Forensic Science International: Genetics 52 (2021) 102474

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Review article

Investigative genetic genealogy: Current methods, knowledge and practice

Daniel Kling[a,b,*], Christopher Phillips[c,**], Debbie Kennett[d], Andreas Tillmar[a,e]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[b] Department of Forensic Sciences, Oslo University Hospital, Oslo, Norway
[c] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain
[d] Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom
[e] Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Genetic genealogy
SNP microarrays
Whole-genome-sequencing
Familial searching
Identity by descent
Forensic DNA analysis
Crime investigation

ABSTRACT

Investigative genetic genealogy (IGG) has emerged as a new, rapidly growing field of forensic science. We describe the process whereby dense SNP data, commonly comprising more than half a million markers, are employed to infer distant relationships. By distant we refer to degrees of relatedness exceeding that of first cousins. We review how methods of relationship matching and SNP analysis on an enlarged scale are used in a forensic setting to identify a suspect in a criminal investigation or a missing person. There is currently a strong need in forensic genetics not only to understand the underlying models to infer relatedness but also to fully explore the DNA technologies and data used in IGG. This review brings together many of the topics and examines their effectiveness and operational limits, while suggesting future directions for their forensic validation. We further investigated the methods used by the major direct-to-consumer (DTC) genetic ancestry testing companies as well as submitting a questionnaire where providers of forensic genetic genealogy summarized their operation/services. Although most of the DTC market, and genetic genealogy in general, has undisclosed, proprietary algorithms we review the current knowledge where information has been discussed and published more openly.

1. Introduction

It is a fundamental principle of genetics that individuals who are closely related will share DNA from their common ancestors; and the more distant the relationship, the less DNA is shared. Familial searching of national DNA databases [1] using 16–22 autosomal STRs will only provide links through partial matches to immediate relatives such as siblings, parent-offspring (50% of DNA shared) or, at most, avuncular relationships, e.g. uncle-nephew (25% shared); although even half-sibling relationships can be difficult to resolve with limited STR data. Once familial searching is extended over a longer range to pairwise comparisons of first cousins, second cousins, third cousins and beyond (12.5%, 3.13% and 0.78% DNA shared, respectively) there is the requirement for genetic variation at much higher densities than the standard forensic tests have been able to achieve up till now. High-resolution commercial direct-to-consumer tests which include a relative-matching feature have been available for more than a decade [2]. These tests are currently analyzed using high-density microarrays genotyping more than 600,000 SNPs, providing matches with both close and distant relatives. By distant we refer to degrees of relatedness exceeding that of first cousins, in contrast to genealogists who use the term distant for relationships beyond 4th or 5th cousins. Genealogists have used these tests routinely since their inception as a tool to help with their family history research, both to confirm existing relationships and find new relatives [3]. Such tests are also used in unknown parentage searches [4,5], with thousands of adoptees, donor-conceived individuals and foundlings successfully using the commercial tests to connect with siblings and identify biological parents. Conversely, tests have revealed unexpected discoveries such as the finding of unknown siblings or the discovery that the social parent is not the biological parent [6]. Therefore, it was only a question of time before the same techniques were applied to forensic DNA from a crime-scene or the remains of missing persons. The barrier hindering the forensic implementation of long-range familial searching was the lack of a method to generate the required high-density SNP data from degraded DNA which would be compatible with the genetic genealogy databases.

Three major factors are necessary to reach the level of effectiveness for relative matching achieved by genetic genealogy: i. large-scale autosomal SNP genotype data with marker numbers in the hundreds

[*] Corresponding author at: Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden.
[**] Corresponding author.
E-mail addresses: daniel.kling@rmv.se (D. Kling), c.phillips@mac.com (C. Phillips).

https://doi.org/10.1016/j.fsigen.2021.102474
Received 2 October 2020; Received in revised form 12 January 2021; Accepted 27 January 2021
Available online 30 January 2021

Table 1
The percentage proportion by country of the ten most frequent GEDmatch uploads for user's country of origin. Web analytics data from Verogen (up to September 2020).

	Country	Users		Country	Users
1	United States	65%	6	Germany	1%
2	United Kingdom	9%	7	Sweden	1%
3	Canada	6%	8	Ireland	1%
4	Australia	4%	9	New Zealand	1%
5	France	2%	10	Netherlands	1%

of thousands and available at an affordable price; ii. large databases of these SNP genotypes open to public access; and iii. a simple but well-founded system for comparing related pairs using this large-scale SNP data. While the use of dense SNP microarray data had already been studied in forensic contexts [7–11], such technology became readily available to the public in 2007 through direct-to-consumer testing companies (the 'DTCs') with the launch of tests from deCODE Genetics and 23andMe, costing nearly $1000 [12]. Early tests were based on the Illumina OmniExpress microarray, but the field is now dominated by the Illumina Infinium Global Screening Array (GSA), which currently has a core set of 654,027 SNPs and the ability to add up to 50,000 custom markers.[1] As the cost of testing decreased and more companies entered the market, SNP databases began to grow exponentially. The inflection point was reached in 2018 and in that year more DNA tests were sold than in all previous years combined [13]. As of August 2020, the four principal genetic genealogy DTC companies have tested over 36 million people (see Table 2, Section 5).

An autosomal SNP-based system of matching relatives in a commercial DNA database first became available in 2009 with the launch of the Relative Finder[2] feature from 23andMe (now known as DNA Relatives). FamilyTreeDNA (FTDNA) introduced their Family Finder test in 2010.[3] AncestryDNA entered the autosomal SNP market in 2012[4] and MyHeritage DNA launched their DNA product in 2016.[5] Of the commercial companies, only FTDNA allows law-enforcement matching within the opted in section of its database. GEDmatch, a citizen science website founded in 2010, proved crucial to the initial development of investigative genetic genealogy. GEDmatch allows DNA profiles to be uploaded from a wide variety of sources, including law enforcement samples, so that cross-company comparisons can be performed using an additional range of tools.

The arrest of Joseph DeAngelo as the suspected Golden State Killer in 2018 brought the investigative use of genetic genealogy to the world's attention [14]. Genetic genealogy has since been used to generate investigative leads in nearly 200 cold cases and some active investigations [2,15–18]. Many of the technical details around the analysis of forensic DNA for long-range familial searching are still not in the public domain, as commercial interests restrict publication of much of the information needed to properly assess how large-scale SNP genotyping techniques are applied to evidential material – typically with DNA limited in quantity and quality. In addition, there is a lack of transparency on the part of law enforcement agencies. IGG is used to generate an investigative lead and the details of the IGG work have not yet been scrutinized in court. Contradictory stories of how the Golden State Killer was caught have been published and further details only became available two years after his arrest from information leaked to the *Los Angeles Times*.[6] Nevertheless, whole genome sequencing to create SNP datasets that mirror microarray-based genotyping has been widely adopted to ensure sensitivity to challenging forensic samples [17]. Many of these techniques adapted the approaches developed to analyze ancient DNA, where sequence targets are much more degraded [19]. While most relative searching systems are centred on matching stretches of shared DNA [20–22] (referred to as segments), alternative analyses exist and are being developed which could offer more viable approaches when insufficient SNP genotypes from poor DNA prevent reliable segment matching [23,24]. In this review, we attempt to fill some of the gaps in knowledge that currently exist, with emphasis on the DNA analysis regimes in use for long-range familial searches. To compensate for the lack of information in the public domain we sent out a questionnaire to some of the forensic science providers in the US. This includes a number of questions relating to the use of technologies and genetic genealogy in their assistance to law enforcement. The answers are submitted from private companies, potentially with conflicts of interests, and we have taken care to peer-review them as far as possible. The responses received to the questionnaire are compiled in Supplementary File S1.

We use the term investigative genetic genealogy (IGG), also known as forensic genetic genealogy, to describe the use of SNP-based relative matching combined with family tree research to produce investigative leads in criminal investigations and missing persons cases. The term forensic genealogy is sometimes used in this context but has a distinct meaning in US genealogical circles and relates to all questions of a legal nature that require genealogical analyses, including disputed inheritance, identification of military personal and citizenship claims.[7] Two papers published in 2019, Greytak et al. [15], and Kennett [2] provide informative overviews of genetic genealogy used in forensic investigations. Useful additional information, an updated review of forensic genetic genealogy practice and a list of many successful crime investigations was provided in 2020 by Katsanis [16]. We also recommend the comprehensive information compiled by the International Society of Genetic Genealogy (ISOGG) in their genetic genealogy wiki portal with 622 articles,[8] including a wealth of information on IGG.

2. Inference of relatedness

There is a wide range of approaches to infer the genetic relationship between two or more individuals [20,25–35]. The aim of relationship inference, as defined in this review, is to determine whether regions of DNA are shared identical by descent (IBD), i.e., through common ancestry. Comprehensive summaries of this topic are provided by Weir et al. [29], and Browning and Browning [36]. Speed and Balding [34] review methods referred to as part of the post-genomic era, which we term exploratory approaches. In contrast, Thompson [33] reviews what we term pedigree-based methods. The following sections provide a brief description of these approaches, summarized in Fig. 1, and an overview of the underlying statistical theory. We do not discuss the number of markers required for each approach in detail and all numbers should be seen as approximate, heavily dependent on the case, the population or other factors. As a rule of thumb, simple versions of exploratory approaches require higher marker numbers evenly distributed across the genome, while pedigree-based methods tend to require fewer markers, but still evenly distributed.

[1] Version 3, details available at: https://emea.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html.

[2] See: https://blog.23andme.com/news/introducing-relative-finder-the-newest-feature-from-23andme/

[3] See: https://thegeneticgenealogist.com/2010/07/19/a-review-of-family-tree-dnas-family-finder-part-i

[4] See: https://www.ancestry.com/corporate/newsroom/press-releases/ancestry.com-dna-launches

[5] See: https://blog.myheritage.com/2016/11/introducing-myheritage-dna/

[6] See: https://www.latimes.com/california/story/2020-12-08/man-in-the-window.

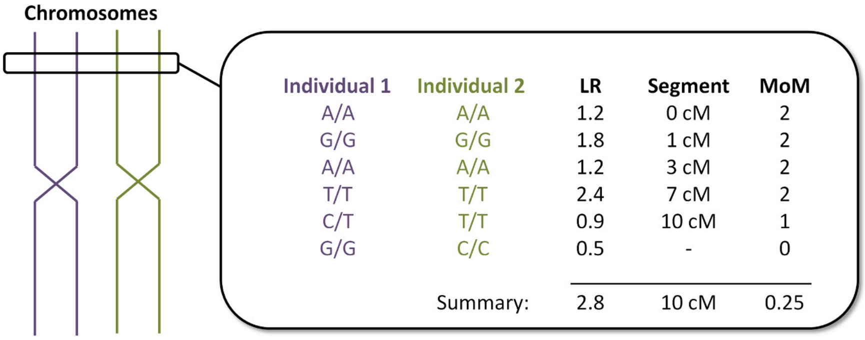
[7] See: https://www.forensicgenealogists.org

[8] ISOGG Wiki: https://isogg.org/wiki

Table 2
Analysis and SNP genotyping details of the four main DTCs and GEDmatch. Information has been compiled from the company websites as well as the scientific publications given in the table. When data was not available 'n/a' is given.

	23andMe	Ancestry.com	FTDNA	GEDmatch	MyHeritage
Website	www.23andme.com	www.ancestry.com/dna	www.familytreedna.com	www.gedmatch.com	www.myheritage.com/dna
Company founded	2006	1996	2000	2010	2003
Sells DNA tests	Yes	Yes	Yes	No	Yes
Launch of microarray-based relative-matching test	2009	2012 (US) and 2015–16 (33 other countries)	2010	n/a	2016
Accepts customer uploads from other companies	No	No	Yes. 23andMe, AncestryDNA, MyHeritageDNA	Yes. Uploads accepted from over 20 companies	Yes. 23andMe, AncestryDNA, FTDNA, Living DNA v1
Law enforcement uploads	No	No	Yes	Yes	No
International availability	50+ countries	34 countries	All countries except Sudan and Iran	Worldwide	All countries except Israel, Iran, Libya, Sudan, Somalia, North Korea, Lebanon and Syria
Database size	12 million	19 million	1.4 million	1.45 million	4.5 million [22]
Chip used	Customised Illumina GSA	Customised Illumina OmniExpress	Customised Illumina GSA	n/a	Customised Illumina GSA
Total SNPs	654,027	~700,000	654,027	n/a	654,027
Autosomal SNPs	621,575	637,639	621,575	n/a	621,575
X- SNPs	27,176	28,892	27,176	n/a	27,176
Autosomal DNA match thresholds	Option 1: 9 cM and at least 700 SNPs for one half-identical region; Option 2: 5 cM and 700 SNPs with at least two half-identical regions being shared	6 cM per segment before the Timber algorithm is applied and a total of at least 8 cM after Timber is applied	Option 1: 9 cM and 500 SNPs for one half-identical region; Option 2: 7.7 cM for the first half-identical region and a total of at least 20 cM (including the shorter matching HIRs between 1 cM and 7 cM); Option 3: 5.5 cM and at least 500 SNPs for the first half-identical region for about 1% of customers who come from specific non-European populations	7 cM. Default SNP count is set to vary dynamically. SNPs down to 3 cM can be seen in the One-to-One tool	8 cM for the first matching segment and at least 6 cM for the 2nd matching segment; 12 cM for the first matching segment in people whose ancestry is at least 50% Ashkenazi
X-DNA match thresholds	For half-IBD segments: Male vs male: 200 SNPs, 1 cM; male vs female: 600 SNPs, 6 cM; female vs female: 1200 SNPs, 6 cM; For full-IBD segments: 500 SNPs, 5 cM	Not applicable	1 cM and 500 SNPs for both males and females; matches must already meet the autosomal DNA matching criteria	7 cM. Default SNP count is set to vary dynamically. SNPs down to 3 cM can be seen in the One-to-One tool	Not applicable
Scientific publications on IBD detection	Durand et al. 2014 [59] and Henn et al. 2012 [20]	Ball et al. 2020 [21]			Petter et al. 2020 [22]

Adapted from Tim Janzen's Autosomal DNA Testing Comparison Chart in the ISOGG Wiki: https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart



Individual 1	Individual 2	LR	Segment	MoM
A/A	A/A	1.2	0 cM	2
G/G	G/G	1.8	1 cM	2
A/A	A/A	1.2	3 cM	2
T/T	T/T	2.4	7 cM	2
C/T	T/T	0.9	10 cM	1
G/G	C/C	0.5	-	0
	Summary:	2.8	10 cM	0.25

Fig. 1. An illustration of three approaches to infer relatedness between a pair of individuals, illustrated in their simplest form. In the likelihood ratio (LR) approach two competing hypotheses are compared and the LR expresses how much more likely the genotypes are given the first hypothesis. In the segment approach stretches of half-identical genotypes are compared and once opposite homozygotes are detected, the segment is terminated. The method-of-moments estimator (MoM) compares the individual genotype states and summarizes them over a large number of SNPs providing estimates of the kinship between the individuals.

2.1. *Exploratory approaches*

The exploratory approach benefits from being able to provide a measure of relatedness without any prior information. Briefly, it uses the observed genotype states and summarizes the number of shared alleles or shared stretches of alleles. Manichaikul et al. [27] describe a method to estimate the so called Cotterman coefficients using dense SNP data. Cotterman coefficients are summarized in the kinship coefficient and probability to share zero alleles IBD. A similar approach is implemented in PLINK [37]. Both can be seen as methods-of-moment estimators. Browning et al. [31,38,39], Gusev et al. [40] as well as Henn et al. [20] outline an alternative model whereby segments of shared DNA are identified, see Fig. 1. The simplest version of this approach utilizes dense SNP data to identify stretches of half-identical genotypes. A half identical stretch is terminated once opposite homozygotes are detected at a certain point. The length of the segment (or haplotype) is recorded as well as the segment's SNP number. The non-probabilistic version of the segment model requires two parameters, the segment length in centiMorgans (cM) and the number of SNPs in a segment.[9] If a segment exceeds a set threshold it is added to the total length of shared segments. Setting the threshold too low can potentially result in higher levels of false matches, whereas higher thresholds may eliminate true matches; although it should be noted that all likelihood-based forensic measurements must establish a threshold to balance false positive and false negative rates accordingly. In relationship tests a false positive result incorrectly includes an unrelated individual, while a false negative result excludes the true relationship, but may incorrectly suggest alternative relationships. Finding appropriate likelihood thresholds, with maximization of this cost/benefit trade-off applies to most statistical evaluations in forensic case work. The segment model has been adopted by all the major direct-to-consumer (DTC) genetic testing companies in different versions[10] and implemented in various freely available tools [31,37,40–45]. Variations of the segment model implement a pre-phasing step whereby the paternal/-maternal origin of each allele is determined and used to potentially improve the accurate detection of IBD segments [38]. The DTC AncestryDNA uses a version of the BEAGLE algorithm [46] to phase short pieces of DNA and subsequently uses phased haplotypes to identify what they term seed segments [21]. Information about the frequency of shared haplotypes can be used to further strengthen the weight of a segment match [36,44]. Haplotype frequency is taken into account in the matching algorithms at AncestryDNA where their so-called Timber algorithm compares segments with a reference panel and down-weights the genetic distance for regions which have unusually high levels of matching [21]. Haplotype frequency estimation could potentially help identify rare shorter segments shared through recent common ancestry [47]. A further refinement, which Browning et al. [38] refer to as probabilistic versions of the segment model, uses a statistical approach (hidden Markov model) to model the IBD states and compute LOD scores determining whether a particular segment is IBD or not. The probabilistic models are likely to perform better for the detection of shorter IBD segments, e.g. below 4–5 cM, but require significantly more computational power [48].

2.2. The likelihood approach

The likelihood approach has its merits as investigators are presented with a probability stating how likely the genetic data are, assuming hypothesis one (H1): the individuals are related as claimed vs. hypothesis two (H2): that they are unrelated or have an alternative relationship. Using likelihood comparisons to determine relatedness has traditionally been part of forensic and medical genetics for some time [49–51]. This approach requires the formulation of hypotheses to assess, for instance:

H1: Two individuals are full cousins.

H2: Two individuals are unrelated.

The likelihood is then computed by conditioning on each hypothesis separately. A likelihood ratio can be formed stating how much more likely or unlikely the observed genotypes are given hypothesis H1 compared to H2 [52]. Evaluating the likelihood is normally associated with computationally intensive algorithms [53,54] for dense SNP data and many typed individuals. For pairwise comparisons the algorithms can be condensed, and results obtained with minimum computational effort. Thompson suggested the use of a maximum likelihood approach (MLE) to estimate the relatedness coefficients for pairs of individuals [55]. However, this method is restricted to non-inbred individuals using unlinked markers. Weir expanded these ideas by including population substructure in the MLE model [56]. The inference of relatedness beyond first cousin level requires expanded marker panels of more than ~10,000 SNPs, and linkage must be accounted for [23,30]. This is in contrast to current forensic practice where unlinked STR or SNP markers are used, though recent progress suggests a move towards more expanded marker panels [57,58]. A maximum likelihood approach accounting for linkage requires the estimation of the relatedness coefficients in combination with inheritance patterns. Genealogical applications normally only provide a range of relationships rather than an exact level of relatedness. Therefore, a discrete grid of relatedness coefficients can be evaluated instead of a continuous optimization, i.e., the MLE approach can compute the likelihood of e.g., the twenty most common degrees of relatedness and then report the highest likelihood, or the top listed likelihoods if these have similar values.

The likelihood approach further benefits from being able to use reduced genotype data, normally comprising pruned genome-wide SNP data. A naïve approach uses only a minimum distance as the inclusion criterion. Closely located SNPs are expected to contain a high degree of redundant information, mainly through the association of alleles in a population. While a large proportion of SNPs with low minor allele frequencies on average convey little information, when a few rare variants are shared they can provide strong support for relatedness. Maximum information (i.e. heterozygosity) is achieved when the minor allele frequency for a bi-allelic marker is 0.5. Therefore, more intricate thinning procedures would utilize measures of allelic associations and population frequency data to prune SNP data.

