Forensic Science International: Genetics 53 (2021) 102525

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Getting the conclusive lead with investigative genetic genealogy – A successful case study of a 16 year old double murder in Sweden

Andreas Tillmar [a,b,*], Siri Aili Fagerholm [c], Jan Staaf [d], Peter Sjölund [e], Ricky Ansell [c,f,**]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[b] Department of Biomedical and Clinical Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden
[c] National Forensic Centre, Swedish Police Authority, Linköping, Sweden
[d] Polisregion Öst, Swedish Police Authority, Linköping, Sweden
[e] Peter Sjölund AB, Härnösand, Sweden
[f] Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Investigative genetic genealogy (IGG)
Forensic genetic genealogy (FGG)
Whole genome sequencing
Forensic DNA
Genotype imputation

ABSTRACT

On the morning of October 19, 2004, an eight-year-old boy and a 56-year-old woman were stabbed to death on an open street in the city of Linköping, Sweden. The perpetrator left his DNA at the crime scene, and after 15 years of various investigation efforts, including more than 9000 interrogations and mass DNA screening of more than 6000 men, there were still no clues about the identity of the unknown murderer. The successful application of investigative genetic genealogy (IGG) in the US raised the interest for this tool within the Swedish Police Authority. After legal consultations it was decided that IGG could be applied in this double murder case as a pilot case study. From extensive DNA analysis, including whole-genome sequencing and genotype imputation, DNA data sets were established and searched within both GEDmatch and FamilyTree DNA genealogy databases. A number of fairly distant relatives were found from which family trees were created. The genealogy work resulted in two candidates, two brothers, one of whom matched the crime scene samples by routine STR profiling. The suspect confessed the murders at the initial police hearing and was later convicted of the murders. In this paper we describe the successful application of an emerging technology. We disclose details of the DNA analyses which, due to the poor quality and low quantity of the DNA, required reiterative sequencing and genotype imputation efforts. The successful application of IGG in this double murder case exemplifies its applicability not only in the US but also in Europe. The pressure is now high on the involved authorities to establish IGG as a tool for cold case criminal investigations and for missing person identifications. There is, however, a continuous need to accommodate legal, social and ethical aspects as well.

1. Introduction

Investigative genetic genealogy (IGG) or forensic genetic genealogy (FGG) has emerged as a powerful forensic tool to generate crucial leads to identify unknown perpetrators and to identify unknown human remains [1–7]. IGG includes the use of large genotype data sets, typically including hundreds of thousands of single nucleotide polymorphisms (SNPs), in combination with large public genealogy DNA databases in order to track biological relatives of an unknown donor by matching segments of shared DNA [1,8–10]. One of the key success elements is that only a fraction of the population of interest needs to be present in the database in order to be able to, in theory, identify every individual in the population by applying genetic genealogy methods. Erlich and colleagues [1] estimated that if 1% of the individuals, in the population of interest, are present in the genealogy DNA database there is more than 90% chance to find at least one 3rd cousin for every individual in the population.

DNA typing using microarrays is an easy, cheap and fast way to establish the genotypes needed. However, the use of microarrays normally requires larger amounts of DNA (in the order of hundreds of nanograms) [11]. Such a high amount of DNA is not always present in forensic samples, which instead may be in the order of nano- to

* Correspondence to: Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Artillerigatan 12, SE-58758 Linköping, Sweden.
** Correspondence to: National Forensic Centre, Swedish Police Authority, SE-58194 Linköping, Sweden.
E-mail addresses: andreas.tillmar@rmv.se, andreas.tillmar@liu.se (A. Tillmar), ricky.ansell@polisen.se (R. Ansell).

https://doi.org/10.1016/j.fsigen.2021.102525
Received 28 January 2021; Received in revised form 26 April 2021; Accepted 28 April 2021
Available online 8 May 2021

subnanogram levels. It is also not uncommon that forensic samples display various levels of degradation and enzymatic inhibition both of which most often have a negative effect on downstream analyses [12–14]. Although successful use of microarrays for IGG purposes has been demonstrated [2], progress in DNA sequencing technologies has significantly increased the possibility to process biological samples with degraded DNA of low quantity [15–17]. In this case study we used whole-genome sequencing for which standard protocols are available for as little as 50 pg of input DNA [18].

If the established SNP data set lacks observed genotypes for a large proportion of SNPs, missing genotypes can be inferred by methods referred to as genotype imputation [19,20]. The aim of genotype imputation is to predict and estimate genotypes for SNPs not typed in the sample. The basic idea is that any two individuals, including apparently unrelated, can share short segments of DNA from a distant common ancestor. Such DNA segments are shared IBD (identity by descent). Factors like high levels of linkage disequilibrium (LD) and low recombination rates within small stretches of chromosomal segments will conserve haplotype variants through generations. Shared segments can be found if the observed genetic variants, in the studied sample, are compared with variants from a panel of reference individuals (e.g. 1000 Genomes Project [21]). From these shared segments, prediction of the missing genotypes in the sample can be performed based on the observed genetic variants in the reference individuals. There are a wide range of software available and a large number of studies have been conducted to study the performance and accuracy of genotype imputation [22,23].

