeDNA (FTDNA) announced that it was now testing samples for the FBI and allowing them to upload profiles to its database [27]. FTDNA currently allows US law enforcement agencies to register for uploads and considers work with non-US agencies on a case-by-case basis [28]. Another company, Othram Inc, which specialises in whole-genome sequencing for forensic purposes, launched a new database in December called DNA Solves that allows users to upload their data to help solve crimes and identify victims [26,29].

In summary, this demonstration of the effectiveness of investigative genetic genealogy provides the initial evidence needed by regulatory bodies outside the US to prompt further consideration of the ethical and regulatory frameworks to enable safe and socially-acceptable introduction of investigative genetic genealogy with appropriate regard for privacy and data security issues. This can facilitate the introduction of this valuable investigative approach to assist in the identification of unidentified human remains as well as the victims and perpetrators of crime in the UK and elsewhere.

References

[1] D. Kennett, Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes, Forensic Sci. Int. 301 (2019) 107–117. https://doi.org/10.1016/j.forsciint.2019.05.016.

[2] International Society of Genetic Genealogy Wiki, Autosomal DNA Testing Comparison Charts, Int. Soc. Genet. Geneal. Wiki. (2016). https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart (accessed January 3, 2020).

[3] T. Arango, The Cold Case That Inspired the ‘Golden State Killer’ Detective to Try Genealogy, New York Times. (2018). https://www.nytimes.com/2018/05/03/us/golden-state-killer-genealogy.html.

[4] S. Augenstein, The tale of the abandoned girl’s DNA that led to a notorious cold case, Forensic Mag. (2017) 1–9. https://web.archive.org/web/20191020024455/http://www.forensicmag.com/article/2017/02/tale-abandoned-girls-dna-led-notorious-cold-case.

[5] M. Kayser, Forensic use of Y-chromosome DNA: a general overview, Hum. Genet. 136 (2017) 621–635. https://doi.org/10.1007/s00439-017-1776-9.

[6] C. Phillips, The Golden State Killer investigation and the nascent field of forensic genealogy, Forensic Sci. Int. Genet. 36 (2018) 186–188. https://doi.org/10.1016/j.fsigen.2018.07.010.

[7] T. Arango, A. Goldman, T. Fuller, To Catch a Killer: A Fake Profile on a DNA Site and a Pristine Sample - The New York Times, New York Times. (2018). https://www.nytimes.com/2018/04/27/us/golden-state-killer-case-joseph-deangelo.html.

[8] 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 (2018) e2006906. https://doi.org/10.1371/journal.pbio.2006906.

[9] E.M. Greytak, C.C. Moore, S.L. Armentrout, Genetic genealogy for cold case and active investigations, Forensic Sci. Int. 299 (2019) 103–113.

https://doi.org/10.1016/j.forsciint.2019.03.039.
[10] Y. Erlich, T. Shor, I. Pe'er, S. Carmi, Identity inference of genomic data using long-range familial searches, Science. 362 (2018) 690–694. https://doi.org/10.1126/science.aau4832.
[11] P. Aldhous, We tried to find 10 buzzfeed employees just like cops did for The Golden State Killer, BuzzFeed News. (2019) 1–15. https://www.buzzfeednews.com/article/peteraldhous/golden-state-killer-dna-experiment-genetic-genealogy (accessed January 3, 2020).
[12] K. V. Brown, A Researcher Needed Three Hours to Identify Me From My DNA, Bloomberg. (2019) 1–5. https://www.bloomberg.com/news/articles/2019-04-12/a-researcher-needed-three-hours-to-identify-me-from-my-dna.
[13] HM Government, Criminal Justice System Exchange Data Standards Catalogue, 2014. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/691544/self-defined-ethnicity-18plus1.pdf.
[14] R.A. Canada, Ancestry V2 Chip, Haplogroup. (n.d.). https://haplogroup.org/ancestry-v2-chip/.
[15] R A Canada, 23andMe V4 Chip, Haplogroup. (n.d.). https://haplogroup.org/23andme-v4-chip/.
[16] 23andMe, Int. Soc. Genet. Geneal. Wiki. (n.d.). https://isogg.org/wiki/23andMe.
[17] GEDmatch raw data upload utility, (n.d.).
[18] Kitty Cooper, GENESIS Basics: GEDmatch reinvented part 1, Kitty Cooper's Blog. (2019). https://blog.kittycooper.com/2019/02/genesis-basics-gedmatch-reinvented-part-1/.
[19] Genealogical Database Growth Slows, DNA Geek. (2019). https://thednageek.com/genealogical-database-growth-slows/.
[20] D.F. Serviski, Updated Eurogenes K13 now at GEDmatch, Eurogenes Genet. Ancestry Proj. (2013) 2–6. http://bga101.blogspot.com/2013/11/updated-eurogenes-k13-at-gedmatch.html.
[21] DNA Painter, (n.d.). https://dnapainter.com.
[22] C. Curtis, 10094220 @ www.abc.net.au, (2018). https://www.abc.net.au/news/science/2018-08-13/online-genealogy-police-dna-databases-golden-state-killer/10094220.
[23] D. Kennett, Updates to the Terms of Service and Privacy Policy at GEDmatch, Cruwys News. 101 (2018). https://cruwys.blogspot.com/2018/05/updates-to-terms-of-service-and-privacy.html.
[24] A. Vaughan, DNA database opts a million people out from police searches, New Sci. (n.d.). https://www.newscientist.com/article/2203857-dna-database-opts-a-million-people-out-from-police-searches/.
[25] M. Taylor, Verogen CEO: 'GEDmatch Will Be Improved, Not Changed,' Forensic. (2019). https://www.forensicmag.com/559058-Verogen-CEO-GEDmatch-Will-Be-Improved-Not-Changed/.
[26] Justin Petrone, Forensic Genomics Market Advances Due to Consumer Databases, Technology Innovation, Genomeweb. (2020). https://www.genomeweb.com/sequencing/forensic-genomics-market-advances-due-consumer-databases-technology-innovation#.XiV96DP7S71.
[27] S. Hernandez, family-tree-dna-fbi-investigative-genealogy-privacy @ www.buzzfeednews.com, (2019). https://www.buzzfeednews.com/article/salvadorhernandez/family-tree-dna-fbi-investigative-genealogy-privacy.
[28] FamilyTreeDNA, Law Enforcement Matching – Frequently Asked Questions, Fam.

Learn. Cent. (2019). https://www.familytreedna.com/learn/ftdna/law-enforcement-faq/.

[29] Andrea Leinfelder, The Woodlands-based Othram applies DNA sequencing to aid investigations, Houst. Chron. (2019). https://www.houstonchronicle.com/business/article/The-Woodlands-based-Othram-applies-DNA-sequencing-14934741.php.