Kling et al. [24] compared exploratory and likelihood approaches (including four degrees of relationships) finding that to identify distant relatives, they provide equal power while the likelihood approach tends to falsely include unrelated individuals as distant relatives to a greater extent than exploratory approaches. Note that Kling et al. used a naïve version of the segment approach, mimicking that of GEDmatch, and better performance would be expected for the more evolved versions [38,40,44,59,60]. As with the likelihood methods, the exploratory approaches do not provide an exact degree of relatedness, but a range of possible relationships which can be investigated through genealogical research. Ultimately, taking a case to court currently requires the formulation of hypotheses and a likelihood ratio which is then converted into a posterior probability stating how likely a certain hypothesis is, given all circumstantial evidence [52,61]. Exploratory approaches are currently only used in forensic analysis to generate investigative leads and are not presented in court, where STR profiling remains the universally accepted way to establish identity or the link between suspect and crime scene. However, Ge and Budowle [62] have suggested that a shift from STRs to dense SNP data could eventually occur which would require establishing new statistical methods in forensic genetics and acceptance as a secure system of identification by courts of law.

2.3. Limitations

In forensic applications, obtaining data for panels of >500,000 SNPs is not always possible, partly due to the nature of forensic samples but also due to the panels and platforms used in routine work. The exploratory approaches require very dense panels of markers to accurately determine relationships. Fig. S1 illustrates that in a small study performed, at least 56,000 SNPs are needed to determine first cousins, while siblings only require 29,000 SNPs. In contrast, the likelihood approach does not rely on as dense a set of markers as the exploratory approaches. It benefits from using allele frequencies to infer relationships and thus, in theory, a few shared rare variants can indicate strong support for relatedness. This could also represent a drawback if inappropriate frequency databases are used, as demonstrated in Kling [23]. Limitations in the number of genotyped SNPs could potentially be overcome by using imputation, described later. A further drawback of the likelihood approach is the need to

[9] The threshold on the number of SNPs in a segment is primarily defined to ensure sufficient marker density in any given region. Further, in a forensic setting, marker density cannot necessarily be ensured, for instance due to low quality DNA samples.

[10] ISOGG at: https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart

account for linkage disequilibrium (LD) when SNP numbers increase. Kling showed that the false positive rate (i.e. false inclusion of true unrelated individuals at various degrees of relationship) is heavily inflated if LD is not accounted for with SNP numbers exceeding 30,000, particularly in some populations [23]. In contrast, LD can be naturally incorporated into the segment approach where SNPs could be in LD (i.e. shared through distant population ancestry) in short segments, but when segments are longer, little LD is detected between their start and stop positions [63–65]. Browning et al. incorporated adjustments for LD in their segment model [38]. Chiang et al. [66] showed that many inferred segments 1–2 cM long actually result from conflation of a number of smaller segments of at least 0.2 cM or longer. AncestryDNA recently illustrated that some longer segments, even up to 50 cM, were identified to be shared by individuals from a common population.[11] They also showed a lack of concordance in matching in mother-father-child trios for inferred IBD up to 30 cM and a 50% discordance rate at 6 cM (Fig. 3.3 in Ball et al. [21]).

The question of how far inference of relatedness can reach was first addressed by Donnelly in 1983 [67]. Donnelly investigated the theoretical probabilities of two people of different degrees of relatedness sharing a portion of their genome identical by descent. This study found that in theory all second cousins should share some DNA identical by descent, but roughly 2% of all third cousins and 30% of all fourth cousins would share no detectable DNA relationship. This work further highlighted the limits of genetic genealogy and the important principle that not all genealogical relationships will be genetic ones [68]. The degree to which relationships can be detected using available genotype data was investigated by Huff et al. [45]. Using a maximum-likelihood method known as ERSA, they identified 80% of sixth- and seventh-degree relatives amongst 169 individuals. Henn et al. [20] investigated IBD sharing in a much larger dataset of over 20,000 individuals drawn from the 23andMe database and HGDP-CEPH panel. Using unphased data, it was possible to detect ~90% of third cousins and 46% of true fourth cousins. There is a considerable overlap between the distribution of shared DNA for distant relatives (see Table 1 in Balding et al. [34] and Ball et al. [21]), which is why DTC reports give ranges of relationships rather than precise inferences. The crowd-sourced initiative “Shared cM Project” (see Section 5) provides a good overview of empirically collected data submitted by DTC customers [69,70]. The use of whole genome sequence (WGS) data has the potential to further improve relationship estimations. Li et al. estimated that WGS data potentially increases the detection power for distant relationships by 5–15% compared with microarray data [71]. Al-Khudahair et al. [72] described the use of whole genome sequence data where distant relatives (8–9th degree) could be detected using very rare genetic variants. Section 3.2 further explores the expected and reported success rates using current databases.

The inference of relatedness is confounded by pedigree collapse and endogamy. Ralph and Coop [73] provided empirical data of the inter-relatedness of all Europeans within the last 1000 years. They found that two European individuals from neighboring populations share between two and 12 genetic ancestors from the last 1500 years and over 100 genetic ancestors within the last thousand years, with substantial regional differences in the level of sharing. They highlighted the difficulties of inferring the age of a single small segment of 10 cM and the impossibility of assigning a genealogical relationship. Gauvin et al. [74] found evidence of genome-wide sharing in the French Canadian population. Carmi et al. [75] found significant IBD sharing on segments over 3 cM and 5 cM in an Ashkenazi Jewish population and Gilbert et al. [76] found elevated levels of segment sharing in the Irish traveller population.

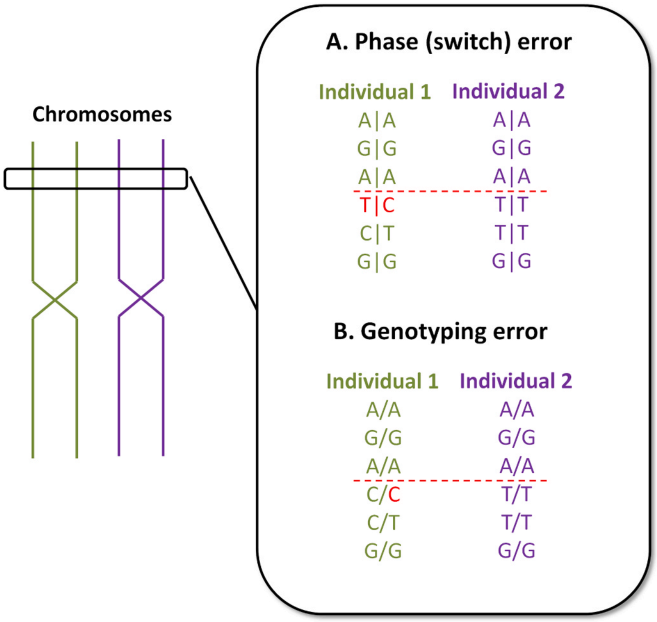


Fig. 2. Illustration of two types of errors. An IBD segment is prematurely terminated at the dashed red line with errors highlighted in red. (A) Phase errors occur when the chromosomes of an individual are separated into maternal and paternal origin and where the process of phasing (wrongly) switches chromosomes. (B) Genotyping errors occur either during the amplification process or at the bioinformatic genotyping level.

Henn et al. [20] explored the effect of endogamy in HGDP-CEPH populations. Very high levels of segment sharing, and therefore very recent common ancestors, were detected in Surui and Karitiana, (Amazonian populations which are essentially extended families). However, high levels of segment sharing were also detected in the much larger Kalash and Yakut populations, indicating the minimum segment length threshold used to analyze IBD needs careful calibration in populations with endogamy or recent bottlenecks [25].

IBD sharing on the X-chromosome was investigated by Buffalo, Mount and Coop [77] and a useful overview of the practical applications and limitations of X-chromosome matching for genetic genealogy is provided by Johnston (see X-DNA techniques and limitations in [78]) An X-chromosome match provided useful additional information in the Golden State Killer case, when a second cousin was found to have an X-chromosome match with the suspect DeAngelo.[12]

2.4. The impact of errors

Various errors can be introduced to SNP genotypes during the process of parsing variants. Such errors are broadly dividable into two subsets: technological errors and induced errors. Technological errors resulting in erroneously called genotypes can occur during DNA amplification and sequencing, or in the bioinformatics pipeline that performs sequence alignment or variant calling, see Fig. 2. Imputation and phase errors fall into the latter category. In the section on imputation we describe a small study where we investigate the errors introduced when inferring missing data. Furthermore, the process of phasing individual chromosomes can introduce errors [59,79], as shown in Fig. 2A. Using data from the 23andMe database, Durand et al. [59] estimated a genotyping error rate of less than 1% and a phasing error rate (using BEAGLE [46]) of less than 0.2%. AncestryDNA further found a phase error rate using the Underdog algorithm of 0.64% with a training set of 502,212 samples and suggested accuracy would improve with larger phasing panels [21].

Kling et al. demonstrated that the likelihood ratio approach is sensitive to all errors, even at low percentages (detectable differences down to

[11] See: https://blogs.ancestry.com/ancestry/2015/6/8/filtering-dna-matches-at-ancestrydna-with-timber/

[12] See: https://www.oxygen.com/crime-news/barbara-rae-vente-reacts-to-golden-state-killer-joseph-deangelos-guilty-plea

0.05%) [23] when these are not accurately modelled. Similarly, de Vries et al. [In submission, 2020] demonstrated that the segment approach is sensitive to wrongly called homozygotes for error rates as low as 0.5% (personal communication). One of the strengths of non-probabilistic versions of segment matching, where phasing is not used, is that it is only sensitive to wrongly called homozygote genotypes, which can prematurely terminate a shared segment. Durand et al. [59] suggest applying a haplotype score incorporating the phase and genotyping error rates. This score could be used as a post-processing step to filter spurious IBD segments. Other researchers have studied and incorporated error rates into their segment models [39,48], and most commercial segment matching implementations are believed to model for errors [20–22].

From a forensic perspective, many contact trace samples are likely to be of low quality and quantity, analyzed with low-depth whole genome sequencing, whereas database samples, commonly analyzed with SNP microarrays, are expected to have significantly lower error rates [80]. To illustrate the effect of genotyping errors and the impact on shared segments we conducted a small study using data from 1000 Genomes samples [81]. We simulated data according to the procedures detailed in Kling et al. [23] and induced errors in one of the genotypes at different levels (2%, 1% and 0.5%), see Supplementary File S2. The results are illustrated in Supplementary File S2, Fig. 1 where no model accounting for errors is used, which show that levels of detectable shared DNA drop rapidly with increasing error rate. At 2% error rates, a pair of full siblings share on average ~500 cM of detectable total segments compared to roughly 2800 cM without errors. Supplementary File S2, Fig. 2 contains an equivalent illustration when a single error per segment is allowed and shows a considerable improvement in terms of detecting broken segments. Furthermore, Supplementary File S2, Fig. 2 demonstrates an implementation of the error model presented in Petter et al. [22]. In our implementation, four homozygote errors per segment are allowed while simultaneously only retaining a match if a segment of above 6 cM without errors is detected. Fig. 3 further illustrates how errors affect the individual segment and indicates that for e.g. full siblings, a few long shared segments are split into multiple shorter segments. Some will disappear, failing to exceed the detection threshold, while others are accumulated into the total length of shared DNA.

2.5. The use of DNA mixtures

In contrast to single-source DNA samples, mixtures of several contributors are common in forensic samples. In terms of using mixtures as court evidence, there are various methods to estimate the evidential weight of a DNA sample [82–86]. Extending such analyses further, studies have examined the viability of using mixtures for familial searching [87–89] indicating feasibility even with common forensic STRs.

In the long-range familial searching process of IGG little has been scientifically documented on the analysis of mixtures. Greytak et al. [15] state that two-person mixtures with one contributor known, were successfully analyzed with microarray data but without exact details disclosed. Furthermore, a *Forensic Magazine* article[13] described the use of WGS analysis of a mixture and subsequent separation through conditioning on the victim's DNA profile, although also lacking details of the method. State-of-the-art methods in forensic DNA analyses use quantitative models where allele peak heights help infer individual contributor genotypes (termed probabilistic genotyping). In current IGG, the search is conducted with a single source DNA profile,[14] so a searchable profile must be obtained by deconvolution of the mixture, either by conditioning on known contributors or by combining a statistical model and information about the balance of allelic signals. As a consequence, the resulting profile used in the search has a level of uncertainty and the analysis benefits from estimation of the false/true positive rates affected by this uncertainty. Standard forensic mixture deconvolution incorporates the uncertainty into a statistical model to potentially allow a search. The current version of CODIS [90] software does not allow for quantitative or qualitative mixture models. However since CODIS allows export of the complete database, external software can be used for this purpose [91].

From a statistical point of view, the pedigree-based approach benefits from being able to consider different genotype combinations (and weights) in the calculations. Dørum et al. [92] demonstrated that linked markers can be used in a qualitative model allowing future expansion of marker panels. Exploratory approaches, on the other hand, rely on large numbers (and segments) of uninterrupted SNPs. IGG relies on the generation of a SNP profile with sufficient genotypes to be accepted into the databases to allow LE matching. The approaches would have to rely on a single deconvolution where the profile of the perpetrator is extracted instead of a more probabilistic approach.

Whole genome sequencing of low-level DNA tends to yield low mean coverage conveying little information on the exact level of individual contributors. However, a statistical model can be developed to extract a contributor in a mixture based on allele dosage (i.e. read counts). Fig. 4 illustrates a two-person mixture and how it is possible to extract the perpetrator based on a known contributor. Without using information on allele dosage, only homozygotes can be called with certainty. If the mixture is a homozygote genotype then the perpetrator must be a homozygote as well, disregarding dropouts, and therefore the second contributor's genotype is irrelevant. For heterozygote mixture genotypes, the perpetrator can be a heterozygote or homozygote for either of the alleles, potentially inferred using information from the second contributor. Inflating the number of erroneous homozygotes is quickly detrimental to genealogy searches, so potential solutions are to always infer a heterozygote genotype for the perpetrator, or to remove these ambiguous genotypes. The former can lead to an increase in the number of false positives, while the latter can potentially increase false negatives since fewer SNPs are called. If information on allele dosage is available, such information can be used if heterozygote genotypes contain a minimum number of reads. Raw data from microarrays contain intensity levels that potentially allow mixture contributors to be separated, as described by Homer et al. [93]. However, we do not recommend the use of such microarrays for forensic analyses (see Section 7.1).

We performed a small study where unrelated individuals from the 1000 Genomes Project were drawn at random in a pairwise approach. The genotypes were mixed (equal proportions) and deconvoluted using three different models, two qualitative and one quantitative, as outlined in Supplementary File S2, section B. Under the assumptions in our study, genotypes could be deduced with 99.9% accuracy when the quantitative model was used, with 4–5% of genotypes dropping out due to uncertainty in the deconvolution process, as shown in Supplementary File S2, Fig. 4. The qualitative models both resulted in an inflation of errors. We did not explore the impact of the deconvolution accuracy on the inference of relatedness but assume that it is minimal for the quantitative model, given the low error rates.

3. Genealogy research

3.1. Genealogical research

Genealogical research is a key component of IGG and generally the most time-consuming part of the process, though time spent on research will vary depending on many factors including closeness of the matches, the supporting network of matches, family size and availability of genealogical records. In a UK pilot study [94] genealogists solved one case which had matches with immediate family members within three hours, while they estimated more complicated cases with matches at third or fourth cousin levels needed 50–100 h of research. Some cases

[13] See: https://www.forensicmag.com/564243-New-Genetic-Genealogy-Technique-Can-Separate-DNA-Mixtures/.

[14] Strictly speaking, since biallelic SNPs are used, it can never be perfectly deduced if a profile is single source or not. However, allelic balances can give information on the number of contributors.

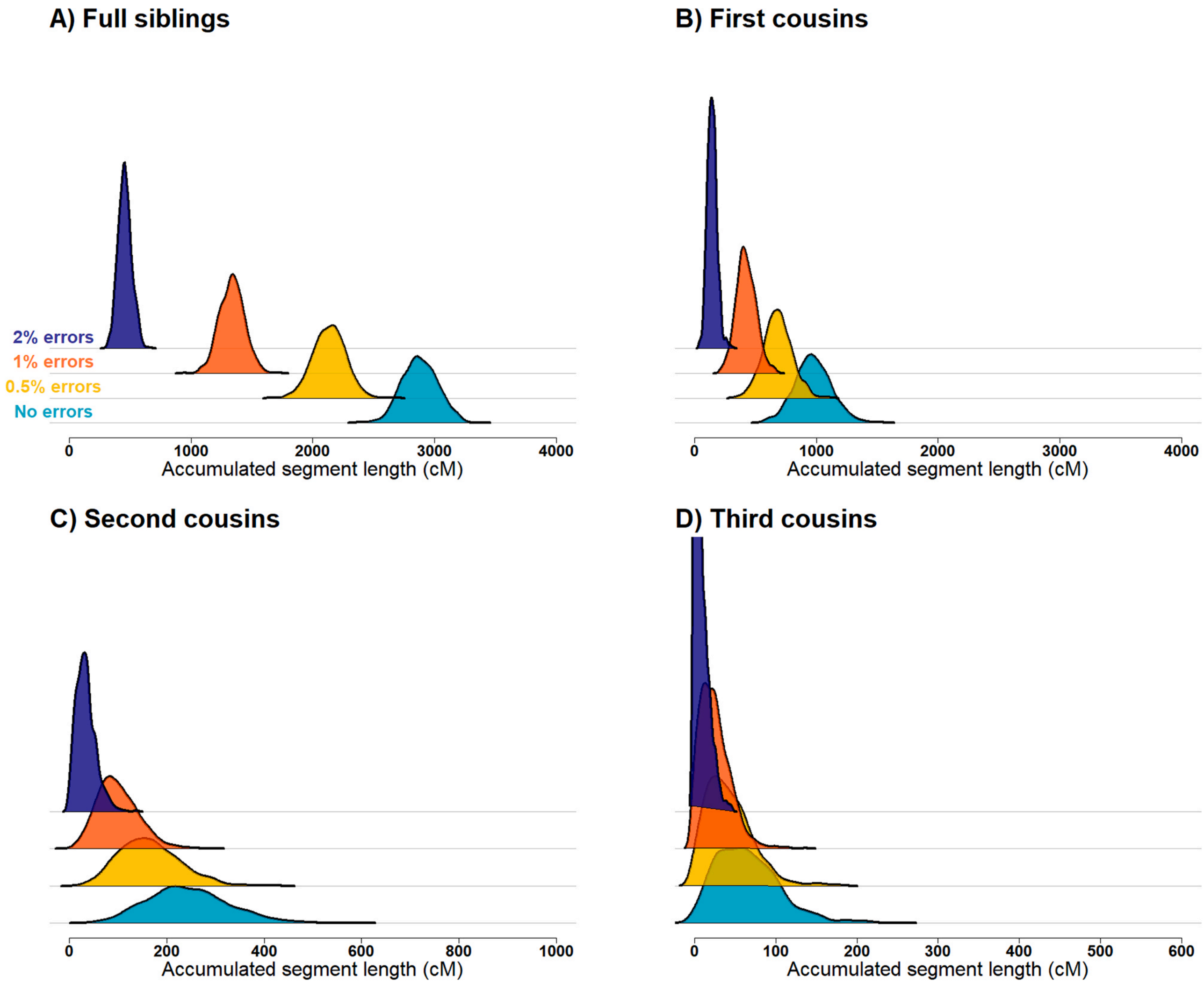


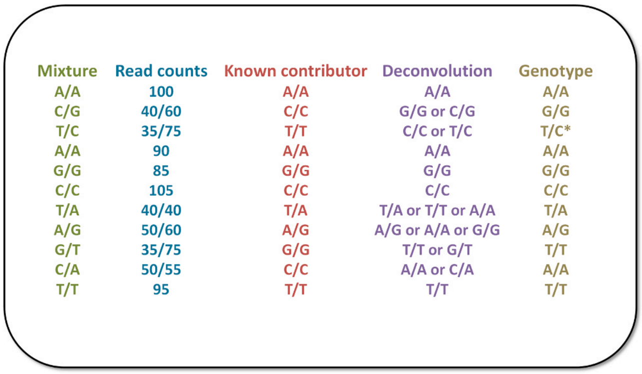
Fig. 3. Results from simulations of 1000 pairs of relatives. For each simulation, errors are induced in one of the profiles at an increasing rate (see legend). The number of shared segments (computed as the total length of shared cM) using 5 cM as a detection threshold, were counted and accumulated towards a total length (x-axis). (A) Full siblings, (B) first cousins, (C) second cousins and (D) third cousins.

analyzed by the DNA Doe Project required hundreds of hours of research by volunteer teams. IGG is only possible because of the large quantities of genealogical records from around the world which have been digitized and indexed in the last two decades. The Church of Jesus Christ of Latter-day Saints has been at the forefront of this process and provides free access to billions of worldwide records through its FamilySearch website (https://www.familysearch.org). The FamilySearch Wiki allows access to information on the availability of worldwide genealogical records and provides articles on the research process. Users can upload family trees, and the site hosts the FamilySearch Family Tree (claimed to be the largest family tree in the world). Commercial companies, such as Ancestry.com, Findmypast, Geneanet and MyHeritage, have also transcribed and indexed billions of records and provide subscription-based online access. These sites also allow users to upload family trees which can then be searched by other users. Therefore, it is now possible to easily access family trees, birth, marriage and death records, censuses, electoral registers, newspaper articles, wills and a variety of other historical records from many different countries. There are also many national and regional archives around the world with growing collections of digitized records which are freely available online. Research which previously took years and required visits in person to archives and repositories can now be done online in a matter of hours.