Although the application of IGG has been shown to be successful, critical concerns have been raised regarding its use for law enforcement purposes. These opinions include issues related to ethical and legal aspects [24–27] and the future use of IGG in the forensic field will therefore not only involve technical challenges. The US Department of Justices (DOJ) published an interim policy in September 2019 in which they, at a general level, described when and how IGG should be used, its limitations, how data should be administrated etc. [28]. In early 2020, recommendations were also published by the Scientific Working Group on DNA Analysis Methods (SWGDAM) on the use of IGG [29]. In addition to the reports in the US, similar reports have been published in Australia [30] and the UK [31] with respect to the potential use in these countries.

The aim of this paper is to summarize and report how IGG successfully was used in a pilot case study to solve a double murder cold case in Sweden. In this paper, we share and discuss details from all parts of the case study including legal and ethical considerations, the extended DNA-analysis, genealogy database searches, the succeeding genealogy and finally the conclusive lead which ultimately resulted in the closure of the second largest criminal investigation in Swedish history. We believe that a high degree of transparency including the disclosure of details, as in our paper, is important to get an informed and fact-based discussion within the forensic community as well as the public.

2. Legal considerations and case description

After the reporting in Swedish media of the successful use of IGG to catch the “Golden State Killer” the question was brought up within the Swedish Police Authority if this tool could be used also by Swedish law enforcement. A legal inquiry concerning the use of IGG in Sweden was initiated in May 2018 by the National Forensic Centre (NFC) and performed in cooperation with the Legal Affairs Department within the Police Authority.

The legal inquiry was finalized in January 2019 and covered a suggested method with defined case inclusion criteria, as well as legal considerations and a methodological framework [32]. As it turned out, the criteria and framework set up for the Swedish pilot case study was much in line with the interim policy later published by the US Department of Justice [28] as well as SWGDAM recommendations [29].

A data protection impact assessment was performed as part of the legal inquiry in accordance with Swedish and European Union laws and regulations. The possible infringement on privacy rights were estimated in the data protection impact assessment to encompass the person who left the DNA at the crime scene, users of the genealogy databases as well as their relatives. In the proportionality assessment, society’s interest of solving a major violent crime was assessed to carry more weight than the risks of infringement of privacy rights for the above mentioned persons or categories of persons.

The legal inquiry further dealt with aspects of the division of responsibilities between the different actors (NFC, the police crime investigators etc.) in regard to the different methodological steps. Genetic data refers to personal data relating to a person’s inherited or acquired genetic characteristics. In Swedish law genetic data is considered to fall within the scope of sensitive personal data. According to the judicial inquiry, a DNA analysis of a trace and the documentation surrounding the analysis are covered by the definition of genetic data. Subsequent processing of the documentation, on the other hand, constitutes a processing of personal data and not genetic data. Furthermore, it was stated that there is legal support for NFC to process genetic data for forensic purposes while other units within the Police Authority do not have such support. According to the judicial investigation, NFC also has legal support to contract other laboratories or other Swedish expert bodies for assistance with expertise.

The legal inquiry also dealt with the provision of absolute necessity to be able to use the method in a specific case as well as the issue of transferring personal data to a third country (i.e. a country outside the European Union). According to legislation, as described in the legal inquiry, the Swedish Authority for Privacy Protection (IMY) would need to be informed in writing after each data transfer to a third country since sensitive personal data information had been transferred. Other important prerequisites were that the database company would not be allowed to use the information received from the Swedish police for any other purpose than the requested searches in the database, and also that following completion of the work all of the data transferred to and processed by the database company should be possible to erase upon request.

Regarding the different steps in the method, ethical concerns were assessed to primarily arise in connection to the use of the commercial genealogy databases. Prior to the pilot case study the Police Authority’s ethical council[1] was thus consulted on the basis of the legal inquiry. The discussion in the council covered the methodology specified in the judicial inquiry as well as possible infringement on privacy rights. The council supported the continuation of the method development work with a so-called pilot case study.

Our aim was that a pilot case should cover a number of key steps, including: 1) to establish DNA datasets that could be used for searches in genetic genealogy databases, 2) to transfer DNA data and search in genealogy databases available for law enforcement (according to user terms and conditions and formal consent of other database users) and, 3) perform genealogy work based on information gathered from the database search, 4) generate investigative leads for the police investigation. In addition, the pilot case study as such also aimed to: 5) illustrate different aspects concerning the handling of sensitive personal data (genetic information), and 6) from both a technical and legal point of view evaluate the workflow as proposed in the legal inquiry.

Furthermore, a legal checklist was compiled. This checklist was a complement to the judicial inquiry and was written as a simpler user support to be used during the pilot case. The checklist was used primarily by NFC and to some extent by the crime investigators.