IGG involves researching not just historical records but tracing lines forward to the present day in what is termed descendancy research or reverse genealogy. This requires access to records on living people. Some modern records are available on the genealogy sites mentioned above but these records can be supplemented by searches on social media, particularly Facebook, which can offer a lot of information about living people and their family relationships. Online obituaries, particularly in the US, often provide complete lists of descendants and relatives of the deceased. People finder sites like BeenVerified and Intelius are particularly useful for US searches.

Successful genetic genealogy searches require not just easy access to genealogical records and a good understanding of how to evaluate genealogical evidence but also considerable experience of interpreting DNA evidence. There are university courses which provide a route to a career as a professional genealogist[15] and several organizations worldwide which provide credentials for genealogists [95]. However, many good professional genealogists are not accredited and have learnt through experience rather than a formal education programme. Genetic genealogy is a new discipline where best practice is being developed slowly through the collective experiences of those who are actively working in the field, many of whom are hobbyists. There are no official genetic genealogy

[15] See: https://www.mdpi.com/2313-5778/1/1/4



Mixture	Read counts	Known contributor	Deconvolution	Genotype
A/A	100	A/A	A/A	A/A
C/G	40/60	C/C	G/G or C/G	G/G
T/C	35/75	T/T	C/C or T/C	T/C*
A/A	90	A/A	A/A	A/A
G/G	85	G/G	G/G	G/G
C/C	105	C/C	C/C	C/C
T/A	40/40	T/A	T/A or T/T or A/A	T/A
A/G	50/60	A/G	A/G or A/A or G/G	A/G
G/T	35/75	G/G	T/T or G/T	T/T
C/A	50/55	C/C	A/A or C/A	A/A
T/T	95	T/T	T/T	T/T

Fig. 4. Illustration of a separation of a *Mixture* of two contributors (deconvolution) illustrated for a stretch of SNPs using a *Known contributor*. Data for the mixture is obtained from a sequencing analysis where sequence read counts are available for each variant. Therefore, the underlying deconvolution model could be probabilistic using the *Read counts* for each allele to deduce the most likely *Genotype* of the unknown contributor. The third marker highlights the inferred genotype with a star (*) indicating that the deconvolution is highly uncertain for this particular marker, with high allelic imbalance once the known contributor has been extracted, further suggesting that this marker should be blanked in the final genotype.

qualifications and no organization which can testify to an individual's ability to work on IGG cases. Many of the leading practitioners in IGG have had no formal genealogy training and have no accreditations. Accreditation with a genealogical organization is no guarantee that an individual has a sufficient level of expertize in genetic genealogy. This lack of professionalization makes it challenging for LE agencies wishing to employ a genetic genealogist to judge whether they have the relevant skills and expertize [2].

The IGG process starts with the upload of a SNP profile to one or more of the three databases where it is currently permitted: GEDmatch, FTDNA and DNASolves. Each company has different protocols for the use of their database by LE agencies, as described below.

The match lists are assessed by the genealogist who determines whether or not a genetic genealogy search is likely to be productive. If the query profile generates one or more matches at the second or third cousin level or closer, then the case is likely to be worth investigating. Second cousins are considered to be the "sweet spot" where identification should be possible [4]. However, much depends on the quality of the matches and whether or not the individuals can be identified through their username and/or e-mail address and by their family tree, if provided. The search will be more difficult if the query profile has ancestry from a country with limited availability of online genealogical records or where access to records on living people is more restricted.

Once the top matches have been identified, a check is made of the shared matches to identify genetic networks (clusters) of related matches. For example, second cousins share a set of great-grandparents in common and any matches which match both the query profile and a second cousin are likely to be related through a common ancestral couple in one specific quadrant of the family tree. The family trees of the shared matches are searched or built out to identify a common ancestral couple for all the people in the cluster. Descendancy research then traces the lines forward to the present day to identify candidates of interest. If additional clusters of related matches can be identified, then the genealogist will look for intersections (triangulations) between clusters, e.g., a marriage involving surnames from two distinct clusters. All the different genetic networks or clusters must be consistent with the identification with each match sharing the appropriate amount of DNA for the hypothesized relationship. However, because full siblings have identical ancestral family trees, genetic genealogy generally only ever narrows down the search to the offspring of a specific couple. It cannot determine which of a number of siblings is the suspect or the missing person, unless additional data for their descendants are available.

If the matches are all more distant (e.g. at third/fourth cousin level or beyond) the family trees can still be worked on, but it is often necessary to perform targeted testing of people identified through the genealogical research as possible closer relatives of the person of interest (e.g. second cousins). The individual is approached and asked to help with the investigation by taking a commercial genetic ancestry test and uploading the results to one of the databases which participates in law enforcement matching. The genealogist can then check that the individual matches the perpetrator in the expected way. Target testing thus helps to confirm that the correct branch of the family tree is being researched and narrows down the search pool, though the practice does have ethical implications, particularly if the DNA sample is obtained without the appropriate informed consent.[16]

The genetic genealogy research process is described in greater detail in Greytak et al. [15] and Thompson et al. [94]. The methodology is also demonstrated in the presentations delivered at the Institute for Genetic Genealogy conferences, with presentation recordings available online.[17] The DNA Adoption website has web pages describing the processes of tree triangulation and connecting trees.[18]

3.2. Success rates

As well as the quality and quantity of forensic DNA in a case, the chances of a successful identification depend on the size of the database plus the number and quality of the cousin matches. Edge and Coop [96] investigated the question of the expected number of genetic cousins at varying degrees in databases of different sizes to assess the chances of success. Using simulations and some simplifying assumptions, their findings indicate that in a database of one million individuals with ancestry from the same population, there is a high probability (>95%) of having at least one genetically detectable third cousin match sharing two or more DNA segments. At that time, the GEDmatch database had nearly one million profiles accessible to LE searches so this study demonstrated that the identification of Joseph DeAngelo as the Golden State Killer was within expectations and that there was a high chance that US individuals with European ancestry could be identified in a database of this size.

A study by Erlich et al. [97], using empirical data from the MyHeritage database (1.28 million SNP profiles at the time of study), found that ~60% of searches for individuals of European ancestry would result in a third-cousin or closer match with a total 100 cM or more shared segments. In 15% of the queries at least 300 cM in total was shared, signifying a second cousin or closer relationship which could provide highly informative investigative leads. They corroborated the results by performing similar queries on a smaller scale in the GEDmatch database which led to ~76% of cases with 100 cM or more shared and ~10% of cases with 300 cM or more shared. Erlich's study estimated that 75% of the MyHeritage database was of Northern European ancestry. The model presented in their study predicted that only 2% of a target population would need to be represented in a DNA database to provide a third cousin match for nearly everyone in the database.

Two studies have demonstrated the potential utility of IGG in a European setting and have validated the methodology. In a pilot study from the UK of ten volunteers, genetic genealogists were able to re-identify four of the ten individuals in the GEDmatch database (1.2 million SNP profiles at the time of study). One of the identified individuals had Indian heritage via St Vincent and the Grenadines, indicating the methods can potentially work for people of non-European descent if the right matches are available [94]. A study from Sweden generated an investigative lead in Croatia

[16] See: https://onezero.medium.com/how-cops-are-using-your-dna-to-catch-criminals-fe27a1d69e85.

[17] The recordings are available for a fee from: https://i4gg.org

[18] See: https://dnaadoption.org/first-timers/step-7/ and: https://dnaadoption.org/first-timers/step-8/

in the case of an unidentified male murdered in 2003 [17]. In a more recent case from Sweden, Daniel Nyqvist was identified as the suspect in a 2004 double murder of a young boy and a woman through matches with fourth cousins and as a result of extensive family tree building.[19]

The searchable portion of GEDmatch which is accessible for investigative purposes changed dramatically in May 2019 following concern amongst some genealogists and users after it was used for a search which was not covered by the existing site policy [98,99]. GEDmatch set to zero the number of 'kits' (herein, a kit refers to an individual's SNP dataset uploaded to GEDmatch, mainly produced and held by the DTCs) against which LE investigators could query and introduced an opt-in framework, where users own the choice to allow their SNP kit to be included in the portion that can be compared for investigative segment matching purposes.[20] Prior to the reset, ~700,000 of the one million or more GEDmatch profiles were available for investigative query. Private profiles, duplicate profiles, those with insufficient SNPs or excessive gaps in SNP coverage and specialized datasets (e.g., surname or ancestry groups) were all excluded from searches. GEDmatch was the subject of a security breach in July 2020,[21] but they have indicated to us that only a minimal number of users have since deleted their accounts, and the database continues to grow. In a presentation at the 31st International Symposium on Human Identification in September 2020,[22] Verogen, who acquired the GEDmatch database in December 2019, said 1.1 million users had uploaded 1.45 million DNA profiles. Over 285,000 users have opted in to LE matching and 83% of new users opt-in to LE matching. Verogen have made internal assessments to test the efficiency of the opted-in profiles for investigative searches. When a small cohort of known investigative SNP kits were compared internally against the opt-in portion of the database, and then against the opted-out portion, the opt-in portion provided equivalent potential leads to the opt-out database in ~80% of cases.

The GEDmatch database is dominated by users of European ancestry, particularly from anglophone countries. Table 1 gives the ten countries with the most GEDmatch uploads based on website analytics (data from Verogen, August 2020). The need for European GDPR compliance is also an influencing factor in the potential success rate as the consent process required EU users to opt in to use the database, following its acquisition by Verogen.

GEDmatch is now supplemented by the FTDNA database where the number of profiles available for LE matching is not known. If the FTDNA database has a similar number of profiles accessible to LE the combined reach of the two databases may be approaching 600,000, though some duplication is likely. In time, critical mass could be reached where nearly any US individual of European descent could potentially be identified through IGG [97].

In response to our questionnaire, Parabon NanoLabs said they had recorded a significant recovery in the informativeness of GEDmatch since the opt out was implemented in May 2019, but match rates had not quite reached the levels available before. However, they indicated the number of cases where investigative leads and actionable information can be provided has not significantly changed, but this often requires uploading to FTDNA as well as GEDmatch. The segment matching evaluations made by Parabon NanoLabs, before and after the GEDmatch LE access changes, are summarized in Supplementary File S3.

On 11th January 2021 Verogen updated the Terms of Service at GEDmatch.[23] The wording was ambiguous but appeared to allow unidentified human remains to be compared against the entire database.[24] The full implications of this change on the availability of profiles for law enforcement cases and the application of GDPR were unclear at the time of writing.

3.3. Ethical considerations of IGG

The use of IGG as an investigative tool raises many ethical and social issues [100,101]. The individual who makes their DNA available for law enforcement matching shares part of their genome with other close relatives and so their decision essentially affects their wider extended family who could potentially be involved in the investigation even though they have never taken a DNA test [2]. The use of surreptitious DNA testing to obtain confirmatory samples from the suspect also raises ethical issues, especially as in some cases the police have put multiple family members under surveillance to obtain these samples. The international nature of the consumer DNA databases and differing approaches to punishment raise ethical and human rights issues, particularly with regard to the death penalty which is still used in a minority of countries and in some US states.[25] The use of IGG to identify and prosecute the mothers of abandoned babies has also been cited as a cause for concern, particularly in jurisdictions where there are no infanticide laws allowing for more lenient and compassionate treatment of mothers.[26] Another emerging ethical issue is that of post-mortem privacy which is not currently protected by law [102]. Advances in technology are now making it possible to extract DNA from hair samples and artefacts of the deceased such as letters or razors.[27] Genealogists are interested in testing the DNA of deceased relatives to help with their family history research, but should they have the ability to make a deceased relative's DNA profile available for LE use? What happens if the descendants have conflicting views on such sharing? Qualitative research looking at the views of UK stakeholders found that there was considerable support for the use of IGG, but many interviewees commented on a range of social and ethical concerns and expressed the need for independent regulatory oversight [18]. While interviewees all expressed the importance of individual informed consent, it was found that it is not an ethical panacea and there is a need for a more societal approach to consent in consultation with the public [103]. We have highlighted some of the key ethical and social issues discussed in the literature which we feel are important, but it is outside the area of expertize of the authors and beyond the scope of this paper to engage with them in depth. Much more research is needed on all these issues by bioethicists and social scientists in consultation with stakeholders and the general public in order to establish a suitable ethical and regulatory framework for the responsible use of IGG.

4. Official guidelines for use of genealogy data in investigative practice

The US Department of Justice (DoJ) released an Interim Policy on Forensic Genetic Genealogical DNA Analysis and Searching in November 2019. The "scientific community and other interested parties" were encouraged to send comments to the FBI [104]. The policy clarifies that

[19] See: https://www.thetimes.co.uk/article/genealogist-uses-ancestry-website-to-track-down-knife-killer-m60rs0j2l

[20] See: https://www.nbcnews.com/news/us-news/police-were-cracking-cold-cases-dna-website-then-fine-print-n1070901

[21] See: https://www.nytimes.com/2020/08/01/technology/gedmatch-breach-privacy.html

[22] See: https://www.ishinews.com/events/gedmatch-a-data-driven-platform-for-forensic-intelligence/

[23] See: https://www.gedmatch.com/Documents/tos_20210111.html

[24] See: https://www.facebook.com/DNADoeProject/posts/2815513718707395

[25] See: https://dnaandfamilytreeresearch.blogspot.com/2019/05/civil-liberties-vs-greater-good.html

[26] See: https://www.watersheddna.com/blog-and-news/mental-health-awareness-baby-doe-cases and https://futurehuman.medium.com/dna-is-now-solving-decades-old-newborn-killings-67dd0f9ccf82

[27] See: https://thegeneticgenealogist.com/2018/11/19/testing-artifacts-obtain-dna-evidence-genealogical-research/

the investigative agency "must have pursued reasonable investigative leads" but it did not make specific recommendations about the need to clear testing backlogs or the need to use familial searching first before resorting to genetic genealogy. The SWGDAM (the Scientific Working Group on DNA Analysis Methods) in the US convened a working group to publish a statement on genetic genealogy and published an Overview of Investigative Genetic Genealogy in February 2020.[28]

Both the DoJ and SWGDAM recommendations emphasize the importance of a 'CODIS first and last' approach in investigative practice. The DoJ policy states: "before an investigative agency may attempt to use genetic genealogy, the forensic profile derived from the candidate forensic sample must have been uploaded to CODIS, and subsequent CODIS searches must have failed to produce a probative and confirmed match". They then emphasize that a CODIS search must complete the investigation, stating: "a suspect shall not be arrested based solely on a genetic association generated by a genealogical service. If a suspect is identified after a genetic association has occurred, STR DNA typing must be performed and the suspect's STR profile must be directly compared to the forensic profile previously uploaded to CODIS". As DNA analysis techniques progress there will eventually be situations where SNP data sufficient for a genealogical analysis will be generated from evidential material where an STR profile has not, e.g., where a hair shaft at a crime scene is submitted for specialist analysis outside of routine crime laboratory testing regimes. At this stage, which may have already been reached, the DoJ and SWGDAM guidelines must be reconsidered to address the way identity is established using SNPs in forensic cases without an STR profile from the crime scene.

With regard to what is described as 'investigative caution' concerning the behaviour of investigators in being transparent about the purpose of relative searches made by genealogical analyses, they state: "Investigative agencies shall identify themselves as law enforcement to genealogical services and enter and search genetic genealogy profiles only in those service suppliers that provide explicit notice to their service users and the public that the law enforcement may use service sites to investigate crimes or to identify unidentified human remains". Furthermore, when obtaining new DNA samples they state: "an investigative agency must seek informed consent from third parties before collecting reference samples that will be used for genealogy, unless it concludes that case-specific circumstances provide reasonable grounds to believe that this request would compromise the integrity of the investigation". The SWGDAM recommendations largely echo those of the DoJ, by saying a CODIS search in state or national databases should be made before instigating genealogical analyses and a CODIS search should conclude the investigation to complete the exclusionary/inclusionary process. On public consent for LE access, SWGDAM state: "policies/procedures should be established which consider applicable privacy policies and the database provider's terms of service, a level of transparency of techniques employed, and maintenance of the public trust".

The UK Biometric and Forensics Ethics Group recently published a report on investigative genetic genealogy which covers the feasibility of using the technique in the UK and ethical issues arising from its use.[29] The National Police Chiefs Council in the UK currently recommends against use of genetic genealogy databases.[30] Forensic scientists in Australia have published a working paper on operationalizing forensic genetic genealogy in an Australian context [105]. Following the resolution of a recent double murder in Sweden assisted by IGG (see above), public pressure to use the method in other cases has emerged. The double murder case was selected as a pilot study, initiated by the Legal Affairs Department at the Swedish Police Authority, to evaluate the suitability of IGG from a Swedish perspective and examine its compliance with current Swedish laws. The experiences from this pilot are currently being evaluated, involving technical, legal and ethical aspects.

5. Direct-to-consumer testing

Most current discussions of genetic genealogy describe four main DTC companies: AncestryDNA; 23andMe; MyHeritage; and FTDNA, each offering SNP microarray-based insights into an individual's health risks and/or ancestral roots, plus the opportunity to find links to previously unknown relatives that match for a pre-set minimum proportion of chromosomal segments. Each company uses a slightly different approach to detect putative IBD segments, commonly without disclosing all details about the exact implementation of their algorithm. They each apply different thresholds for declaring a match, but none report matches that share less than 7 cM. With the limitations of microarray technology, it is estimated that 20% of matches are false positives [22]. Most DTC's relative-searching analyses require customers to opt-in. AncestryDNA and 23andMe restrict matching to customers who have directly tested with the company. FTDNA and MyHeritage permit the upload of raw SNP data from 23andMe and AncestryDNA to expand the potential number of links to relatives.