For the pilot case study a double murder cold case was selected. Early

[1] The Swedish Police Authority Ethical Council constitutes external experts appointed by the government for periods of four years. The Ethical Council has an advisory role and is headed by the National Police Commissioner.

in the morning October 19th, 2004 an 8-year-old boy was on his way to school when he was fiercely attacked by an unknown perpetrator who stabbed him to death. A 56-year-old woman had just come out of her home close by and witnessed the assault. The perpetrator then attacked her and she received several stab wounds. The attacks were fatal and both the boy and the woman died from their injuries. The murder weapon, a butterfly knife, was found left at the crime scene and seized for forensic examination.

During the forensic examination of the butterfly knife, the police's technicians and experts, from the National Laboratory of Forensic Science (SKL) the predecessor of today's NFC, found DNA traces from three persons. It was a mixture of DNA that matched the two victims as well as DNA from an unknown person. The unknown person's DNA was found on several additional exhibits seized in the case, together confirming the relevance of that trace.

The DNA profile from the unknown person has subsequently been searched for and was also searched continuously during the pilot case in the national DNA database as well as internationally against for example European countries through the Prüm Treaty. Extended analysis was carried out with multi-dimensional scaling (MDS) analysis based on 24 ancestry-informative autosomal SNPs and with four global reference populations, resulting in an assessment of the perpetrator's biogeographical origin to Western Eurasia. Y- and mtDNA-SNP analysis was also performed as well as hair and eye color predictions using the HIrisPlex system. Considering the results from the analysis it was concluded that the person likely was of European ancestry. From a knitted cap (probably worn by the perpetrator) left near the crime scene blond hair was recovered and witnesses testified that the perpetrator looked Swedish. Altogether, the investigation assumed the perpetrator to be of northern European origin. This assumption was not used for exclusion but for prioritizing between persons to interrogate. A familial search was also carried out in the national DNA database early 2019 without success. In the case, an extensive DNA-sampling of more than 6000 individuals had been carried out throughout the years (testing of selected persons was also ongoing during the project) and more than 9000 persons had been interrogated.

NFC decided, with the legal inquiry in mind, and together with the officer in charge of the murder investigation, that this murder case could be used in the pilot case study. The decision was based on the fact that relevant DNA traces were still available, and the extent of which available forensic DNA tools had been used to try and solve the case as well as case circumstances in general all met the criteria to justify IGG being tested in this specific case.

The case had already been discussed in connection with, and used as a case to "lean on", in the development of the proposed methodology in the legal inquiry. The methods used for the DNA analysis, database searches and genealogy are described in detail in the next section. After the work was completed NFC made requests to the genealogy database companies to delete all entered DNA data files, account information, etc. This was done by both the database companies involved and confirmed within a couple of days.

An evaluation report was written after the end of this pilot case. The work was presented as well as experiences gained, conclusions and suggestions for the continuous work. It was concluded, that under the right circumstances and conditions, the use of IGG can truly be an extremely powerful tool for Swedish criminal investigations although it must be used with extensive care. With the goal of having a high degree of transparency the evaluation report, written in Swedish, was made publically available in November 2020 [33]. The pilot case study was the product of a successful cooperation between different parts of the Swedish Police Authority, including the Legal Affairs Department, region Öst and the National Forensic Centre, together with expertise from the National Board of Forensic Medicine, an external laboratory and a contracted genealogist.

3. Material and methods

3.1. Samples and DNA extraction

DNA was extracted, with an in-house developed organic extraction method and/or a Chelex based extraction method, from three blood stains from a knitted cap that was found near the crime scene. The first and second WGS analyses were performed on one of these samples, "DNA extract 1″. The third WGS analysis was performed on a pool of DNA extracts from two other blood stains on the cap. This pool was further washed with EB buffer (Qiagen) and filtered using Amicon Ultra 2 (Merck Millipore). The pool is hereafter referred to as "DNA extract 2″. The DNA was quantified with Quantifiler® HP DNA Quantification kit (Thermo Fisher Scientific). The integrity of the DNA was analyzed with TapeStation (Agilent).

3.2. STR analysis and targeted SNP analysis

STR analysis was performed using the AmpFlSTR SGM Plus kit (Applied Biosystems) for DNA extract 1 and the PowerPlex ESX 16 Fast System kit (Promega) for DNA extract 2. The amplified products were analyzed using an ABI PRISM 3100 Genetic Analyzer (CE) instrument (Applied Biosystems) for DNA extract 1, and an ABI 3500 Genetic Analyzer (CE) instrument (Applied Biosystems) for DNA extract 2. GeneMapper ID Software v. 3.1 (Thermo Fisher Scientific) and GeneMapper ID-x Software v.1.6 (Thermo Fisher Scientific) were used for DNA extract 1 and 2, respectively.

In order to study the quality of DNA extract 1 prior to the WGS analysis, and to be able to cross-validate the WGS-established SNP data sets, a SNP profile comprising 131 targeted SNPs was obtained using massively parallel sequencing (MPS) as previously described by Grandell and colleagues [34]. In brief, the DNA library was constructed using the GeneRead™ DNAseq Targeted Panels V2 library preparation workflow (Qiagen) with the QIAseq Investigator 140 SNP panel primer set. One positive control (2800M Control DNA), an extraction blank and a PCR negative control were analyzed together with DNA extract 1. The sequencing was performed on a MiSeq FGx instrument (Verogen) with Reagent Kit v3. The bioinformatic analysis was performed from FASTQ files using the Biomedical Genomics Workbench v 2.1.1 (Qiagen). The minimum coverage for genotype calling was set to 200X and the heterozygote balance acceptable for genotype calling were based on allele read frequency (ARF) values as described earlier [34] (0.4–0.6 for a heterozygous genotype, 0–0.1 or 0.9–1 for a homozygous genotype, otherwise no genotype was called).

3.3. Whole-genome sequencing, bioinformatics and genotype calling

A total of three runs of whole-genome sequencing were performed. From DNA extract 1, 20 ng was used to prepare three libraries using the ThruPLEX® DNA-seq 48 S Kit (R400427, Takara Bio). No fragmentation of the DNA was performed prior to the library preparation and 6 PCR cycles were used for the library amplification step. This set of libraries ("DNA library 1.1″) was sequenced on a HiSeq X with v 2.5 sequencing chemistry (Illumina) using paired-end sequencing and 150 base pair (bp) read length. For the second run, another three libraries were prepared as above, except that 3 PCR cycles were used for the library amplification step. This set of libraries ("DNA library 1.2″) was sequenced on NovaSeq 6000 SP with v 1 sequencing chemistry (Illumina) using paired-end sequencing and 150 bp read length. From DNA extract 2, 20 ng was used to prepare three libraries using the SMARTer ThruPLEX® DNA-seq 48 S Kit (R400676,[2] Takara Bio). The DNA was fragmented to 350–400 bp, and the library preparation was performed

[2] The kit R400676 is a newer version of the kit R400427 with improvements of the reagents incorporated in the kit.

according to the manufacturer's protocol [18]. This set of libraries ("DNA library 2.1″) was sequenced on NovaSeq 6000 SP with v 1 sequencing chemistry (Illumina) using paired-end sequencing and 150 bp read length.

FASTQ files were used as the input for the bioinformatic analysis which was performed with the software Biomedical Genomics Workbench v 5.0.1 (Qiagen). Adapters were trimmed from the raw reads and low quality reads were removed. The remaining sequences were aligned to the reference genome hg19 (human_g1k_v37.fasta). The mapped sequences were locally re-aligned after which PCR duplicates were removed. The final SNP genotype calling was then performed with an in-house developed R script (R version 3.5.0, www.r-project.org). The genotype calling was based on the parameters and criteria presented in Supplementary Table 1.

In total, 1378,481 SNPs were selected as targets for the complete WGS datasets. This selection of SNPs comprises the commonly used SNPs by the larger DTC vendors and is described in more detailed in Tillmar et al. [5].

3.4. Genotype imputation

Genotype imputation was performed with the software Beagle 5.1 [19,35] on the first and second WGS datasets (from DNA library 1.1 and 1.2). Genotypes for approximately 6 million SNPs were used as observed genotypes, from which the missing genotypes were imputed. The software *Conform*[3] was applied prior to the imputation in order to check the file format, consistencies between the target and reference SNP definitions and allele definitions etc. Genotype data from the 1000 Genomes Project (including all populations, "ALL") was used as the reference dataset (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/). Beagle was run with the following parameter settings: burn in = 6, iterations = 12, phase-states = 280; imp-states = 1600, imp-segment = 6.0, imp-step = 0.1, imp-n steps = 7, cluster = 0.005, ap = true, gp = true, ne = 1000,000, window = 400 cM and overlap = 4.0. Imputed genotypes were added to the set of observed genotypes if the genotype imputation probability was equal to or above the threshold Q_{gp}. Different datasets were established for which the parameter Q_{gp} was set to 0.9, 0.95 or 0.99, respectively.

3.5. Database search and genealogy

The established autosomal genotype data sets were used for searches in GEDmatch and/or FTDNA databases. For the searches in GEDmatch, the datasets were uploaded according to GEDmatch site policy.[4] During the upload, GEDmatch was informed that the DNA was obtained and authorized by law enforcement to identify a perpetrator of a violent crime against another individual and "research" settings were used. For the searches in FTDNA, an application was sent to Gene by Gene / FTDNA in accordance with their "FamilyTreeDNA Law Enforcement Guide".[5] The case was accepted and a Terms of Service (TOS) contract was established, for this single case, and signed by both parties. Due to a change in their policy, FTDNA did not accept DNA datasets to be transferred from non US countries for a period of time. In February 2020 FTDNA informed that they now accepted this case, and the imputation datasets and later on also the third WGS dataset were transferred for searches in the FTDNA database.