The DTCs provide lists of matches and the suggested range in which the relationship might occur. The matches only provide a rough guideline and the genealogist makes further interpretation of the most probable degree of relatedness based on genealogical information and the related genetic network of matches. The analytical tools provided by the DTCs for estimating relationships can be supplemented by additional tools. The Shared cM Tool on the DNA Painter website (https://dnapainter.com/) reports cM value ranges and averages. It allows the user to enter the total cM shared and generate a table of probabilities for the possible range of relationships (probabilities inferred from the AncestryDNA Matching White Paper [21]). The Shared cM Project collects and summarizes crowd-sourced data on the range of sharing for various degrees of self-reported relationship [69]. The project, last updated in March 2020, has almost 60,000 submissions for nearly 50 different relationships.[31] Although participants have the opportunity to state if endogamy is suspected in their own family tree, they may have underestimated the degree of endogamy occurring. Therefore, the average total shared cM and upper range limits collected by the project are likely to be inflated. Some outlying values were removed from undetected misattributed parentage and data entry errors. Nevertheless, the compiled values and their distribution as histograms of average total cM (excluding alleged relationships without shared DNA) provide valuable aids for the interpretation of segment sharing data and a useful point of comparison with the predicted relationships given by the DTCs and GEDmatch. It should be noted that since DTCs use different detection thresholds which change over time, these numbers are only rough estimates reflecting that particular method and parameters.

The four DTC's microarray compositions are summarized in Table 2. Note that ISOGG list 32 separate genetic testing companies, but we concentrate on the four with the largest customer databases. The two next largest DTCs are the Genographic Project and Living DNA. Although Genographic had more than one million participants, it ceased making analysis data available to customers in June 2020. However, many participants have transferred Genographic data to FTDNA.[32] Living DNA has a worldwide customer base but is focused on Britain and

[28] See: https://www.swgdam.org/publications and the publication: "Overview of Investigative Genetic Genealogy."

[29] See: https://www.gov.uk/government/publications/use-of-genetic-genealogy-techniques-to-assist-with-solving-crimes

[30] See: https://www.thetimes.co.uk/article/police-wont-use-genealogy-sites-for-cold-cases-vvk0rbhqg

[31] See: https://thegeneticgenealogist.com/2020/03/27/version-4-0-march-2020-update-to-the-shared-cm-project/

[32] See: https://learn.familytreedna.com/imports/already-tested-genographic-project-can-join-family-tree-dna/

Ireland. It introduced a relative-matching feature called Family Networks in February 2018, initially restricted to close matches.[33] The matching range was expanded in May 2020 to provide matches with more distant cousins, but the number of matches obtained is modest in comparison to those of the four main DTCs. Several companies, such as Dante Laboratories, Full Genomes Corporation, Nebula Genomics and YSEQ, now offer WGS direct to the consumer, and the cost of WGS services continues to fall. However, there is no database which can fully leverage the information contained in WGS data to infer relationships. Advanced users can extract specific SNP profiles for upload to GEDmatch. While FTDNA accepts SNP profiles generated from WGS from LE accounts, such uploads cannot be made by regular customers. In theory, customers could mimic the file formats for DTC microarrays to upload a WGS-generated SNP profile.

Each of the four main DTCs has specific rules of engagement for their interactions with LE investigators seeking to match forensic DNA with their customer's data. These frameworks are well covered in the ISOGG Investigative Genetic Genealogy FAQs[34] and we summarize their current positions along with the SNP testing features of each DTC below in descending order of SNP profile database size.

5.1. AncestryDNA

AncestryDNA (http://ancestry.com) is by far the biggest DTC in the genealogy field, with nearly 20 million SNP profiles. It provides autosomal SNP testing based on a modified Illumina OmniExpress – changed in 2016 from v1 with 682 K SNPs to v2 (~300 K underperforming or uninformative SNPs swapped, reducing total SNPs to 637 K). Although AncestryDNA's microarray includes X-chromosomal SNPs, these are not used in their analyses, but are available through raw data download to the customers. AncestryDNA previously offered Y-DNA and mtDNA tests but discontinued them in 2014. Company policy is "not to allow law enforcement to use Ancestry's services to investigate crimes or to identify human remains", but when a warrant or subpoena is issued, "data relating to the DNA of an AncestryDNA user will be released only pursuant to a valid search warrant from a government agency with proper jurisdiction", and "when we receive a request our team reviews it to make sure it satisfies legal requirements and our policies. If we believe the request is overly broad, we will try to narrow it to the extent legally permitted"[35] Although not permitting LE access to the DNA database, the genealogical records and family trees held by AncestryDNA's parent company Ancestry.com are used extensively in IGG searches. In some cases, target testing to narrow down the search pool is done first at AncestryDNA before uploading to those databases allowing LE searches.

AncestryDNA has been the most transparent in outlining the process used to identify IBD segments by releasing a white paper which describes the principles and processes well [21]. Fig. S2 summarizes the steps taken to define each segment match. SNP data is phased into sections of sequentially arranged alleles using an adaptation of BEAGLE developed into a more efficient algorithm called Underdog. A separate algorithm, Timber, handles haplotype frequency estimation from the millions of profiles this DTC holds.

5.2. 23andMe

23andMe (http://www.23andme.com) has more than 12 million users as of January 2021, over 80% of whom have opted in to participate in research. They currently offer one test (used for multiple analysis options)[36] using the Illumina GSA with additional customized SNPs providing Y-chromosome DNA and mitochondrial DNA ancestry reports. The GSA was preceded by four different configurations of the Illumina OmniExpress microarray (v1, 2007–8, indeterminate custom SNP set; v2 2008–10, 556 K SNPs; v3, 2010–13, 930 K SNPs; v4, 2013–17, 585 K SNPs). Of the four main DTCs, 23andMe is the only one to fully incorporate X-chromosome data in their relative-matching process. The emphasis of 23andMe is on trait and disease risk associations from the collective SNP data compiled by the company, with customers self-reporting their lifestyle/behaviour, disease histories and known characteristics, which in turn provides some input to the forensic phenotyping knowledge base [106,107]. 23andMe were the first DTC to introduce segment matching tests to link customers to unknown relatives on the 23andMe database. The initial segment analysis regime to identify so-called 'cryptic relatives' was published by Henn et al. in 2011 [20].

23andMe do not give any access to customer information from requests by LE authorities, stating: "use of the 23andMe Personal Genetic Service for casework and other criminal investigations falls outside the scope of our services intended use. However, 23andMe must respond to, and is expected to comply with, court orders, subpoenas, or search warrants for genetic and personal data. 23andMe state: "they would use every legal remedy possible" to challenge a request for such legally enforced access to 23andMe customer data.[37] 23andMe's transparency report includes details of LE requests for information and is updated on a quarterly basis. As of May 2020, it had received seven requests, all from US agencies, pertaining to 10 users, and all were refused, with no data passed to LE authorities (https://www.23andme.com/transparency-report/).

5.3. MyHeritage

MyHeritage (http://www.myheritage.com) is estimated to have over 4.5 million SNP profiles, and accepts free data transfers from 23andMe, AncestryDNA, FTDNA and Living DNA. They originally used the Illumina OmniExpress but moved to the GSA in 2019. They launched a new microarray-based health test in June 2019 and the GSA has been customized to provide ancestry and health informative SNPs.[38] Although X-chromosome data is available in the raw DNA data download it is not currently incorporated into the relative-matching service.

The MyHeritage database unwittingly provided the breakthrough match in the Golden State Killer case. The terms and conditions have since been changed and the company now specifically "prohibits law enforcement use of its DNA services" and states "we will not provide information to law enforcement unless required by a valid court order or subpoena for genetic information".[39] However, because MyHeritage accepts uploads from people who have tested at other companies it is theoretically possible that they are vulnerable to unauthorized LE uploads, though a high-quality and near-complete SNP profile would be needed to pass quality control checks. To prevent such potential breaches, Yaniv Erlich of MyHeritage and colleagues [97] proposed a system of cryptographic signatures for the raw data text files provided by valid DTCs but to date there is little interest in adopting such a scheme. Erlich et al. recently published [22] details on the new matching algorithm of MyHeritage which builds on a similar model to

[33] See: https://livingdna.com/blog/family-networks-a-new-dna-driven-matching-system-and-family-tree-reconstruction-method/

[34] See: https://isogg.org/wiki/Investigative_genetic_genealogy_FAQs

[35] AncestryDNA, Ancestry guide for law enforcement: https://www.ancestry.com/cs/legal/lawenforcement

[36] Customer options are ancestry and traits or ancestry and health. The 23andMe health service is only available in the US, Canada, UK, Denmark, Finland, Ireland, Sweden and The Netherlands.

[37] See: https://blog.23andme.com/news/our-stance-on-protecting-customers-data/

[38] See: https://education.myheritage.com/article/an-introduction-to-the-myheritage-dna-health-test/

[39] MyHeritage, Privacy policy, July 2020: https://www.myheritage.com/privace-policy/

AncestryDNA using shorter phased seed segments and extending them using unphased data. The study also contained details about models for errors without disclosing exactly what the standard parameters in the matching algorithms were.

5.4. FTDNA

FamilyTreeDNA (FTDNA) launched in 2000 and were the first company in the US to offer direct-to-consumer ancestry testing [108]. The initial focus was on Y-chromosome and mitochondrial DNA testing as a tool for genealogical research, with Y-DNA results focused on surname projects. FTDNA were the second company to offer an autosomal DNA test for finding relative matches with the Family Finder test launched in 2010.[40] FTDNA initially used an Affymetrix Axiom microarray with ~563,000 autosomal SNPs. They changed to the Illumina OmniExpress in February 2011 with ~710,000 autosomal SNPs.[41] FTDNA moved to a customized Illumina GSA in spring 2019.

FTDNA appears to apply a system of half-identical segment matching with unphased genotypes, although their algorithms are proprietary, and no technical details have been published. The original threshold for a match was set at 20 cM total shared and a minimum longest segment of 7.69 cM for 99% of customers and 5.5 cM for the other 1%. Thresholds were updated in 2016, comprising a reduced minimum shared cM total but at least one segment required to be 9 cM or longer.[42] Matching segments must have at least 500 overlapping SNPs.[43] Matches are reported in a list with information on total number of shared cM, length of the longest segment and the predicted relationship range. FTDNA provides relationship predictions in four ranges: immediate matches, close matches, distant matches and speculative matches.[44] Although the minimum segment size for a match is set at 9 cM, once a match has been declared all segments down to 1 cM are included in the cM total. The majority of these small segments are either false matches (pseudo-segments) because of the lack of phasing or they are not genealogically relevant. Genetic genealogists normally recalculate the total cM shared to exclude segments under 7 cM to obtain a more realistic number. Matches for Ashkenazi Jews are down-weighted to account for the underlying endogamy in the population, though the technical details of the algorithms have not been published.[45] Users can download a list of their matches and the shared segment data. The problem of small false segments is seen when viewing known relations from different generations, in the chromosome browser, as shown in Fig. 5A and B.

FTDNA do not include X-DNA in total cM shared and an X match is only reported when two individuals have an autosomal DNA match. Once an X-DNA match is declared, FTDNA reports X-DNA matches down to 1 cM. There is a high false positive rate with these small X-DNA matches which is partly explained by the low SNP density on the X-chromosome on current microarrays. The false positives are clearly evident when comparing the low number of male X-DNA matches, which are naturally phased, with the unusually high number of female X-DNA matches. A small study found major discrepancies between the number of male and female X-DNA matches.[46]

FTDNA has accepted autosomal transfers from other testing companies since 2013.[47] In March 2018 FTDNA announced it was collaborating with BC Platforms who would provide a solution for incorporating genotype data from multiple chips into the database and dealing with backwards compatibility of historical data.[48] No details of the methods used have been published to date.

FTDNA's ability to accept third-party uploads inevitably made them susceptible to unauthorized uploads from non-genealogical sources. In January 2019 it was revealed that the FBI had infiltrated the FTDNA database and FTDNA had agreed to collaborate and continue to provide FBI access.[49] However, this meant existing customers not wishing to make their profiles available for LE use were denied access to the matching database for their own genealogical research. Following a backlash, in March 2019 FTDNA allowed customers to opt out of LE matching. EU citizens were opted out to comply with GDPR rules but could choose to opt back in.[50] Concerns remain about the lack of informed consent for the remaining customers who were automatically opted in, as many were unaware of the FBI breach.[51] New customers worldwide can choose whether to participate in LE matching when they activate their kit. In December 2020 further details of the Golden State Killer case emerged and it transpired that FTDNA had tested the rape kit and allowed the FBI to upload the profile to the FTDNA database as part of a covert operation. The FBI had invoked a legal privilege to prevent the disclosure of this information, thus raising concerns about the transparency and accountability of the FBI.[52] The FTDNA database has over 1.4 million SNP profiles,[53] though the number available for LE matching is unknown.

LE agencies wishing to use the FTDNA database are required "to register all forensic samples and genetic files prior to uploading to the FTDNA database. Permission to use the service is only granted after the required documentation is submitted, reviewed, and approved." Permission to use the FTDNA database for law enforcement purposes is only granted "to identify the remains of a deceased individual" and "to identify a perpetrator of homicide, sexual assault, or abduction".[54] FTDNA works with US LE agencies but will consider working with agencies outside the US on a case-by-case basis. Gene By Gene (https://genebygene.com/forensics/), the parent company of FTDNA, has its own testing laboratory in Houston, Texas, which has established a forensics division performing DNA extraction and testing in house. LE uploads are also accepted for a fee when testing has been done elsewhere.[55] LE kits are not visible to other FTDNA users regardless of whether they have opted in or opted out, and LE agencies receive a more restricted match list than regular customers. However, and similar to MyHeritage, FTDNA is theoretically susceptible to unauthorized LE uploads seeking to gain access to the entire database rather than the restricted LE matching portion.

[40] See: https://thegeneticgenealogist.com/2010/07/19/a-review-of-family-tree-dnas-family-finder-part-i/

[41] See: the archive version of the FTDNA FAQs: https://web.archive.org/web/20110927060537/http://www.familytreedna.com/faq/answers.aspx?id=39

[42] See: https://thegeneticgenealogist.com/2016/05/24/family-tree-dna-updates-matching-thresholds/

[43] See FTDNA learning centre article: https://learn.familytreedna.com/autosomal-ancestry/universal-dna-matching/number-snps-ibd-segment/

[44] See: https://learn.familytreedna.com/autosomal-ancestry/universal-dna-matching/possible-relationships-family-finder-match/

[45] See: https://learn.familytreedna.com/jewish-dna-testing/autosomal-dna-research-challenging/

[46] See: http://blog.kittycooper.com/2014/12/small-segments-on-the-x-by-kathy-johnston/

[47] See: https://www.familytreedna.com/autosomal-transfer

[48] See: https://www.prnewswire.com/news-releases/gene-by-gene-selects-bc-platforms-to-enhance-its-world-leading-genomic-data-processing-services-675954313.html

[49] See: for instance https://www.buzzfeednews.com/article/salvadorhernandez/family-tree-dna-fbi-investigative-genealogy-privacy

[50] See: https://www.newscientist.com/article/2196433-home-dna-testing-firm-will-let-users-block-fbi-access-to-their-data/

[51] See: https://www.nytimes.com/2019/02/04/business/family-tree-dna-fbi.html

[52] See: https://www.latimes.com/california/story/2020-12-08/man-in-the-window

[53] See: https://ggi2013.blogspot.com/2020/02/how-big-is-familytreedna-database.html

[54] See: https://www.familytreedna.com/legal/law-enforcement-guide and https://learn.familytreedna.com/ftdna/law-enforcement-faq/

[55] See: https://www.theatlantic.com/science/archive/2019/10/genetic-genealogy-dna-database-criminal-investigations/599005/)

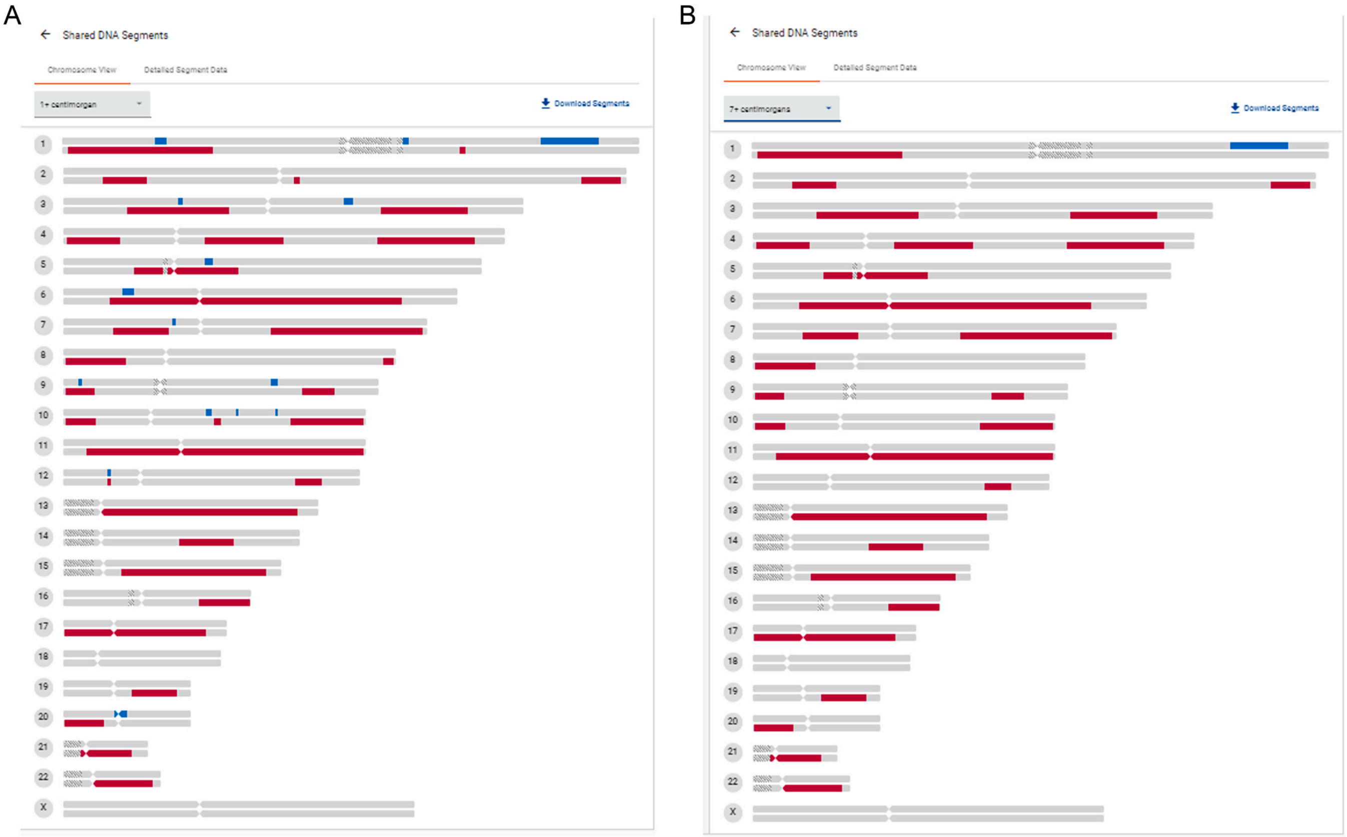


Fig. 5. An FTDNA chromosome browser comparison of an individual compared with their paternal grandfather (red) and a paternal 3rd cousin twice removed (blue). Fig. 5A shows only matching segments of 7 cM or more, while Fig. 5B shows all the unphased segments down to 1 cM in size. It is apparent that many of the small segments do not overlap with the segments inherited from the grandfather. The individual shares 1513 cM across 33 segments with their grandfather when all the segments down to 1 cM are included. When segments under 7 cM are removed, they share 1497 cM across 28 segments. The individual shares 60 cM across 15 segments with their third cousin twice removed when all the segments down to 1 cM are included, but when excluded, they share just one 27 cM segment containing 6900 SNPs.

6. Third-party services

In addition to the databases provided by the four main DTCs, there are two third-party services – GEDmatch and DNASolves – which do not sell DNA test kits but provide databases that accept uploads and can be used for LE matching. Below we address GEDmatch in particular since this portal has been the main entry point for LE up till now. Three additional third-party databases – DNA.Land (https://dna.land/), Geneanet (https://en.geneanet.org/) and Geni (https://www.geni.com/) – accept autosomal DNA uploads and could be vulnerable to unauthorized uploads, but these databases are all very small and less likely to be the focus of investigations.

6.1. GEDmatch

GEDmatch (https://www.gedmatch.com/) was founded in 2010 as a hobbyist website by genealogists Curtis Rogers and John Olson to supplement the tools provided by the DTCs and to help in particular with unknown parentage searches.[56] GEDmatch is a freemium website with both free and paid-for tools that perform a series of comparisons to other uploaded profiles and provide additional functionalities. GEDmatch was acquired in December 2019 by Verogen, a private forensic genomics company.[57]

The portal allows the user to search for matches with people who have tested on different platforms and at different testing companies. GEDmatch now accepts SNP profiles from over 20 DTC providers and is able to accept raw data from both microarrays and whole genome sequencing. They further allow users to upload DNA profiles obtained from ancient DNA (aDNA) samples or from the testing of artefacts of deceased people (e.g. testing the tooth or bone of a deceased parent or the DNA of a deceased relative obtained from a letter) as long as some quality criteria (i.e. number or density of genetic markers) are fulfilled. Artefact testing for genealogical purposes is still in its infancy but is likely to be a growth sector in the future.[25]. GEDmatch also has a tool available in the Tier 1 subscription which allows the user to combine SNP sets from multiple testing platforms into a 'superkit' to maximize the potential/reach of the search.[58]

GEDmatch has a dedicated law enforcement portal known as GEDmatch Pro (https://pro.gedmatch.com/) which was launched in December 2020. Law enforcement are now charged a fee to upload a SNP profile and LE uploads are no longer accepted on the main GEDmatch website. On 11th January 2021 Verogen subtly updated the site

[56] See: https://www.theatlantic.com/science/archive/2018/06/gedmatch-police-genealogy-database/561695/

[57] See: https://www.forensicmag.com/559058-Verogen-CEO-GEDmatch-Will-Be-Improved-Not-Changed/

[58] See: https://www.beholdgenealogy.com/blog/?p=2963

policy to allow unidentified human remains to be compared against the entire database.[59] Profiles uploaded to identify the perpetrator of a violent crime will continue to be matched only against the opt-in portion of the database. This change was made without the consent of the users and was a reversal of the decision taken in May 2019 to opt out the entire database from law enforcement matching and seek fresh consent from users. It is not clear how the distinction between offenders and unidentified remains will be enforced. It also not clear how Verogen can effectively identify LE users and prevent unauthorized uploads. Although LE uploads are expected to be declared as such, there is no regulation of this process outside of the guidelines and code of practice issued by the US DoJ and SWGDAM (see Section 4).

6.1.1. GEDmatch SNP uploads and analyses

Uploading a set of SNP genotypes, whether from a DTC raw data file or as compiled variants from a microarray or WGS analysis of forensic DNA, initiates the GEDmatch processing, which begins with the parsing of SNP data to ensure viability, followed by the assignment of a kit number. LE uploads are marked as research kits and so are excluded from comparisons made by individual users with their own kits and are not visible to other users. The SNP data are subjected to a process called tokenization, creating a compressed site-specific binary format which allegedly would not be possible to de-code in a security breach. As part of this process, health-related SNPs, SNPs with low minor allele frequencies and SNPs with no calls are removed in the tokenization. All comparisons in the database are done with the token files. Details of the tokenization process are given in Supplementary File S4. Once tokenized, the original upload is deleted, so it is not possible for SNP allele calls to be accessed directly either by a user or through malicious attacks on the site. However, if the phenotype of the query profile is known it is still possible to infer that matches have a particular trait of interest as demonstrated by Leah Larkin in a cystic fibrosis case study.[60]

The DNA File Diagnostic Utility can be used to verify the number of SNPs used in the token files. There are two different versions. The standard token file is used for all the GEDmatch tools with the exception of the One-to Many comparisons which uses the slim token file. To save processing time, heterozygous SNPs are removed from the slim token file. These SNPs would produce universal matches so do not provide any additional information.

GEDmatch data viability checks reject SNP numbers below 50,000 as insufficient for reliable segment comparisons, so potentially useful datasets from very degraded DNA may require troubleshooting of the DNA extract to increase call rates or a proportion of genotypes may be inferred by SNP imputation (see Section 8). Although imputation can ‘rescue’ scant genotype coverage when analyzing very challenging forensic DNA, in the most extreme cases uploading heavily imputed data leads to excessive numbers of false associations – commonly observed as a high proportion (up to 25%) of the reference profiles associating with the query file.

The One-to-Many tool is used to search for matches in the GEDmatch database. There are two different versions of this tool: the standard version and a beta version which has enhanced functionality and some additional features limited to Tier 1 subscribers. The One-to-Many tools look for all the SNPs in common between any two kits and then uses a simple system of half-identical matching to look for matching segments. The basic tools provide a list of 3000 matches, while additional matches can be viewed with the Tier 1 tool. Segments under 7 cM are excluded by default, but there is an option to set more specific thresholds both for length (cM) and number of required SNPs in each segment. The One-to-Many reports also include X-chromosome matches. The match list provides information on the length of the largest segment, the total cM shared, plus the number of generations between the pair suggested by their segment overlaps. The number of overlapping SNPs is also reported. If there is a low overlap, as happens for example when comparing a GSA kit with an OmniExpress kit, the overlap is marked in pink in the beta tool to highlight that caution is needed in the interpretation of results. All GEDmatch kits marked as ‘private’ or ‘research’ are excluded from the matching process. For LE kits the One-to-Many comparison is made with the subset of profiles permitting LE access. More details on the number of matching segments, their individual cM lengths/SNP numbers, and the bounding genome co-ordinates of these segments are given in the follow-up One-to-One searches made for each of the most closely related individuals. The One-to-One X-DNA comparison tool provides additional information on X-chromosome matches. Smaller segments down to 3 cM can be seen in the One-to-One tools.

The listed individuals in GEDmatch are identified in the match list by their kit number (assigned to each member’s uploaded SNP profile), self-designated name or alias, and an e-mail address. It is common for one member with a single e-mail to manage a large number of individual kits. The ‘People who match both, or 1 of 2 kits tool’ can be used to identify the list of matches shared between two kits which is the foundation of cluster building. It is possible to see who matches the query profile but clicking on their kit numbers reports the matches of each match. Therefore, it is potentially possible to build up an extended network of matches. An automated clustering tool is available as a Tier 1 tool. Finally, Q-matching is available in Tier 1 tools, in which the Q process considers the individual statistical characteristics of each SNP, gaps in coverage, and several other factors to provide a more evidence-based analysis of segments before they are reported.[61]

6.1.2. Data security and user privacy

As with all entities storing sensitive information, public genetic databases are particularly susceptible to breaches, either as a way to obtain the genetic data itself or to upload forged profiles potentially misleading LE investigators. The security of users’ kit permissions in GEDmatch is now centre stage in discussions about the consequences of the security breach of July 2020.[62] Malicious attacks on GEDmatch could seek to target information on three types of data: i. SNP genotypes which are not accessible online; ii. kit numbers and the associations produced relative to other kit numbers; and iii. users’ names or aliases and e-mail addresses. The July 2020 breach reset each user’s permissions so that private and research kits, including LE kits, became part of the segment matching comparisons and kit numbers were displayed among putative familial networks. Therefore, if queries were conducted in the 3-hour period of the attack, the ramifications are that putative associations were displayed temporarily to include kit number, name/alias and e-mail of each. Other information potentially accessed included the DTC used to create a kit. It is not clear whether other information was obtained from the GEDmatch attack, as the low numbers of targeted MyHeritage users would suggest the email addresses were obtained by running One-to-Many queries where the source DTC is indicated. Online user forums have recommended GEDmatch members delete their kits and re-upload with new kit numbers for the same SNP data, as a security measure which resets links between kit numbers and personal information. Since their recent acquisition of GEDmatch, Verogen have made repeated assurances that the SNP file reconfiguration process in GEDmatch makes a person’s genetic data secure from data mining attacks. These SNP data include the rs-number; GRCh37 coordinate; and allele calls. Health sensitive and low minor allele frequency sites are stripped

[59] See: https://futurehuman.medium.com/deleted-dna-data-just-reappeared-on-a-popular-database-e1f43587f7ec This article also highlights the second data breach at GEDmatch in January 2021.

[60] See: https://thednageek.com/cystic-fibrosis-a-case-study-in-genetic-privacy/

[61] See: https://www.gedmatch.com/Documents/Qdocs.pdf

[62] See: https://www.nytimes.com/2020/08/01/technology/gedmatch-breach-privacy.html

from the SNP file that is uploaded. This is an important point, as it has been reported that the One-to-One comparison tool can be used to mine genotypes by using artificial SNP datasets designed to find known relatives and estimate the genotypes when mismatches are found. Recent studies by Edge and Coop [109] and Ney et al. [110] have demonstrated an attacker could upload artificial files and attempt to extract the large majority of allele calls of other GEDmatch kits. They concluded that the visualizations and other results such as segment boundaries leak enough information for attackers to infer over 90% of the SNPs used in the comparisons. Therefore, this could potentially indicate a significant privacy violation for the targeted individuals. Verogen has stated that a series of measures are now in place, which effectively block such attacks.

The system of using kit numbers at GEDmatch to access DNA profiles and match lists brings a risk of sensitive accounts, e.g. LE accounts, being exposed as a result of user error, for example, if the kit number of a private or research kit is inadvertently shared or published. In the case of James Curtis Clanton, LE published the GEDmatch kit number of the crime scene profile in a publicly available affidavit for his arrest warrant, along with the initials of the people in the match list. As a result, all of the suspect's family members could easily be identified by anyone with a basic understanding of genealogy.[63] Although the kit was removed after LE personnel were alerted, many people would have had access to the match list up to this time, and such a breach could potentially compromise an investigation.

6.2. DNASolves

Set up in December 2019, DNASolves (https://dnasolves.com) is run by Othram (https://www.othram.com), and is intended to be a dedicated SNP database for LE use. As of March 2020, there were estimated to be several thousand profiles in the database.[64] DNASolves accepts SNP data from the four main DTCs and sequencing data in other formats (BAM/SAM, FASTQ or VCF). Some of the database plans were revealed in a podcast with David Mittelman, CEO and founder of Othram, on the Genialis website (aired March 2020).[65] Users contribute data to DNASolves solely to solve crime; there is no public-facing search and users cannot be matched with relatives or access anyone else's data but their own. People can voluntarily submit their name, date of birth and their parents' names as data points to help investigators. When LE agencies submit data for a case, their credentials are validated. The matching algorithm is similar to that of AncestryDNA (personal communication, David Mittelman). The database is currently a grassroots effort with a user group on Facebook where new features are discussed.[66] Although DNASolves is now actively accepting uploads there are no reported cases of it yet being used to produce investigative leads.

7. Technologies that generate a SNP dataset from forensic DNA

There are three ways that SNP genotype datasets suitable for relative searches can be generated from forensic DNA: i. using the same type of SNP microarrays as those adopted by the DTCs; ii. whole-genome sequencing (WGS) to obtain sufficient sequence coverage to reliably call heterozygote variant sites; iii. use of massively parallel sequencing (MPS) to perform targeted sequencing. This latter category can be further divided into amplicon-based methods (amplifying a smaller subset of SNPs with higher overall informativeness than those genotyped by a full SNP microarray) and hybridization capture methods. These different genotyping technologies are described below, and Fig. 6 illustrates the main steps included in the workflows. The genotyping technologies have different characteristics including analysis cost; availability of off-the-shelf assays; instrumentation requirements; data handling capacities required; and protocols available and optimized for the analysis of low quantity/low quality DNA. Note that it is possible to perform relative searches, which are not necessarily based on segment matching, but neither GEDmatch nor the DTCs currently use alternatives to the measurement of shared IBD segments. Therefore, when a subset of a typical DTC SNP dataset is assembled comprising 10,000, 20,000 or 50,000 SNPs, all uploads to GEDmatch are rejected due to insufficient data for IBD segment matching. This SNP density limit will potentially change in the near future as a result of initiatives by Verogen to develop smaller SNP sets for forensic analysis which will be suitable for their ForenSeq MPS platform, but informative enough to make reliable relationship inferences (see Section 7.4).

7.1. SNP microarrays

SNP microarrays have been the system of choice for over 15 years to genotype large numbers of SNP sites in a single workflow [111–113]. The basis of SNP allele detection with microarrays is to let fragmented sample DNA sequences hybridize to oligonucleotide sequences bound to a surface or to beads. In Illumina's BeadArray technology these oligo-sequences are designed to end prior to the SNP position, and the variant nucleotide(s) in the test sample are identified by single base extension. The Affymetrix (now part of Thermo Fisher Scientific) microarray technology uses fragmented DNA labelled with fluorescent dyes and then hybridized to a dense panel of allele-specific capture probes on the microarray surface. Hybridization of the DNA fragments containing the target SNP nucleotides, to one or both allele capture probes, produces dye signals detected by microscopic examination of the microarray surface corresponding to each genotype. There are multiple replicated capture probes per allele and SNP to ensure reliable consensus genotype analysis. In early microarray versions there was a degree of non-specific hybridization, but the sensitivity and reliability of the probe designs has improved markedly and, with optimized signal processing pipelines in place, microarrays deliver a very reliable system for SNP genotyping. The two most commonly used microarrays for SNP genotyping are Illumina GSA (654,000 target SNPs) and Illumina CytoSNP (850,000 target SNPs). Prior to the introduction of the GSA in 2016, the Illumina OmniExpress was the most commonly used microarray. Affymetrix provides the most commonly used alternative SNP microarray technology to the Illumina system. All current DTC analyses are based on Illumina OmniExpress or GSA microarrays with the addition of up to 50, 000 customized SNPs, with the exception of Living DNA (using Affymetrix). The CytoSNP microarray has more SNP targets, but unfortunately only provides a 104 K marker overlap with the GSA. Therefore, imputation is required to increase the overlap and facilitate relative searches (see Section 8). The CytoSNP microarray is used by Parabon NanoLabs both for phenotype predictions and for upload to genealogy databases [15]. Generally, SNP microarrays require 20–100 times more DNA than would be considered a standard input quantity of 1 nanogram (ng) for other forensic DNA tests, and this has hindered the technology's adoption for forensic analysis since it was first developed. Wendt et al. recently demonstrated in a titration experiment (1–200 ng of input DNA), using the Infinium Omni2.5Exome-8 chip, that high genotype concordance and call rates can be obtained down to 25 ng of input DNA [114]. However, the quality of the DNA used, in terms of degradation, levels of inhibitory substances and ratios of bacterial to human DNA in the sample, has a much greater effect on SNP call rates and their reliability (i.e. the genotype concordance recorded) than pushing DNA input levels below recommended quantities, as shown by the Bode Technology experiments outlined in Section 7.3. Two advantages with SNP microarray typing are the comparatively low cost and the ease of variant calling compared to the much more demanding bioinformatic analysis required by whole-genome sequencing.

[63] The affidavit is available online at: https://denver.cbslocal.com/wp-content/uploads/sites/15909806/2019/12/1216_DougCo-Cold-Case-Arrest_James-Curtis-Clanton.pdf

[64] See: https://onezero.medium.com/how-cops-are-using-your-dna-to-catch-criminals-fe27a1d69e85

[65] See: https://www.genialis.com/2020/03/04/dna-solves-the-oldest-cold-case-with-david-mittelman-podcast-18/

[66] See: https://www.facebook.com/groups/DNASolves

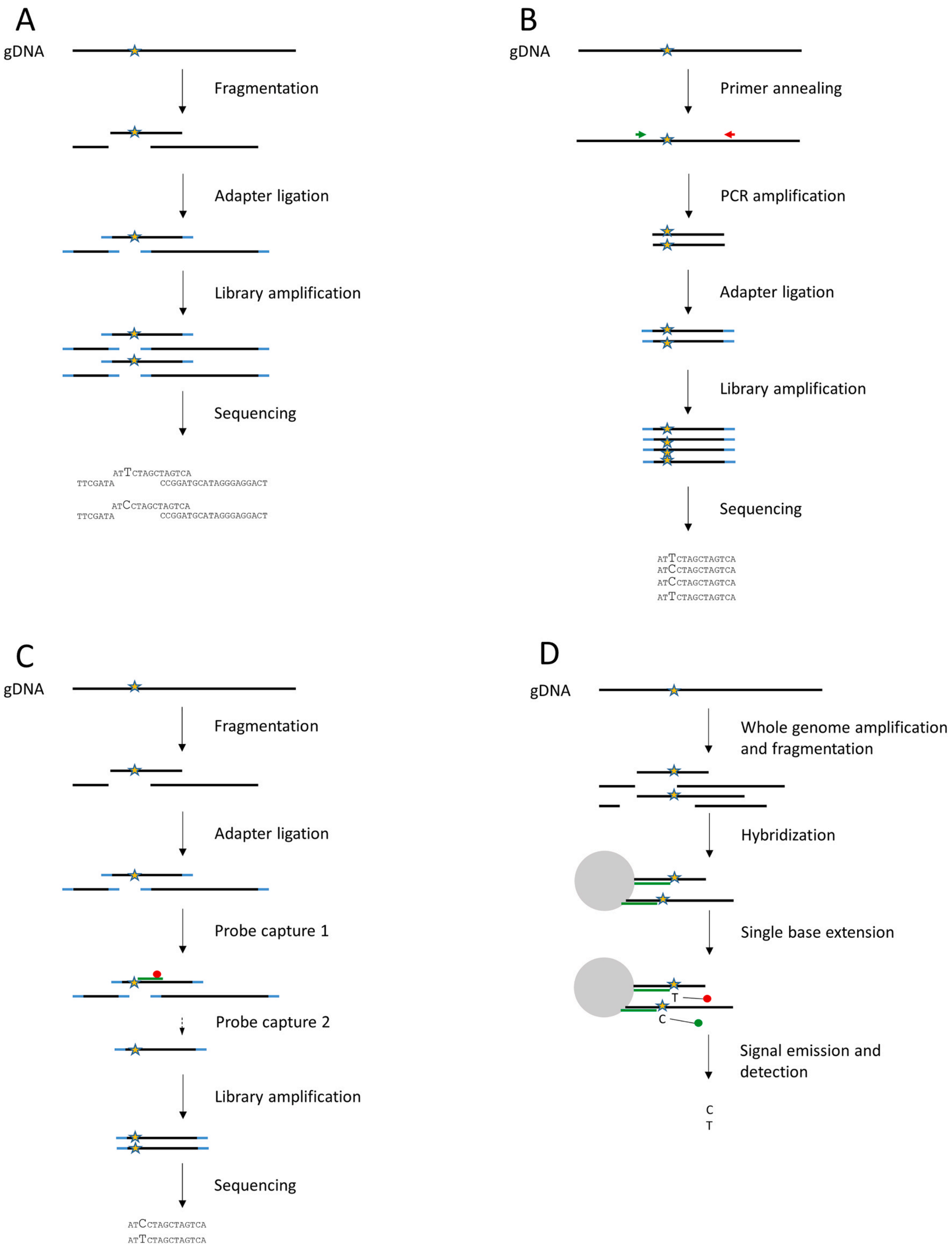


Fig. 6. Illustration of four different categories of technology for genotyping a large number of SNPs. In the diagrams the "star" represents a SNP position of interest. In each category there are a variety of different methods and the figure just illustrates the main steps. (A) Whole-genome sequencing starts with fragmentation of the DNA. This can be performed with sonication or restriction enzymes. This creates a library of random, shorter segments of DNA to which adapters are ligated. The adapters contain sequences for the sequencing step, the library amplification step and barcodes to aid sample multiplexing. After amplification of the library it is ready for sequencing. (B) Amplicon-based methods start with the enrichment of the targets using traditional PCR amplification. Adapters are then ligated to the amplicons and, after library amplification, the targeted region is ready for sequencing. (C) Hybridization capture methods start with fragmentation of the DNA. After adapter ligation the fragments of interest are captured via probe hybridization. The probes are then removed and after library amplification the sequencing can be performed. The sequencing step for A, B, and C is performed using standard MPS on any existing platform (e.g. Illumina NextSeq/NovaSeq, Thermo Fisher Scientific Ion GeneStudio). (D) Microarray typing (here illustrated by Illumina BeadArray technology) starts with the amplification of the whole genome (whole genome amplification, WGA) after which the amplified DNA is fragmented. The fragmented DNA is then hybridized to complementary sequences which are attached to beads. These oligo sequences are designed to end prior to the SNP position. Via single base extension the variant nucleotide(s) in the test sample can be identified using fluorophore-labelled nucleotides.