Family pedigree building and other genealogy related work were primarily performed from the hit list produced by the last search in FTDNA. The vast majority of the top hits could be identified from their aliases and/or email addresses. From these individuals, family pedigrees were created using information from sources such as public address registers, ArkivDigital[6] and Riksarkivets digitala forskarsal.[7]

4. Results & discussion

The Results & Discussion section focus mainly on analytical and technical aspects encountered during the course of the pilot case work.

Regarding the DNA extract used, the DNA concentration was measured to 0.9 ng/μl and 10.8 ng/μl for DNA extract 1 and DNA extract 2, respectively. Analysis of the integrity of the DNA showed that the DNA was heavily degraded in both DNA extracts (Supplementary Fig. 1). Complete STR profiles were however obtained for both DNA extracts. The STR analysis also showed a single contributor. Genotypes for 129 out of the 131 SNPs in the MPS based targeted SNP assay met the quality criteria. The genotype calls were shown to have high coverage and to be well balanced (Supplementary Fig. 2).

The first WGS analysis resulted in a much lower coverage and higher duplication rate than expected (Table 1), and merely 155,000 SNPs met the quality criteria for the genotype calling (Table 2). Due to the highly degraded DNA, the median insert size was only around 60 bp. Thus, the 150 bp reads only partially contained the actual DNA of interest. Standard Chelex-based extraction methods generate single stranded DNA [36,37], however it is expected that a proportion of the DNA is renaturated after extraction, resulting in a mixture of double and single stranded DNA. As the library preparation starts from double stranded DNA it may have had an influence on the performance of the WGS. Also, impurities in the Chelex extract might have influenced the process [14,38]. Furthermore, a relatively high degree of genotype errors were estimated when cross-validating this first WGS SNP dataset with the genotypes obtained from the targeted SNP assay. However, this estimate was interpreted with much care since only seven SNPs (out of the 131 SNPs) met the quality criteria in the WGS SNP data set. One of these genotypes was an allelic drop-out in the WGS SNP dataset (A/G in the targeted assay, A/A in the WGS SNP dataset). The conclusion from this analysis was that the WGS SNP dataset, and the potential results from database searches, should be handled with care. The search in GEDmatch, with this dataset, resulted only in very distant relatives (less than 30 cM in total shared segment lengths for top hits) from which limited genealogy work was performed.

Genotype imputation was applied on the dataset in order to increase the number of genotypes. Approximately 6 million genotypes were called and used as the observed genotypes from which the missing SNP genotypes were imputed. After the imputation, the number of genotypes increased from roughly 155,000 to 864,000 (Table 2). Interestingly, the estimated genotype error rate decreased at the same time to approximately 6%. The search in GEDmatch resulted, despite the many more SNPs, in a similar matching pattern as for the initial dataset (e.g. less than 30 cM in total shared segment lengths for top hits).

Next, a new WGS was performed on a replicate library preparation (DNA library 1.2). The output was similar as for the first WGS run

Table 1
Summary statistics from the WGS runs.

Parameter	WGS run 1 (DNA library 1.1)	WGS run 2 (DNA library 1.2)	WGS run 3 (DNA library 2.1)
Average Coverage	13X	10X	60X
Duplication rate	~ 70%	~ 60%	< 10%
Median insert size	~ 60 bp	~ 60 bp	~ 180 bp

[3] https://faculty.washington.edu/browning/conform-gt.html
[4] https://www.gedmatch.com/tos.htm
[5] https://www.familytreedna.com/legal/law-enforcement-guide
[6] www.arkivdigital.se/
[7] sok.riksarkivet.se/digitala-forskarsalen

Table 2
Summary of the established dataset and database searches.

Sample/ library preparation	Dataset	Database	Number of genotypes[a] (approx.)	Total shared segment length for top hits
DNA library 1.1	WGS analysis	GEDmatch	155,000–269,000	Less than 30 cM
DNA library 1.1	WGS analysis and genotype imputation	GEDmatch	864,000–1026,000	Less than 30 cM
DNA library 1.1 + DNA library 1.2	WGS analysis and genotype imputation	GEDmatch	908,000–1050,000	Less than 30 cM
DNA library 2.1	WGS analysis	GEDmatch	1279,000	Less than 30 cM
		FTDNA	1861,000	~350 cM, ~100 cM, ~60 cM and decreasing

[a] The range represents subsets for which different quality thresholds had been applied. The smallest number, in each range, represents the dataset with the most stringent quality thresholds.

regarding coverage and duplication rate (Table 1). The output reads were therefore merged into the first dataset and a new imputation round was performed. This resulted in a slightly increased number of genotypes (Table 2) with an overall genotype error rate of approximately 4%. However, the search in GEDmatch still resulted in a similar matching pattern as with the initial dataset (e.g. less than 30 cM in total shared segment lengths for top hits). Despite the absence of close relatives, the matching lists were analyzed by the contracted genealogist and a cluster of potentially distant relatives originating from northern Germany was discovered. This trail proved hard to investigate further, and other actions were eventually made in order to bring the investigation forward. It should be noted that any German origin was not observed in the succeeding and final analyses. One can conclude that there is a certain risk that an investigation be led in the wrong direction by the genealogy searches and for future work it is important to establish quality parameters and thresholds for the genealogy data analysis.