7.2. Whole-genome sequencing

The ability to sequence the whole genome of an individual in a viable single workflow has been refined much more recently than SNP microarray technologies. Initially confined to genetic research, WGS is now increasingly applied to clinical studies – where a rapid and practical sequencing system is required. The workflow is quite straightforward and starts by shearing the DNA into smaller fragments, achieved by, e.g. sonication. Sequencing adapters and sample indexes are then ligated onto these fragments. The library is amplified and after one or several purification steps it is ready for sequencing. Illumina and Thermo Fisher Scientific each offer whole-genome-scale sequencing systems with greatly expanded nucleotide reading throughput. Both companies have adapted the chemistry and nucleotide detection of their MPS targeted sequencing solutions now increasingly applied to forensic DNA analysis with the MiSeq-based and Ion S5 systems. Of the two options, Illumina have dominated the approaches to sequencing human whole genomes with the HiSeq X and NovaSeq systems, and each has been successfully applied to sequence-challenging forensic DNA samples ([17] and Section 7.3, respectively). There are kits and protocols available to analyze as little as 50 pg of DNA, e.g. using the ThruPLEX DNA-Seq Kit (Takara), although such a low amount requires pure, good quality DNA. One advantage of the WGS protocol is that the input in the library preparation is fragmented DNA, which increases the probability to obtain results from degraded forensic samples. Brandhagen et al. [115] employed a whole-genome shotgun sequencing approach on rootless human hair shafts and showed that complete mitochondrial genomes (mtGenomes) could be recovered from aged hair shafts in reasonable quantities. However, the sequencing data in their study was not sufficient to provide any reasonable depth of coverage across the nuclear genome. A probable cause was that their libraries were sequenced on a MiSeq with considerably less sequencing capacity than high-throughput NextSeq or NovaSeq platforms. A method has recently been developed to extract DNA from rootless hair shafts to create SNP genotype datasets from WGS for upload to GEDmatch. The methodology has not been published to date but has reportedly been used to identify two murder victims.[67]

An additional advantage with whole-genome sequencing is that, if high coverage data are obtained, it is possible to design and extract genotypes for any custom SNP panel. Thus, there is no need to target specific primers or probes. The disadvantages remain the high cost and computational workload when performing the bioinformatics and genotype calling.

7.2.1. Application of whole-genome sequence SNP genotyping to a real case

The application of HiSeq X WGS analysis of DNA extracted from the femur of a 2003 murder victim by Tillmar et al. [17] was documented in detail. The researchers used carefully constructed validation measures throughout the process to ensure there was sufficient sequence coverage for robust SNP genotype calling from these data. They were able to build a SNP dataset of more than 1.3 million variants that allowed efficient querying of the GEDmatch database. We describe the process used in detail next, since this was achieved by a forensic laboratory that was already investigating the case with conventional DNA analyses, rather than a commercial supplier who may not wish to disclose proprietary methods in such detail.

The sequencing pipeline followed three steps. First, DNA was extracted by standard phenol-chloroform methods from two grams of bone powder [116]. A critical part of this preparatory step was the checks made of the DNA quality prior to WGS. As well as quantitation with NanoDrop and checks of DNA integrity with Agilent TapeStation tests, MPS-based genotyping of an established forensic ID-SNP panel [117] ensured the DNA extracts would be suitable input for WGS. Prior genotyping by MPS also provided a concordance check of the SNP calls made by WGS at lower average sequence coverage. The MPS system tested the Qiagen QIAseq Investigator 140-SNP identification panel (131 SNPs used), set to a minimum coverage threshold of 200X and allele read frequency limits of 0.4–0.6 for heterozygotes and 0.1–0.9 for homozygotes. Additional GlobalFiler STR profiling was run alongside the MPS tests. Second, three library preparations were made in parallel from the single bone powder DNA extract with the ThruPLEX® DNA-seq 48S Kit (R400427, Takara). Each library used 3 ng of fragmented DNA prepared by sonication to produce fragments with a mean size of 400 basepairs (bp). Sequencing was performed with an Illumina HiSeq X instrument and v2.5 sequencing chemistry using paired-end sequencing and read lengths of 150 bp. Third, a bioinformatics pipeline was created to compile the three FASTQ files by alignment to human genome build hg19 and was recalibrated to adjust for potential misalignments caused by flanking indels, etc., with GATK. Duplicated, broken and non-specifically mapped reads were removed using Qiagen Biomedical Genomics Workbench v5.0.1.

For the GEDmatch relative searches 1378,481 SNPs were selected from the complete WGS variant dataset based on an optimum intersect of SNPs from the DTCs' adapted GSA and Illumina OmniExpress sets, comprising: 23andMe v5; AncestryDNA v2; and FTDNA/MyHeritage v1 microarrays. SNP genotyping quality was checked by applying four QC criteria to the HiSeq X sequence output: sequence coverage; allelic balance; Q-score and forward-reverse read ratio. Threshold values for these were, respectively: 10 or more homozygote reads, 5 or more per allele heterozygote reads; 0.5–0.7 heterozygote allele ratios; Q-scores higher than 25; and a read ratio of at least 0.2. From almost 3 billion reads, 86.7% were successfully aligned to the reference genome with a mean coverage of 32.2X. In the WGS-based SNP genotypes, 122/127 (WGS/MPS QC passed) cross-check genotypes passed their respective thresholds and were concordant. Approximately 75% of targeted SNPs passed the above thresholds, leading to a total of 1035,274 SNPs which were considered to be reliably called and compiled into the query profile for this case.

The GEDmatch relative search was marked for LE purposes and made across the full database of ~1.2 million reference profiles, i.e. before the opt-in setting was applied from member's choice to permit LE access. Searches returned several thousand putative relatives, but these were refined by choosing the top 100 matches that had >10 cM matching the query profile and 7 cM in common with others in the match list. This led to 36 putative relatives being analyzed further which created four clusters of individuals estimated to have been linked by their relationships to common grandparents. Information for some of the matched relatives indicated a Croatian origin and in fact, could be located more precisely to an area of ~40 km radius in NW Croatia. At this stage, meaningful investigative leads could be given to the police for their enquiries. These analyses are particularly important in establishing a benchmark for the validation of a new approach to forensic SNP genotyping and its application to IGG. They show the value of a forensic laboratory performing the WGS analyses who are well versed in applying multiple QC checks to novel techniques, as well as the diligence and depth of experience necessary for the handling of limited evidential material.

7.3. Evaluation of technology for forensic samples

There are few scientific studies on the suitability of each technology applied to forensic casework. Bode Technologies published a webinar ("Forensic Genealogy: Unlocking the Science of Genealogy") outlining an evaluation of Illumina GSA/CytoSNP microarrays and WGS for forensic DNA analysis in a valuable series of comparative tests [62]. Evaluations were based on the traditional measure of forensic sensitivity using dilution series of control DNA and artificial degradation of the same samples by progressively longer periods of sonication. Quality of SNP genotyping was measured by concordance rates (% of concordant genotypes with 250 ng input) and call rates (% of genotypes called), with input DNA quantities of 250; 50; 10; 2; 1; 0.5; and 0.25 ng extracted from blood and sperm. From the results presented in the webinar,

[67] See: https://www.nytimes.com/2019/09/16/science/hair-dna-murder.html

microarray technology (1x CytoSNP, duplicated GSA runs from two different analysis laboratories supplying data to Bode) was able to provide high call rates from 100% (250 ng−50 ng) reducing to 95% for 10 ng of blood-based input DNA, which only dropped to 90–95% at 2 ng input (1–0.25 ng not reported or analyzed). The lowest concordance rates were 89% in the 2 ng sample, but this lowest rate was an outlier value for general rates of 100–95% concordant genotypes, and one GSA analysis laboratory was consistently higher for both values indicating that established expertize and experience with handling low level microarray input affects the quality of results obtained. Apart from these differences, no discernible differences were detected between CytoSNP and GSA results. Sperm fraction DNA gave much lower call rates ranging from 90% down to ~65% (50 ng duplicates to 2 ng, respectively; no 100 ng analyses made), and concordance of 100–92% was similar to that from blood-based DNA. Degraded DNA produced by sonication and measured by Degradation Index (DI) had high concordance in microarray analyses of 100–98%. However, once DI values reached 6.6, 11.1 and 21 (from 1.4) call rates dropped to between 90% and 58%, meaning microarray technology struggled to detect damaged DNA with SNP target fragments of sufficient size to hybridize successfully.

WGS analysis was provided by a laboratory specializing in this technique using the Illumina NovaSeq 6000 system. Call rates for blood and sperm extracts were slightly lower with WGS than microarrays (97–87% in blood DNA, 93–41% in sperm fraction DNA), but concordance was consistently high at all input levels at: 100% (250 ng) to 98% (2 ng) in blood, and not dropping below 99% in all inputs (no 250 ng input). When very low levels of input DNA were examined this high sensitivity was maintained; duplicates of 2, 1, 0.5 and 0.25 ng of blood DNA had 92–91% call rates and were at or close to 100% concordance. This translated to GSA microarrays reaching $>$ 91% concordance with WGS data in blood and $>$ 81% in semen. Therefore, the overall trend in sensitivity measurements indicated WGS was more sensitive than microarrays and this sensitivity was more consistent—concordance dropped much less markedly as input DNA was reduced despite slightly fewer calls being made.

Following these experiments, an interesting evaluation of the SNP dataset informativeness obtained from each technique was performed using GEDmatch to examine known matches to kits from the control DNA used. The low template DNA SNP dataset obtained from the GSA microarray matched 9/13 kits in GEDmatch, compared with 11/13 with WGS, indicating more extensive SNP genotypes from WGS with minimal input DNA. The difference in the performance of WGS vs microarrays was much more marked when uploading SNP datasets from degraded DNA, with no kit matches amongst the top 18 with GSA-analyzed DNA having DI values of 6, 11 and 21; in contrast to WGS, with matches to all the top 18 with DNA at DI values of 1 and 21. Bode Technologies concluded from these studies that WGS is the system of choice for forensic DNA as it is a more accurate and sensitive SNP genotyping system for degraded DNA, matching or exceeding genotype call rates from microarrays.

7.4. Capture-based massively parallel sequencing technologies

Massively parallel sequencing (MPS) based methods using hybridization capture can genotype a large number of SNPs. These techniques have many similarities with whole-genome sequencing but, instead of including all fragments in the library to be sequenced, only those from regions of interest are captured and sequenced. The major advantages with this approach are that only relevant sequences are analyzed and deeper coverage is consequently obtained for those targets [118]. However, one disadvantage is that efforts are needed for the design of the probes (or "baits") used to capture the sequences of interest. Several different hybridization capture methods exist and have been developed. The main steps of the method are similar (see Fig. 6) but variation exists, especially in the way the targets are captured [119]. Some examples of commercial hybridization capture technologies are SureSelect (Agilent), HaloPlex (Agilent), Nextera (Illumina), myBaits (Arbor Biosciences), Twist technology (Twist Bioscience) and SeqCap (Kapa HyperExplore).

In hybridization capture methods the template genomic DNA is first randomly sheared by e.g. sonication or restriction enzymes. Sequencing adapters (which can also include sequences for library amplification, sample barcoding, etc.) can then be ligated to the fragmented DNA. The sequences of interest are captured using oligonucleotide probes. These synthetic probes are hybridized to the regions of interest and these hybridized regions are further captured by, e.g. magnetic beads, enabling non-targeted DNA fragments to be washed out. The probes are then removed from the targets prior to library amplification and sequencing. Sequencing can be performed using standard MPS approaches such as MiSeq/NextSeq/NovaSeq (Illumina) or Ion GeneStudio (Thermo Fisher Scientific). The number of targets, the desired depth of coverage per target, the level of sample multiplexing and other variables determine the level of sequencing capacity needed.

The main advantage with capture approaches, apart from offering high multiplexing capabilities, is that they are amenable to all sample types, from high-quality genomic DNA to severely degraded DNA (e.g. [120,121]). DNA from forensic samples and human remains is often of poor quality and, as a result of degradation, the DNA is already broken up into fragments, so such approaches are particularly suitable for forensic analysis. However, hybridization capture is more costly than amplicon-based approaches, but has been shown to be superior when testing mtDNA from human remains [122]. Many of the existing hybridization capture methods were initially optimized for research studies and clinical testing where large quantities of DNA are available. Nevertheless, several protocols have been adjusted for lower DNA input [123]. Hybridization capture methods have been used for a long time to study human aDNA, for which the DNA quantity/quality may be similar to forensic samples [124,125]. Very large custom-designed SNP panels have been developed and employed on a variety of different aDNA sample types. For example, Mathieson and colleagues [126] used a hybridization capture method to target approximately 1240,000 SNPs to analyze historical genetic variation among 230 West Eurasians dating between 6500 and 1000 BCE. Almost 600,000 of these SNPs were included on the Affymetrix Human Origins microarray. The samples in their study comprised teeth, petrous bones, femurs and other sources. Interestingly, they compared their data with that of a similar study using whole-genome sequencing, indicating that while the mean number of reads generated per sample with the capture approach was ~40 times lower, median coverage per analyzed SNP was ~4 times higher. Feldman et al. [127] used the same capture assay to successfully produce genotype data from Bronze/Iron age individuals.

Although most of the hybridization capture companies offer custom-made panels, we have not found any large-scale (>100 K SNPs) studies on forensic samples combined with genealogically relevant SNPs. However, Shih et al. [128] analyzed a custom SeqCap assay (Roche) to capture the mtDNA genome and a smaller number of autosomal SNPs (~400). They tested their assay on forensic samples (telogen hairs, mock stain samples, etc.) and obtained highly accurate SNP genotype data. We expect more studies and case reports to be published in the near future in which hybridization capture methods are applied to forensic analyses. Lastly, Ancestry.com launched an MPS-based AncestryHealth test in August 2020[68] using a large hybridization capture assay developed by Twist Biosciences.[69]

At the time of writing, a targeted SNP genotyping system, using MPS to generate data for SNP sets at a much-reduced scale of approximately ten thousand loci, was being developed by Verogen following their

[68] See: https://www.genomeweb.com/sequencing/ancestry-rolls-out-sequencing-based-health-offering-focused-common-conditions; https://news.thomasnet.com/fullstory/new-ngs-technology-powering-ancestryhealth-enables-fast-genetic-screening-40038209

[69] See: https://www.biospace.com/article/releases/twist-bioscience-launches-new-ngs-solutions-highlights-customers-at-2020-advances-in-genome-biology-and-technology-conference/.

acquisition of GEDmatch. The new assay, named the ForenSeq Kintelligence Kit (https://verogen.com/products/forenseq-kintelligence-kit/) was announced in January 2021, and comprises <10,250 SNPs which exclude medically important loci or those with low minor allele frequencies. The kit is based on the established ForenSeq library preparation approach using the MiSeq FGx forensic genomics system (validated for forensic use [129]). To develop the ForenSeq Kintelligence Kit, Verogen performed detailed bioinformatic analyses of the relative performance of component SNPs on various Illumina microarrays uploaded to GEDmatch, in order to gain knowledge of optimum candidates for smaller, forensically relevant SNP sets. Verogen will use a new IBS (identical by state)-based analysis tool supporting data from the new assay and used in the LE portal. The advantage of developing a 'built for purpose' SNP set for kinship analysis is not only an improved performance with challenging DNA (expected to work on forensic samples with DNA concentrations at sub-nanogram levels), but it is also feasible that SNP details can be encoded at each stage of the genotyping and query processes enabling better protection of kits in GEDmatch used for investigative purposes.

8. SNP genotype imputation

As discussed previously, DTC companies as well as scientific studies use a variety of different SNP microarrays and their marker configurations can change over time. Although some microarrays have a large proportion of overlapping SNPs, others have a considerable amount of non-overlapping SNPs which may reduce the power in database searches and segment analyses. One example is the transition from Illumina's OmniExpress to their GSA microarray which have fewer than 200,000 SNPs in common.[70] Missing data may also be the result of low-quality or degraded DNA. A way to increase the number of genotypes, and increase the proportion of overlapping SNPs, is to predict the missing genotypes with a method known as genotype imputation. Genotype imputation may also be relevant to apply when an analysis of DNA of low quantity and/or quality results in a large proportion of missing genotypes. A database search may be impossible to conduct if the number of missing SNPs is too large.

The aim of imputation is to predict the genotypes for SNPs not directly genotyped in a sample. One of the first studies using genotype imputation was in connection with the identification of genetic risk variants for type 2 diabetes. The study compared their results with those from similar studies conducted with different genotyping microarrays [130]. Since then, imputation has become a standard analysis tool for various purposes and has been used mainly in the field of medical genetics [131]. Some of the most common applications of imputation have been to increase the power obtained from genome wide association studies (GWAS) [132], to improve the resolution in fine-mapping studies (e.g. [133]) and to facilitate meta-analysis by combining data from different sets of genotyping microarrays into one unified set of SNPs (e.g. [134]). Although rare, there are also examples of genotype imputation for forensic STR typing purposes. Edge et al. as well as Kim et al. recently published two studies in which they demonstrated that a standard STR profile can be used to impute genome-wide SNP data (and vice versa) [135,136].

The underlying principle of genotype imputation is that any two individuals, including those who are apparently unrelated, will share short segments of DNA from a distant common ancestor. Factors like high levels of linkage disequilibrium (LD) and low recombination rates within small stretches of chromosomal segments will conserve haplotype variants through many generations. Shared segments can be identified if the observed genetic variants in the studied individual are compared with variants from a panel of reference individuals. From these shared segments, missing data in the sample can be predicted based on the observed genetic variants in the reference individuals. In practice, the genotype data (for both test individuals and reference individuals) is first converted into haploid format (i.e. haplotypes) by phasing methods [137]. There is a wide range of phasing software and many of them now combine phasing and imputation. The principle of phasing is illustrated in Fig. 7, and population haplotype frequencies are used to probabilistically estimate the most likely haplotype configuration. Many of the phasing models use hidden Markov models (HMM) for this inference.

Once phasing is completed, the missing variants in the test sample can be predicted from the variants present in the reference individuals with matching haplotypes (see Fig. 8). A studied haplotype will be a mosaic of the reference haplotypes where changes may represent historical recombination events, but differences may also represent historical mutations, gene conversions and genotyping errors. Most of the imputation methods also utilize an HMM framework and may differ in the parameters and the setup of the HMM [131]. A key component in the development of new methods and models for phasing and imputation is to decrease the computational burden to better handle larger reference data sets and to speed up the computations without loss of accuracy. A selection of available software is presented in Table S1. The genotype predictions are not always perfect, and many of the programs provide a prediction probability along with the imputed genotype which corresponds to the uncertainty of the imputed variants [138].

The performance of imputation depends on several factors. For example, the imputation error rate increases as the minor allele frequency decreases [138,139]. The reason for this is that rare alleles will be observed less often in the reference data and they tend to have lower levels of LD with common variants, which increases the uncertainty in the imputation [140]. Other factors that affect the error rate are the number, density and quality of the observed genotypes in the DNA analysis [46,141]. An additional important factor is the size and the population origin of the reference panel. Larger reference panels have increased imputation accuracy, and since genotype imputation depends on finding haplotype segments shared between reference and target haplotypes, a matching reference population is relevant to use [131,138,142]. Having a reference panel with very little genetic similarity with the test sample can decrease the imputation accuracy. At present, several public or partially public reference panels exist and include the HapMap project [143], 1000 Genomes phase 3 [81] UK10K project [144,145], and Trans-Omics for Precision Medicine (TOPMed) [146]. The TOPMed project includes more than 100,000 sequenced samples.