From this later dataset, a separate SNP dataset was established and sent to FTDNA for evaluation. This dataset did however not meet FTDNA's internal quality evaluation and were therefore not used in any search.

Due to absence of useful hits, it was decided to start over with the analyses using a completely new DNA extract. The new WGS analysis (DNA library 2.1) resulted, in contrast to the previous attempts, in a high coverage dataset with a low duplication rate (Table 1). The genotypes for approximately 1.3 million and 1.9 million SNPs were called (GEDmatch and FTDNA SNPs, respectively) and no genotype errors were detected when compared with the targeted 131 SNPs. Searches were performed in GEDmatch and FTDNA. Despite the now apparently good WGS SNP dataset still no close relatives were found in GEDmatch and, as with previous searches, all top hits had less than 30 cM in total shared segment lengths. The search in FTDNA was however more fruitful and yielded several hits that were used for family pedigree building and genealogy. In total, 890 hits were obtained in the first search of which the top 28 individuals were used for the genealogy analyses (top two hits shared about 60–100 cM with the unknown and later turned out to be 2nd cousin once removed and 3rd cousin once removed to the perpetrator). Family pedigrees were built back to the late 18th century, in search of common ancestors, and matching DNA segments were mapped looking for triangulation. During the process, 15 volunteers with known origin from a specific part of Sweden (that emerged as of high interest due to the genealogy work performed) provided their DNA samples to FTDNA whereas one of them turned out to be a closer match (shared about 347 cM in total) [33]. From the subsequent mapping of descendants of the common ancestors, including investigative information such as year of birth, a pair of brothers remained as candidates to be the unknown perpetrator. Buccal swabs were subsequently, following prosecutors decision, obtained from both brothers and with comparative routine STR profiling, one of these brothers was confirmed to match the crime scene sample. The suspect confessed and was later convicted for the double murder.

In this case report we describe how IGG was used to obtain conclusive investigative leads which led to the arrest and conviction of the perpetrator of a double murder cold case. From many aspects, this was not a trivial case. Not only due to legal and ethical challenges but also due to the application of DNA analysis methodologies not normally used in forensic genetic analysis.

From a technical point of view there were two main obstacles that needed to be addressed in order to take the investigation forward. Both of which were due to the limited quality and the low quantity of the DNA. Firstly, the large number of SNPs with missing genotypes. When performing searches in genealogy databases, with a limited number of SNPs, there is a risk that the output does not match the expected pattern as if one had searched with a complete SNP dataset. A low SNP density could result in a decreased total segment length since some shared segments will not meet the SNP density criteria and thus will not be included and added to the total shared segment length estimate. The consequence of this could be that true close relatives may be estimated as more distant relatives and therefore be ignored (e.g. considered not worth further investigation) by the genealogist. Ultimately, distant relatives may go completely undetected. In contrast, false positives could appear if the SNP density threshold is set too low, which can result in falsely shared segments being added to the total shared segment length [9]. These issues were discussed with the genealogist in the team and database searches were performed with this in mind so that the obtained shared cM was interpreted with care.

The second obstacle was the presence of genotype errors in the established datasets. As noted above, since we had the possibility to cross-validate the WGS SNP datasets, with the genotypes from the targeted SNP assay, we could get a rough estimate of the proportion of genotype error and also the type of error (e.g. allelic drop-out [false homozygous], allelic drop-in [false heterozygous] etc.). We believe that it is crucial to have such a possibility to assess the quality of the established WGS SNP dataset to be able to make informed decisions for the intended usage. When it comes to the impact of genotype error for the segment sharing estimations, different types of errors will have different impacts. A larger proportion of false heterozygous genotypes may create false matching segments and/or too long matching segments, and thus increase the risk of false positive hits. False homozygous genotypes may, on the other hand, prematurely terminate shared segments and could cause false negatives, in a similar way to the low SNP density situation. In our case, the majority of the errors comprised of allelic dropouts, which is expected when dealing with forensic samples of low quality and quantity [39,40]. Most of the segment analysis algorithms do however allow mismatches of this type to a certain degree. In GEDmatch, for example in the one-to-one tool, the user may define the mismatch threshold manually depending on the quality of the dataset. If allelic drop-outs are not taken into account close relatives would share smaller segments than expected (because of breakdown of matching segments due to lack of allele sharing). Also, some of the shared segments can, due to the same reasoning, become shorter than the cutoff for a single segment to be included in the total shared segment length, and therefore a close relative may appear more distant than expected. Similarly, distant relatives may go undetected. As previously noted, searches were made with this in mind, and discussed with the genealogist beforehand.