What level of accuracy can be expected in practice? Based on the factors outlined above, it is hard to be certain, but for the microarrays and SNPs included in genealogy testing a reasonable estimate may be ~1% or less [21]. However, the commercial companies have databases of millions of customers and will therefore have much larger reference panels for imputation than are currently available to academic researchers. For illustrative purposes we performed a genotype imputation test on a SNP profile from one of the co-authors. The genotypes of the

	Genotypes	**Phase alt A**	**Phase alt B**
SNP1	A/G	A G	A G
SNP2	C/T	C T	T C
Haplotype frequency		0.12 0.04	0.54 0.30
Posterior probability of haplotype pair		2*0.12*0.04/ (2*0.12*0.04+ 2*0.54*0.30)= *0.028777*	2*0.54*0.30/ (2*0.12*0.04+ 2*0.54*0.30)= *0.971223*

Fig. 7. Phasing of genotype data into haplotype format using a probabilistic approach. Assume that genotypes have been observed for two SNPs ("SNP1" and "SNP2") and they require phasing into haplotypes. In theory, there are two pairs of phase alternatives (alt A and alt B). If there is population data with haplotype frequencies it is possible to estimate the probability for each pair of phase alternatives. In this example, HWE is assumed and the result shows that variant B is the most probable phase alternative given the observed genotypes and underlying population data. This example is based on a similar example in Browning et al., 2013 [38].

[70] See: https://isogg.org/wiki/Autosomal_SNP_comparison_chart

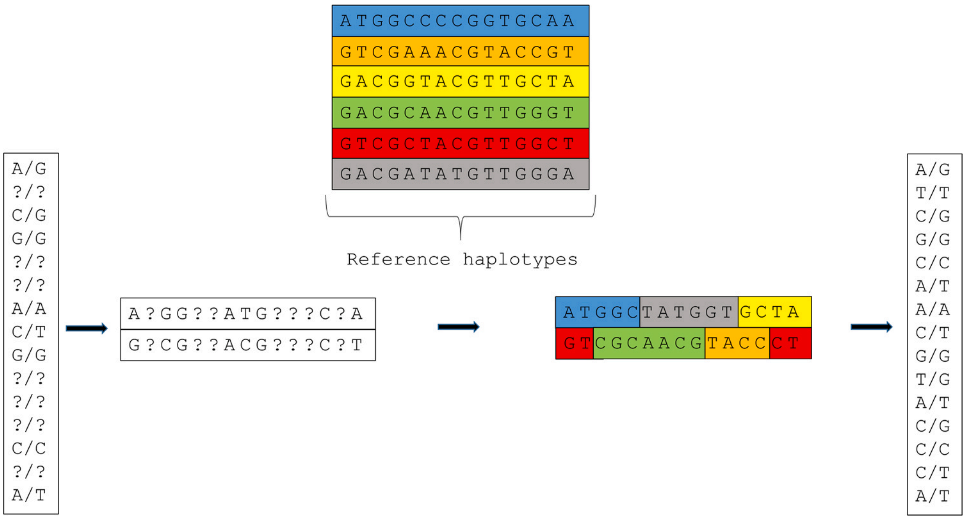


Fig. 8. An illustration of how genotype imputation works. The original SNP data is shown on the left. This set consists of SNP genotypes observed directly from the DNA analysis but also missing genotypes (marked with "?/?"). First, the observed diploid data is phased into haplotypes, which are then compared with haplotypes from a reference panel (e.g. phased 1000 Genomes data). Second, missing SNP alleles can be predicted using the matching haplotypes in the reference data and a probabilistic model. The final genotype set, which includes both original and imputed genotypes, is shown on the right.

SNPs included in the current version of the microarray used by AncestryDNA were used to estimate the error rate when reducing the number of observed SNPs for imputation and thus increasing the masked SNPs which required imputation. Although no direct conclusions can be drawn from this single experiment, the imputation error rate was around 1% or less, even when the number of SNPs was reduced to less than 100, 000 (Supplementary File S2, Fig. 5).

It has been noted that several of the companies use genotype imputation, at least to some degree, either to accept transfers from different companies or to ensure backwards compatibility with the OmniExpress microarray, though no published details are available. The MyHeritage imputation process has been described in a company blog post.[71] However, the extent of imputation amongst DTC and forensic service providers, and its specific application in genealogical analyses beyond MyHeritage are not known.

9. Concluding remarks: proportionality and the CODIS gap

While IGG is an exciting and powerful forensic genetic technique which has led to the successful resolution of many long-standing cold cases, its use has highlighted systemic problems in the US criminal justice system. It has become apparent that many cold cases could have been solved much earlier using existing DNA technologies. It is estimated there are many thousands of profiles from convicted offenders which are legally mandated but have yet to be collected.[72] In Washington State alone, 30,000 convicted offenders are thought to be missing from the CODIS database, highlighting an increasing 'CODIS gap' in the US.[73] In a growing number of cases, IGG has identified suspects who had a criminal history but were not in CODIS.[74] This includes two criminals in Texas who were executed without having their DNA taken.[75]

The problem of untested sexual assault kits has been well documented and publicized, but although many grants have been awarded over the years to clear the backlog, it continues to increase.[76] Testing on its own will not solve the problem if funding is not also provided to pay for the costs of investigating and prosecuting cases.[77]

The CODIS gap is further exacerbated by the piecemeal use of familial searching in the US. It is currently confined to 12 of the 50 US states (Arizona, California, Colorado, Florida, Michigan, New York, Ohio, Texas, Utah, Virginia, Wisconsin, and Wyoming), is explicitly prohibited in Maryland and Washington DC, and is not permitted in the federal CODIS database.[78] Familial searches are restricted to people who are already in the CODIS database and, because they have been convicted of a crime, are considered to have forfeited some rights to privacy. Therefore, familial searching has far fewer privacy implications than IGG, which extends searches to both close and distant relatives who are in a genealogy DNA database, not all of whom have given specific consent for their profiles to be used. IGG can also involve networks of related family members who have not had their DNA tested but are approached for target testing.[79] It is therefore sobering to find that some cases where IGG was used could have been solved much earlier if familial searching had been implemented. For example, Patrick Leon Nicholas was identified as a suspect in the murder of Sarah Yarborough in Washington State, yet could have been caught through familial searching because his brother's DNA was entered into CODIS in 2005.[80] Joseph DeAngelo, the Golden State Killer, had a brother who was convicted of a felony. He could have been caught many years earlier if familial searching had been in use at the time and if the law had been in place to allow police to take DNA from arrestees [2].

Proportionality is a key concept in forensics when attempting to balance the privacy rights of the individual versus the need for public safety [147]. It is vital to ensure that the least privacy-invasive methods are used first and IGG is used as a last resort and not to compensate for

[71] See: https://blog.myheritage.com/2018/01/major-updates-and-improvements-to-myheritage-dna-matching/.

[72] See: https://www.corrections1.com/products/police-technology/investigation/biometrics-identification/articles/hidden-in-prison-thousands-of-inmates-not-in-dna-databases-8r9qlgaNHXfvA4aJ/

[73] See: https://www.atg.wa.gov/news/news-releases/ag-ferguson-wins-additional-25-million-fund-sexual-assault-kit-initiative-program

[74] See: https://www.nbcnews.com/news/us-news/national-disgrace-holes-dna-databases-leave-crimes-unsolved-decades-n1236748

[75] See: https://web.archive.org/web/20190202013800/https://www.forensicmag.com/news/2019/02/police-discover-oregon-cold-case-killer-was-executed-texas-1999 and: https://www.kwtx.com/content/news/DNA-confirms-ID-of-man-who-killed-area-real-estate-agent-in-1981–562531361.html

[76] See: https://promega.foleon.com/theishireport/november-2019/pursuing-justice-the-state-of-the-sexual-assault-kit-backlog-in-the-united-states/ and https://www.channel4.com/news/us-rape-survivors-wait-for-justice-amid-backlog-of-dna-testing-kits

[77] See: https://web.archive.org/web/20160303071244/http://www.forensicmag.com/articles/2015/12/going-beyond-codis-illusion-rape-kit-testing-panacea

[78] See: https://www.nbcnews.com/news/us-news/familial-dna-puts-elusive-killers-behind-bars-only-12-states-n869711

[79] See: https://www.nbcnews.com/news/us-news/they-lied-us-mom-says-police-deceived-her-get-her-n1140696

[80] See https://www.kiro7.com/news/local/wednesday-at-5-30-why-isnt-washington-using-dna-tool-to-solve-crimes-/1010311730/s-/1010311730/

systemic failures.

Acknowledgements

We are indebted to Brett Williams (CEO), Cydne Holt and Nicola Oldroyd-Clark of Verogen for sharing information about the GEDmatch database and Verogen's future plans. We thank Ellen Greytak of Parabon NanoLabs for the detail provided in responses to the questionnaire and for permission to use the data and graphs in Supplementary File S3. We also thank the other service provider representatives who gave helpful information in their answers to the questionnaire. We thank Rock Harmon for informative discussions. Finally we thank the two reviewers for their helpful and constructive comments.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102474.

References

[1] C.N. Maguire, L.A. McCallum, C. Storey, J.P. Whitaker, Familial searching: a specialist forensic DNA profiling service utilising the National DNA Database® to identify unknown offenders via their relatives—the UK experience, Forensic Sci. Int.: Genet. 8 (1) (2014) 1–9.
[2] 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.
[3] M. Stallard, J. de Groot, "Things are coming out that are questionable, we never knew about": DNA and the new family history, J. Fam. Hist. 45 (3) (2020) 274–294.
[4] C. Moore, The history of genetic genealogy and unknown parentage research: an insider's view, J. Genet. Geneal. 8 (1) (2016) 35–37.
[5] J.C. Harper, D. Kennett, D. Reisel, The end of donor anonymity: how genetic testing is likely to drive anonymous gamete donation out of business, Hum. Reprod. 31 (6) (2016) 1135–1140.
[6] L. Copeland, *The Lost Family: How DNA Testing Is Upending Who We Are*, Blackstone Publishing, Abrams Press,, 2020.
[7] D. Kling, J. Welander, A. Tillmar, Ø. Skare, T. Egeland, G. Holmlund, DNA microarray as a tool in establishing genetic relatedness – current status and future prospects, Forensic Sci. Int.: Genet. 6 (3) (2012) 322–329.
[8] M. Sun, M.A. Jobling, D. Taliun, P.P. Pramstaller, T. Egeland, N.A. Sheehan, On the use of dense SNP marker data for the identification of distant relative pairs, Theor. Popul. Biol. 107 (2016) 14–25.
[9] C. Morimoto, S. Manabe, T. Kawaguchi, C. Kawai, S. Fujimoto, Y. Hamano, R. Yamada, F. Matsuda, K. Tamaki, Pairwise kinship analysis by the index of chromosome sharing using high-density single nucleotide polymorphisms, PLoS One 11 (7) (2016), e0160287.
[10] U.A. Perego, M. Bodner, A. Raveane, S.R. Woodward, F. Montinaro, W. Parson, A. Achilli, Resolving a 150-year-old paternity case in Mormon history using DTC autosomal DNA testing of distant relatives, Forensic Sci. Int. Genet. 42 (2019) 1–7.
[11] M.V. Lareu, M. García-Magariños, C. Phillips, I. Quintela, Á. Carracedo, A. Salas, Analysis of a claimed distant relationship in a deficient pedigree using high density SNP data, Forensic Sci. Int. Genet. 6 (3) (2012) 350–353.
[12] G. Pálsson, Decode me! Anthropology and personal genomics, Curr. Anthropol. 53 (S5) (2012) S185–S195.
[13] A. Regalado. More than 26 Million People have taken an at-Home Ancestry Test, *MIT Technology Review*, 2019.
[14] C. Phillips, The Golden State Killer investigation and the nascent field of forensic genealogy, Forensic Sci. Int. Genet. 36 (2018) 186–188.
[15] E.M. Greytak, C. Moore, S.L. Armentrout, Genetic genealogy for cold case and active investigations, Forensic Sci. Int. 299 (2019) 103–113.
[16] S.H. Katsanis, Pedigrees and perpetrators: uses of DNA and genealogy in forensic investigations, Annu. Rev. Genom. Hum. Genet. 21 (2020) 535–564.
[17] A. Tillmar, P. Sjölund, B. Lundqvist, T. Klippmark, C. Älgenäs, H. Green, Whole-genome sequencing of human remains to enable genealogy DNA database searches – a case report, Forensic Sci. Int. Genet. 46 (2020), 102233.
[18] G. Samuel, D. Kennett, The impact of investigative genetic genealogy: perceptions of UK professional and public stakeholders, Forensic Sci. Int. Genet. 48 (2020), 102366.
[19] J. Dabney, M. Meyer, S. Pääbo, Ancient DNA damage, Cold Spring Harb. Perspect. Biol. 5 (7) (2013), a012567.
[20] 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 (4) (2012), e34267.
[21] Ball, C.A., et al. AncestryDNA matching white paper. 2020; Available from: 〈https://www.ancestrycdn.com/support/us/2020/08/matchingwhitepaper.pdf〉.
[22] Petter, E., et al., Relative matching using low coverage sequencing. bioRxiv, 2020: p. 2020.09.09.289322.
[23] D. Kling, On the use of dense sets of SNP markers and their potential in relationship inference, Forensic Sci. Int. Genet. 39 (2019) 19–31.
[24] D. Kling, A. Tillmar, Forensic genealogy—a comparison of methods to infer distant relationships based on dense SNP data, Forensic Sci. Int. Genet. 42 (2019) 113–124.
[25] A. Kong, G. Masson, M.L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D.F. Gudbjartsson, H. Stefansson, K. Stefansson, Detection of sharing by descent, long-range phasing and haplotype imputation, Nat. Genet. 40 (9) (2008) 1068–1075.
[26] H. Li, G. Glusman, C. Huff, J. Caballero, J.C. Roach, Accurate and robust prediction of genetic relationship from whole-genome sequences, PLoS One 9 (2) (2014), e85437.
[27] 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 (22) (2010) 2867–2873.
[28] M.P. Epstein, W.L. Duren, M. Boehnke, Improved inference of relationship for pairs of individuals, Am. J. Hum. Genet. 67 (5) (2000) 1219–1231.
[29] B.S. Weir, A.D. Anderson, A.B. Hepler, Genetic relatedness analysis: modern data and new challenges, Nat. Rev. Genet. 7 (10) (2006) 771–780.
[30] Ø. Skare, N. Sheehan, T. Egeland, Identification of distant family relationships, Bioinformatics 25 (18) (2009) 2376–2382.
[31] S.R. Browning, B.L. Browning, High-resolution detection of identity by descent in unrelated individuals, Am. J. Hum. Genet. 86 (4) (2010) 526–539.
[32] Thompson, E.A., Statistical inference from genetic data on pedigrees. NSF-CBMS regional conference series in probability and statistics. 2000: JSTOR. i-169.
[33] E.A. Thompson, Identity by descent: variation in meiosis, across genomes, and in populations, Genetics 194 (2) (2013) 301–326.
[34] D. Speed, D.J. Balding, Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. 16 (1) (2015) 33–44.
[35] M.S. McPeek, L. Sun, Statistical tests for detection of misspecified relationships by use of genome-screen data, Am. J. Hum. Genet. 66 (3) (2000) 1076–1094.
[36] S.R. Browning, B.L. Browning, Identity by descent between distant relatives: detection and applications, Annu. Rev. Genet. 46 (2012) 617–633.
[37] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (3) (2007) 559–575.
[38] B.L. Browning, S.R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data, Genetics 194 (2) (2013) 459–471.
[39] B.L. Browning, S.R. Browning, Detecting identity by descent and estimating genotype error rates in sequence data, Am. J. Hum. Genet. 93 (5) (2013) 840–851.
[40] A. Gusev, J.K. Lowe, M. Stoffel, M.J. Daly, D. Altshuler, J.L. Breslow, J. M. Friedman, I. Pe'er, Whole population, genome-wide mapping of hidden relatedness, Genome Res. 19 (2) (2009) 318–326.
[41] A. Albrechtsen, T. Sand Korneliussen, I. Moltke, T. van Overseem Hansen, F. C. Nielsen, R. Nielsen, Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium, Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc. 33 (3) (2009) 266–274.
[42] M.D. Brown, C.G. Glazner, C. Zheng, E.A. Thompson, Inferring coancestry in population samples in the presence of linkage disequilibrium, Genetics 190 (4) (2012) 1447–1460.
[43] L. Han, M. Abney, Identity by descent estimation with dense genome-wide genotype data, Genet. Epidemiol. 35 (6) (2011) 557–567.
[44] B.L. Browning, S.R. Browning, A fast, powerful method for detecting identity by descent, Am. J. Hum. Genet. 88 (2) (2011) 173–182.
[45] 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 (5) (2011) 768–774.
[46] S.R. Browning, B.L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, Am. J. Hum. Genet. 81 (5) (2007) 1084–1097.
[47] S. Hochreiter, HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data, Nucleic Acids Res. 41 (22) (2013), e202 e202-e202.
[48] A. Dimitromanolakis, A.D. Paterson, L. Sun, Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE, Am. J. Hum. Genet. 105 (1) (2019) 78–88.
[49] G.R. Abecasis, S.S. Cherny, W.O. Cookson, L.R. Cardon, Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (1) (2002) 97–101.
[50] D.F. Gudbjartsson, K. Jonasson, M.L. Frigge, A. Kong, Allegro, a new computer program for multipoint linkage analysis, Nat. Genet. 25 (1) (2000) 12–13.
[51] A.L. Leutenegger, E. Génin, E.A. Thompson, F. Clerget-Darpoux, Impact of parental relationships in maximum lod score affected sib-pair method, Genet. Epidemiol. 23 (4) (2002) 413–425.
[52] D.W. Gjertson, et al., ISFG: recommendations on biostatistics in paternity testing, Forensic Sci. Int. Genet. 1 (3–4) (2007) 223–231.
[53] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (6) (1971) 523–542.
[54] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. U.S.A. 84 (8) (1987) 2363–2367.
[55] E. Thompson, The estimation of pairwise relationships, Ann. Hum. Genet. 39 (2) (1975) 173–188.

[56] A.D. Anderson, B.S. Weir, A maximum-likelihood method for the estimation of pairwise relatedness in structured populations, Genetics 176 (1) (2007) 421–440.
[57] A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int.: Genet. 26 (2017) 58–65.
[58] C. Phillips, J. Amigo, A.O. Tillmar, M.A. Peck, M. de la Puente, J. Ruiz-Ramírez, F. Bittner, Š. Idrizbegović, Y. Wang, T.J. Parsons, M.V. Lareu, A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel, Forensic Sci. Int.: Genet. 46 (2020), 102232.
[59] E.Y. Durand, N. Eriksson, C.Y. McLean, Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis, Mol. Biol. Evol. 31 (8) (2014) 2212–2222.
[60] D.W. Bjelland, U. Lingala, P.S. Patel, M. Jones, M.C. Keller, A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data, Eur. J. Hum. Genet. 25 (5) (2017) 617–624.
[61] C. Champod, et al., ENFSI guideline for evaluative reporting in forensic science, a primer for legal practitioners, Crim. Law Justice Wkly. 180 (10) (2016) 189–193.
[62] J. Ge, B. Budowle, How many familial relationship testing results could be wrong? PLoS Genet. 16 (8) (2020), e1008929.
[63] D.E. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, E.S. Lander, Linkage disequilibrium in the human genome, Nature 411 (6834) (2001) 199–204.
[64] J.K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: models and data, Am. J. Hum. Genet. 69 (1) (2001) 1–14.
[65] G.R. Abecasis, E. Noguchi, A. Heinzmann, J.A. Traherne, S. Bhattacharyya, N. I. Leaves, G.G. Anderson, Y. Zhang, N.J. Lench, A. Carey, L.R. Cardon, M. F. Moffatt, W.O.C. Cookson, Extent and distribution of linkage disequilibrium in three genomic regions, Am. J. Hum. Genet. 68 (1) (2001) 191–197.
[66] C.W. Chiang, P. Ralph, J. Novembre, Conflation of short identity-by-descent segments bias their inferred length distribution, G3: Genes, Genomes, Genet. 6 (5) (2016) 1287–1296.
[67] K.P. Donnelly, The probability that related individuals share some section of genome identical by descent, Theor. Popul. Biol. 23 (1) (1983) 34–63.
[68] M.D. Edge, G. Coop, Donnelly (1983) and the limits of genetic genealogy, Theor. Popul. Biol. 133 (2020) 23–24.
[69] B. Bettinger, The shared cM project – Version 4.0. 2020; Available from: 〈https://thegeneticgenealogist.com/2020/03/27/version-4-0-march-2020-update-to-the-shared-cm-project/〉.
[70] B.T. Bettinger, The shared cM project: a demonstration of the power of citizen science, J. Genet. Geneal. 8 (1) (2016) 38–42.
[71] H. Li, G. Glusman, H. Hu, Shankaracharya, J. Caballero, R. Hubley, D. Witherspoon, S.L. Guthery, D.E. Mauldin, L.B. Jorde, L. Hood, J.C. Roach, C. D. Huff, Relationship estimation from whole-genome sequence data, PLoS Genet. 10 (1) (2014), e1004144.
[72] A. Al-Khudhair, S. Qiu, M. Wyse, S. Chowdhury, X. Cheng, D. Bekbolsynov, A. Saha-Mandal, R. Dutta, L. Fedorova, A. Fedorov, Inference of distant genetic relations in humans using “1000 genomes”, Genome Biol. Evol. 7 (2) (2015) 481–492.
[73] P. Ralph, G. Coop, The geography of recent genetic ancestry across Europe, PLoS Biol. 11 (5) (2013), e1001555.
[74] H. Gauvin, C. Moreau, J.F. Lefebvre, C. Laprise, H. Vézina, D. Labuda, M.H. Roy-Gagnon, Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population, Eur. J. Hum. Genet. 22 (6) (2014) 814–821.
[75] S. Carmi, K.Y. Hui, E. Kochav, X. Liu, J. Xue, F. Grady, S. Guha, K. Upadhyay, D. Ben-Avraham, S. Mukherjee, B.M. Bowen, T. Thomas, J. Vijai, M. Cruts, G. Froyen, D. Lambrechts, S. Plaisance, C. Van Broeckhoven, P. Van Damme, H. Van Marck, N. Barzilai, A. Darvasi, K. Offit, S. Bressman, L.J. Ozelius, I. Peter, J.H. Cho, H. Ostrer, G. Atzmon, L.N. Clark, T. Lencz, I. Pe’er, Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins, Nat. Commun. 5 (1) (2014) 1–9.
[76] E. Gilbert, S. Carmi, S. Ennis, J.F. Wilson, G.L. Cavalleri, Genomic insights into the population structure and history of the Irish Travellers, Sci. Rep. 7 (2017) 42187.
[77] V. Buffalo, S.M. Mount, G. Coop, A genealogical look at shared ancestry on the X chromosome, Genetics 204 (1) (2016) 57–75.
[78] *Advanced Genetic Genealogy*, in: D.P. Wayne (Ed.), Wayne Research, Texas, 2019.
[79] A. Gusev, P.F. Palamara, G. Aponte, Z. Zhuang, A. Darvasi, P. Gregersen, I. Pe’er, The architecture of long-range haplotypes shared within and across populations, Mol. Biol. Evol. 29 (2) (2012) 473–486.
[80] I.W. Saunders, J. Brohede, G.N. Hannan, Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference, Genomics 90 (3) (2007) 291–296.
[81] Genographic Project Consortium, 2015. A global reference for human genetic variation, Nature, 526 (7571), 68–74.
[82] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44.
[83] J.-A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, J. Buckleton, Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles, Forensic Sci. Int. Genet. 23 (2016) 226–239.
[84] H. Haned, Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (4) (2011) 265–268.
[85] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B. W. Duceman, Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[86] E. Alladio, M. Omedei, S. Cisana, G. D’Amico, D. Caneparo, M. Vincenti, P. Garofano, DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods’ performances on challenging samples, Forensic Sci. Int. Genet. 37 (2018) 143–150.
[87] K. Slooten, Familial searching on DNA mixtures with dropout, Forensic Sci. Int. Genet. 22 (2016) 128–138.
[88] Y.K. Chung, Y.Q. Hu, W.K. Fung, Evaluation of DNA mixtures from database search, Biometrics 66 (1) (2010) 233–238.
[89] Y.-K. Chung, W.K. Fung, Y.-Q. Hu, Familial database search on two-person mixture, Comput. Stat. Data Anal. 54 (8) (2010) 2046–2051.
[90] L. Bradford, J. Heal, J. Anderson, N. Faragher, K. Duval, S. Lalonde, Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0, Forensic Sci. Int. Genet. 5 (4) (2011) 291–296.
[91] D. Kling, S. Füredi, The successful use of familial searching in six Hungarian high profile cases by applying a new module in Familias 3, Forensic Sci. Int. Genet. 24 (2016) 24–32.
[92] G. Dørum, D. Kling, A. Tillmar, M.D. Vigeland, T. Egeland, Mixtures with relatives and linked markers, Int. J. Leg. Med. 130 (3) (2016) 621–634.
[93] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D.A. Stephan, S.F. Nelson, D.W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS Genet. 4 (8) (2008), e1000167.
[94] J. Thomson, T. Clayton, J. Cleary, M. Gleeson, D. Kennett, M. Leonard, D. Rutherford, An empirical investigation into the effectiveness of genetic genealogy to identify individuals in the UK, Forensic Sci. Int. Genet. 46 (2020), 102263.
[95] P. Gorry, *Credentials for Genealogists: Proof of the Professional*, Gorry Research,, Wicklow, Ireland, 2018.
[96] Edge, M. and G. Coop, How lucky was the genetic investigation in the golden state killer case? 2019. bioRxiv 531384; doi: 10.1101/531384.
[97] Y. Erlich, T. Shor, I. Pe’er, S. Carmi, Identity inference of genomic data using long-range familial searches, Science 362 (6415) (2018) 690–694.
[98] S. Skeva, M.H. Larmuseau, M. Shabani, Review of policies of companies and databases regarding access to customers’ genealogy data for law enforcement purposes, Pers. Med. 17 (2) (2020) 141–153.
[99] C. Arnold, The controversial company using DNA to sketch the faces of criminals, Nature 585 (7824) (2020) 178–181.
[100] D.S. Court, Forensic genealogy: Some serious concerns, Forensic Sci. Int. Genet. 36 (2018) 203–204.
[101] B.E. Berkman, W.K. Miller, C. Grady, Is it ethical to use genealogy data to solve crimes? American College of Physicians,, 2018.
[102] L. Edwards, E. Harbina, Protecting post-mortem privacy: reconsidering the privacy interests of the deceased in a digital world, SSRN Electron. J. 32 (2013) 83.
[103] G. Samuel, D. Kennett, Problematizing consent: searching genetic genealogy databases for law enforcement purposes, N. Genet. Soc. (2020) 1–21.
[104] T.F. Callaghan, Responsible genetic genealogy, American Association for the Advancement of Science,, 2019.
[105] N. Scudder, R. Daniel, J. Raymond, A. Sears, Operationalising forensic genetic genealogy in an Australian context, Forensic Sci. Int. 316 (2020), 110543, https://doi.org/10.1016/j.forsciint.2020.110543.
[106] N. Eriksson, J.M. Macpherson, J.Y. Tung, L.S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, I. Pe’er, J. Mountain, Web-based, participant-driven studies yield novel genetic associations for common traits, PLoS Genet. 6 (6) (2010), e1000993.
[107] P.G. Hysi, A.M. Valdes, F. Liu, N.A. Furlotte, D.M. Evans, V. Bataille, A. Visconti, G. Hemani, G. McMahon, S.M. Ring, G.D. Smith, D.L. Duffy, G. Zhu, S.D. Gordon, S.E. Medland, B.D. Lin, G. Willemsen, J. Jan Hottenga, D. Vuckovic, G. Girotto, I. Gandin, C. Sala, M.P. Concas, M. Brumat, P. Gasparini, D. Toniolo, M. Cocca, A. Robino, S. Yazar, A.W. Hewitt, Y. Chen, C. Zeng, A.G. Uitterlinden, M.A. Ikram, M.A. Hamer, C.M. van Duijn, T. Nijsten, D.A. Mackey, M. Falchi, D.I. Boomsma, N.G. Martin, D.A. Hinds, M. Kayser, T.D. Spector, Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability, Nat. Genet. 50 (5) (2018) 652–656.
[108] J.K. Wagner, J.D. Cooper, R. Sterling, C.D. Royal, Tilting at windmills no longer: a data-driven discussion of DTC DNA ancestry tests, Genet. Med. 14 (6) (2012) 586–593.
[109] M.D. Edge, G. Coop, Attacks on genetic privacy via uploads to genealogical databases, Elife 9 (2020), e51810.
[110] Ney, P., L. Ceze, and T. Kohno, Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference. in Network and Distributed System Security Symposium (NDSS). 2020.
[111] C.P. Schaaf, J. Wiszniewska, A.L. Beaudet, Copy number and SNP arrays in clinical diagnostics, Annu. Rev. Genom. Hum. Genet. 12 (2011) 25–51.
[112] T. LaFramboise, Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, Nucleic Acids Res. 37 (13) (2009) 4181–4193.
[113] R.M. Rajagopalan, J.H. Fujimura, Variations on a chip: technologies of difference in human genetics research, J. Hist. Biol. 51 (4) (2018) 841–873.
[114] F.R. Wendt, A.L. Rahikainen, J.L. King, A. Sajantila, B. Budowle, A genome-wide association study of tramadol metabolism from post-mortem samples, Pharm. J. 20 (1) (2020) 94–103.
[115] M.D. Brandhagen, O. Loreille, J.A. Irwin, Fragmented nuclear DNA is the predominant genetic material in human hair shafts, Genes 9 (12) (2018) 640.

[116] G. Holmlund, et al., Experiences from DNA Analysis in Sweden for the Identification of Tsunami Victims. International Congress Series, Elsevier,, 2006.

[117] I. Grandell, R. Samara, A.O. Tillmar, A SNP panel for identity and kinship testing using massive parallel sequencing, Int. J. Leg. Med. 130 (4) (2016) 905–914.

[118] E. Samorodnitsky, B.M. Jewell, R. Hagopian, J. Miya, M.R. Wing, E. Lyon, S. Damodaran, D. Bhatt, J.W. Reeser, J. Datta, S. Roychowdhury, Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing, Hum. Mutat. 36 (9) (2015) 903–914.

[119] E. Samorodnitsky, J. Datta, B.M. Jewell, R. Hagopian, J. Miya, M.R. Wing, S. Damodaran, J.M. Lippus, J.W. Reeser, D. Bhatt, C.D. Timmers, S. Roychowdhury, Comparison of custom capture for targeted next-generation DNA sequencing, J. Mol. Diagn. 17 (1) (2015) 64–75.

[120] M.C. Ávila-Arcos, E. Cappellini, J.A. Romero-Navarro, N. Wales, J.V. Moreno-Mayar, M. Rasmussen, S.L. Fordyce, R. Montiel, J.P. Vielle-Calzada, E. Willerslev, M.T.P. Gilbert, Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA, Sci. Rep. 1 (1) (2011) 1–5.

[121] M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing, Cold Spring Harb. Protoc. 2010 (6) (2010) pdb.prot5448-pdb.prot5448. prot5448.

[122] C. Marshall, K. Sturk-Andreaggi, J. Daniels-Higginbotham, R.S. Oliver, S. Barritt-Ross, T.P. McMahon, Performance evaluation of a mitogenome capture and Illumina sequencing protocol using non-probative, case-type skeletal samples: Implications for the use of a positive control in a next-generation sequencing procedure, Forensic Sci. Int. Genet. 31 (2017) 198–206.

[123] M.L. Carpenter, J.D. Buenrostro, C. Valdiosera, H. Schroeder, M.E. Allentoft, M. Sikora, M. Rasmussen, S. Gravel, S. Guillén, G. Nekhrizov, K. Leshtakov, D. Dimitrova, N. Theodossiev, D. Pettener, D. Luiselli, K. Sandoval, A. Moreno-Estrada, Y. Li, J. Wang, M.T.P. Gilbert, E. Willerslev, W.J. Greenleaf, C. D. Bustamante, Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries, Am. J. Hum. Genet. 93 (5) (2013) 852–864.

[124] A.W. Briggs, J.M. Good, R.E. Green, J. Krause, T. Maricic, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, R. Schmitz, V.B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, S. Paabo, Targeted retrieval and analysis of five Neandertal mtDNA genomes, Science 325 (5938) (2009) 318–321.

[125] C.W. Knapp, J. Dolfing, P.A.I. Ehlert, D.W. Graham, Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940, Environ. Sci. Technol. 44 (2) (2010) 580–587.

[126] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S.A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, K. Sirak, C. Gamba, E. R. Jones, B. Llamas, S. Dryomov, J. Pickrell, J.L. Arsuaga, J.M.B. de Castro, E. Carbonell, F. Gerritsen, A. Khokhlov, P. Kuznetsov, M. Lozano, H. Meller, O. Mochalov, V. Moiseyev, M.A.R. Guerra, J. Roodenberg, J.M. Vergès, J. Krause, A. Cooper, K.W. Alt, D. Brown, D. Anthony, C. Lalueza-Fox, W. Haak, R. Pinhasi, D. Reich, Genome-wide patterns of selection in 230 ancient Eurasians, Nature 528 (7583) (2015) 499–503.

[127] M. Feldman, D.M. Master, R.A. Bianco, M. Burri, P.W. Stockhammer, A. Mittnik, A.J. Aja, C. Jeong, J. Krause, Ancient DNA sheds light on the genetic origins of early Iron Age Philistines, Sci. Adv. 5 (7) (2019) eaax0061.

[128] S. Shih, N. Bose, A. Gonçalves, H. Erlich, C. Calloway, Applications of probe capture enrichment next generation sequencing for whole mitochondrial genome and 426 nuclear SNPs for forensically challenging samples, Genes 9 (1) (2018) 49.

[129] R. England, S. Harbison, A review of the method and validation of the MiSeq FGx™ Forensic Genomics Solution, WIREs Forensic Sci. 2 (1) (2020), e1351.

[130] L.J. Scott, K.L. Mohlke, L.L. Bonnycastle, C.J. Willer, Y. Li, W.L. Duren, M. R. Erdos, H.M. Stringham, P.S. Chines, A.U. Jackson, L. Prokunina-Olsson, C. J. Ding, A.J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X.Y. Li, K.N. Conneely, N. L. Riebow, A.G. Sprau, M. Tong, P.P. White, K.N. Hetrick, M.W. Barnhart, C. W. Bark, J.L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T.A. Buchanan, R. M. Watanabe, T.T. Valle, L. Kinnunen, G.R. Abecasis, E.W. Pugh, K.F. Doheny, R. N. Bergman, J. Tuomilehto, F.S. Collins, M. Boehnke, A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants, Science 316 (5829) (2007) 1341–1345.

[131] S. Das, G.R. Abecasis, B.L. Browning, Genotype imputation from large reference panels, Annu. Rev. Genom. Hum. Genet. 19 (2018) 73–96.

[132] C.C.A. Spencer, Z. Su, P. Donnelly, J. Marchini, Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip, PLoS Genet. 5 (5) (2009), e1000477.

[133] C. Wang, X. Zheng, R. Tang, C. Han, Y. Jiang, J. Wu, Y. Shao, Y. Gao, J. Yu, Z. Hu, Z. Zang, Y. Zhao, N. Dai, L. Liu, X. Wu, J. Nie, B. Jiang, M. Lin, L. Li, Y. Wei, Y. Li, Y. Gong, Y. Dai, L. Wang, N. Ding, P. Xu, S. Chen, P. Jiang, L. Wang, F. Qiu, Q. Wu, M. Zhang, R. Jawed, R. Chen, Y. Zhang, X. Shi, Z. Zhu, H. Pei, L. Huang, Y. Tian, K. Zhang, H. Qiu, W. Zhao, M.E. Gershwin, W. Chen, M.F. Seldin, X. Liu, X. Ma, L. Sun, Fine mapping of the MHC region identifies major independent variants associated with Han Chinese primary biliary cholangitis, J. Autoimmun. 107 (2020), 102372.

[134] P.I.W. De Bakker, M.A.R. Ferreira, X. Jia, B.M. Neale, S. Raychaudhuri, B. F. Voight, Practical aspects of imputation-driven meta-analysis of genome-wide association studies, Hum. Mol. Genet. 17 (R2) (2008) R122–R128.

[135] M.D. Edge, et al., Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets, Proc. Natl. Acad. Sci. 114 (22) (2017) 5671–5676.

[136] J. Kim, M.D. Edge, B.F.B. Algee-Hewitt, J.Z. Li, N.A. Rosenberg, Statistical detection of relatives typed with disjoint forensic and biomedical loci, Cell 175 (3) (2018) 848–858, e6.

[137] S.R. Browning, B.L. Browning, Haplotype phasing: existing methods and new developments, Nat. Rev. Genet. 12 (10) (2011) 703–714.

[138] J. Marchini, B. Howie, Genotype imputation for genome-wide association studies, Nat. Rev. Genet. 11 (7) (2010) 499–511.

[139] B.L. Browning, Y. Zhou, S.R. Browning, A one-penny imputed genome from next-generation reference panels, Am. J. Hum. Genet. 103 (3) (2018) 338–348.

[140] G. Gibson, Rare and common variants: twenty arguments, Nat. Rev. Genet. 13 (2) (2012) 135–145.

[141] S. Shi, N. Yuan, M. Yang, Z. Du, J. Wang, X. Sheng, J. Wu, J. Xiao, Comprehensive assessment of genotype imputation performance, Hum. Hered. 83 (3) (2018) 107–116.

[142] G. Pistis, E. Porcu, S.I. Vrieze, C. Sidore, M. Steri, F. Danjou, F. Busonero, A. Mulas, M. Zoledziewska, A. Maschio, C. Brennan, S. Lai, M.B. Miller, M. Marcelli, M.F. Urru, M. Pitzalis, R.H. Lyons, H.M. Kang, C.M. Jones, A. Angius, W.G. Iacono, D. Schlessinger, M. McGue, F. Cucca, G.R. Abecasis, S. Sanna, Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs, Eur. J. Hum. Genet. 23 (7) (2015) 975–983.

[143] K.A. Frazer, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (7164) (2007) 851–861.

[144] J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J.L. Min, P. Danecek, G. Malerba, E. Trabetti, H.F. Zheng, G. Gambaro, J.B. Richards, R. Durbin, N. J. Timpson, J. Marchini, N. Soranzo, Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel, Nat. Commun. 6 (1) (2015) 1–9.

[145] U.K. consortium, The UK10K project identifies rare variants in health and disease, Nature 526 (7571) (2015) 82–90.

[146] J.A. Brody, A.C. Morrison, J.C. Bis, J.R. O'Connell, M.R. Brown, J.E. Huffman, D. C. Ames, A. Carroll, M.P. Conomos, S. Gabriel, R.A. Gibbs, S.M. Gogarten, N. Gupta, C.E. Jaquish, A.D. Johnson, J.P. Lewis, X. Liu, A.K. Manning, G. J. Papanicolaou, A.N. Pitsillides, K.M. Rice, W. Salerno, C.M. Sitlani, N.L. Smith, S.R. Heckbert, C.C. Laurie, B.D. Mitchell, R.S. Vasan, S.S. Rich, J.I. Rotter, J. G. Wilson, E. Boerwinkle, B.M. Psaty, L.A. Cupples, Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology, Nat. Genet. 49 (11) (2017) 1560–1563.

[147] R.A. Wickenheiser, Forensic genealogy, bioethics and the Golden State Killer case, Forensic Sci. Int. Synergy 1 (2019) 114–125.