Since no useful hits were obtained for the first three rounds of searches in GEDmatch we could not, at the time of the searches, exclude the risk of not detecting any relatives due to the low SNP density or due to the observed genotype error rate. But since useful hits were not found

with the last SNP dataset either, we can conclude that the absence of hits was most probably due to the absence of relatives in the GEDmatch database, and not due to low SNP density or genotype errors in the WGS datasets. The absence of useful relatives in GEDmatch highlights, however, another important aspect; the content and the size of the databases available for law enforcement searches. The GEDmatch database is heavily weighted towards individuals living in the US [7], and the number of individuals who have "opted in" is currently (December 2020) only around 325,000 [41]. Verogen, the owner of GEDmatch, has recently released a law enforcement dedicated portal, GEDmatch Pro (https://pro.gedmatch.com/). Verogen states that this new portal "… separates police comparisons of GEDmatch data from standard genealogy activities and offers a range of tools most relevant to help further investigations." This portal is also likely to include measures related to the previously identified security breaches [42,43]. Hopefully, the transformation of GEDmatch will also attract new users, some of whom will allow law enforcement searches.

5. Summary

Herein, a historical part of Swedish criminal investigation and forensic science has been reported. With the support of a legal inquiry, a pilot case study was successfully completed with the use of one of the most powerful emerging tools in police work, investigative genetic genealogy. The assessment made in the case was that the disclosure of DNA data to genealogy databases available for law enforcement use, from a legal standpoint was an absolute necessity to move the investigation forward. The handling of sensitive DNA data was in this specific case considered proportionate and to outweigh the risks for infringement on privacy rights. The DNA data used was protected by security measures and searches in databases were made with regard to user privacy and rules set for law enforcement. Upon completion, all DNA and other processed data was removed from the databases. During this work many challenges were identified (legal, ethical and technical). Although they were resolved in this case, many issues still remain and will return in future cases. Evaluation of the use of IGG in other Swedish criminal cases is currently in progress within the Swedish Police Authority, in cooperation with the National Board of Forensic Medicine. This will include the set-up of national guidelines that cover criteria and conditions on the cases to be selected, properties and characteristics of the DNA, expectations based on the database composition and possibility to perform genealogy work (availability of national records etc.). Such national guidelines, with preset conditions and criteria, is one way forward to transparently control and balance the utilization of this important tool so that IGG can be used in accordance with user privacy, database user terms and conditions, only when absolutely necessary and taking legal aspects and ethical concerns into account. It is also relevant to do a cost-benefit analysis as there will be a significant cost associated with analyzes and work to perform IGG. This must however be weighed against the cost for other investigative measures and the benefit of resolving cold cases.

Acknowledgments

We would like to thank, Sara Markstedt (Legal Affairs Department, Swedish Police Authority) for her invaluable work on the legal inquiry, David Kummel (Legal Affairs Department, Swedish Police Authority) as well as Helena Trolläng (National Forensic Centre) for their support and contributions during the course of the case study, and the contact staff members at FTDNA and GEDmatch for valuable support. We would also like to thank two anonymous reviewers for their constructive comments which improved the manuscript.

Whole genome sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102525.

References

[1] 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.
[2] E.M. Greytak, C. Moore, S.L. Armentrout, Genetic genealogy for cold case and active investigations, Forensic Sci. Int. 299 (2019) 103–113.
[3] S.H. Katsanis, Pedigrees and perpetrators: uses of DNA and genealogy in forensic investigations, Annu. Rev. Genom. Hum. Genet. 21 (2020) 535–564.
[4] 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.
[5] 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.
[6] 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.
[7] D. Kling, C. Phillips, D. Kennett, A. Tillmar, Investigative genetic genealogy: current methods, knowledge and practice, Forensic Sci. Int. Genet. 52 (2021), 102474.
[8] 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) 34267.
[9] 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.
[10] 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), 0160287.
[11] A. Patel, S.W. Cheung, Application of DNA microarray to clinical diagnostics, Methods Mol. Biol. 1368 (2016) 111–132.
[12] R. Alaeddini, S.J. Walsh, A. Abbas, Forensic implications of genetic analyses from degraded DNA–a review, Forensic Sci. Int. Genet. 4 (3) (2010) 148–157.
[13] T. Lindahl, Instability and decay of the primary structure of DNA, Nature 362 (6422) (1993) 709–715.
[14] M. Sidstedt, P. Radstrom, J. Hedman, PCR inhibition in qPCR, dPCR and MPS-mechanisms and solutions, Anal. Bioanal. Chem. 412 (9) (2020) 2009–2023.
[15] S.E. Levy, R.M. Myers, Advancements in next-generation sequencing, Annu. Rev. Genom. Hum. Genet. 17 (2016) 95–115.
[16] S.T. Park, J. Kim, Trends in next-generation sequencing and a new era for whole genome sequencing, Int. Neurourol. J. 20 (Suppl 2) (2016) S76–S83.
[17] B.S. Petersen, B. Fredrich, M.P. Hoeppner, D. Ellinghaus, A. Franke, Opportunities and challenges of whole-genome and -exome sequencing, BMC Genet. 18 (1) (2017) 14.
[18] ThruPLEX® DNA-Seq Kit User Manual. Takara Bio USA "Available from: https://www.takarabio.com/".
[19] 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.
[20] J. Marchini, B. Howie, Genotype imputation for genome-wide association studies, Nat. Rev. Genet. 11 (7) (2010) 499–511.
[21] A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J. L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.
[22] 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.
[23] S. Das, G.R. Abecasis, B.L. Browning, Genotype imputation from large reference panels, Annu. Rev. Genom. Hum. Genet. 19 (2018) 73–96.
[24] C. Phillips, The Golden State Killer investigation and the nascent field of forensic genealogy, Forensic Sci. Int. Genet. 36 (2018) 186–188.
[25] G. Samuel, D. Kennett, The impact of investigative genetic genealogy: perceptions of UK professional and public stakeholders, Forensic Sci. Int. Genet. 48 (2020), 102366.
[26] D. Syndercombe Court, Forensic genealogy: some serious concerns, Forensic Sci. Int. Genet. 36 (2018) 203–204.
[27] R.A. Wickenheiser, Forensic genealogy, bioethics and the Golden State Killer case, Forensic Sci. Int. 1 (2019) 114–125.
[28] Interim Policy on Forensic Genetic Genealogical DNA Analysis and Searching, (2019). Available from: ⟨https://www.justice.gov/olp/page/file/1204386/download⟩.
[29] SWGDAM, Overview of Investigative Genetic Genealogy, (2020). Available from: ⟨https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_6cc9e7c82ccc4fc0b5d10217af64e31b.pdf⟩.
[30] N. Scudder, R. Daniel, J. Raymond, A. Sears, Operationalising forensic genetic genealogy in an Australian context, Forensic Sci. Int. 316 (2020), 110543.

[31] BFEG Group, Should we be making use of genetic genealogy to assist in solving crime?, (2020). Available from: ⟨https://www.gov.uk/government/publications/use-of-genetic-genealogy-techniques-to-assist-with-solving-crimes/should-we-be-making-use-of-genetic-genealogy-to-assist-in-solving-crime-a-report-on-the-feasibility-of-such-methods-in-the-uk-accessible-version⟩.
[32] DNA-spår och släktforskning, Legal inquiry, the Swedish Police Authority, A637.388/2018, (2019).
[33] S.A. Fagerholm, R. Ansell, A. Tillmar, J. Staaf, Pilot: Dna-spår och släktforskning, (2020). Available from: ⟨https://nfc.polisen.se/om-nfc/nyhetsarkiv/2020/november/dna-baserad-slaktforskning-kan-bli-nationellt-anvand-metod⟩.
[34] 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.
[35] 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.
[36] J. Casquet, C. Thebaud, R.G. Gillespie, Chelex without boiling, a rapid and easy technique to obtain stable amplifiable DNA from small amounts of ethanol-stored spiders, Mol. Ecol. Resour. 12 (1) (2012) 136–141.
[37] N. Simon, J. Shallat, C. Williams Wietzikoski, W.E. Harrington, Optimization of Chelex 100 resin-based extraction of genomic DNA from dried blood spots, Biol. Methods Protoc. 5 (1) (2020) 009.
[38] P.S. Walsh, D.A. Metzger, R. Higuchi, Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material, Biotechniques 10 (4) (1991) 506–513.
[39] J.M. Butler. Advanced Topics in Forensic DNA Typing: Methodology, 3rd ed., Elsevier Science, 2011.
[40] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Sci. Int. Genet. 3 (4) (2009) 222–226.
[41] M. Molteni, Cops are getting a new tool for family-tree sleuthing, (2020). Available from: ⟨https://www.wired.com/story/cops-are-getting-a-new-tool-for-family-tree-sleuthing/⟩.
[42] P. Ney, L. Ceze, T. Kohno, Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference. Network and Distributed System Security Symposium (NDSS), (2019); Available from: ⟨https://dnasec.cs.washington.edu/genetic-genealogy/ney_ndss.pdf?fbclid=IwAR292NYdtPKb_3yXJewUtD3PzYFBbOxpZ2S_P7UySSblofBsj3ppZ17wWZc⟩.
[43] M. Taylor, Two Security Breaches at GEDmatch Open All Users' Profiles to Law Enforcement, (2020). Available from: ⟨https://www.forensicmag.com/566543-Two-Security-Breaches-at-GEDmatch-Open-All-Users-Profiles-to-Law-Enforcement/⟩.