Forensic Engineering Science: Developing tools for human identification

By

Shaun Fernando & Harrison Boyce

ENG 4001 Research Project

Progress report submitted for the degree of

Bachelor of Engineering (Honours)

in

Electrical and Electronic Engineering

Faculty of Science, Engineering and Technology

The University of Adelaide

Project ID 2022s1-EEE-UG-13223

Supervisor: Professor Derek Abbott

Collaborator: Colleen Fitzpatrick

Date submitted: 30/05/2022

Pages in main-matter content:16

1. Executive Summary

Forensic scientists working to discover the identities of unidentified remains often use genealogy as part of their investigations. A key method in the genealogical process is genetic analysis, commonly through the use of files containing key information about individuals related to the deceased. These files, called Genealogical Data Communication (GEDCOM) files can be quite unintuitive to look through as they can contain thousands of events such as births and deaths of relatives as well as many other less useful pieces of information. Due to the nature of GEDCOM files being designed for use as just a way of transferring information between programs, the files are not useful without a way to better display the information. This project aims to create a program capable of taking several GEDCOM files and producing an interactive map displaying the locations of birth and death events, displayed using Google Earth. By creating an interactive, visual experience, potential locations of interest can be identified as potential birth locations of the deceased much more easily than via viewing the data in other mediums such as GEDCOM files or in Comma-Seperated Values (.csv) files such as those used in Microsoft Excel. To produce this program, three necessary subsystems were identified. These are; a system to parse the input GEDCOM files and create a .csv file, a system to translate from geographic information such as city and state names to co-ordinate data, and finally a system to produce the interactive map of markers showing each location noted in the original GEDCOM files. Programs currently exist to do these steps individually but are somewhat unintuitive and the need to bounce between multiple programs and multiple files results in the process being much less productive than it otherwise would be. For this reason the project has created a plan to combine these three subsystems together so a user simply adds the GEDCOM files needed and receives the interactive map. The parser subsystem is to be created in Python and will operate by assessing each line of the GEDCOM file to determine if it contains relevant information about an event and then writing this to a .csv file via a writer object using the writerow() function for all of the data on each individual event. This results in a .csv file where each row is a list of information about an individual event, including but not limited to, name of person, type of event, date, location and GEDCOM identification number. This location information data in this .csv file is then ran through the Google Geocoding API to convert from location names to co-ordinate data. A sanity check is done on the resulting co-ordinates to ensure the API has determined a reasonable location based off of the other locations in the GEDCOM file. This is necessary due to the existence of multiple locations having the same name when not enough information is supplied, ie. “Perth” could be either “Perth, Western Australia” or “Perth, Scotland”. The end user is given the choice to either include the selection made or to ignore it if the location is unreasonable. With this .csv file then being converted into the file format used to display geographic data in Google Earth, Keyhole Markup Language (KML), the final interactive map is produced, including all relevant information about each event in the initial GEDCOM files, viewable in a visual medium. As of now the project team have created a basic prototype for the third subsystem and have created thorough plans for the design and completion of the first two subsystems. Working alongside project collabotaror and expert in the field Colleen Fitzpatrick will ensure the program meets all requirements and is user friendly.

1. Table of Contents

2. Abbreviations

DNA - Deoxyribonucleic acid
SNP- Single nucleotide polymorphism
KML- Keyhole Markup Language
GEDCOM- Genealogical Data Communication
CSV- Comma-Seperated Values
STR- Short Tandem Repeat

3. Key Words

DNA, GEDCOM, KML

4. List of Figures

5. List of Tables

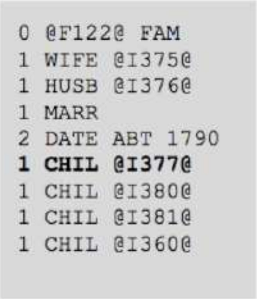
1. Introduction

The following section gives an introduction to the topic of Forensic engineering science and discusses about the aim, motivation, objectives and the overview of the project.

1.1 Background

Human identification has been present for a prolonged period of time. Presently, some of the biometric technologies such as fingerprint, DNA sequence matching, face recognition ,iris identification and retina identification are used to identify humans [3]. The main identification methods used in the present are fingerprint analysis, dental analysis and DNA analysis [4]. However, finding distant relatives of victims is hard with the use of the current methods used to identify humans [1]. Furthermore, techniques used in the past such as visual identification take a large amount of time compared to the rest of the methods [2]. Hence, the use of a software to automate the data visualisation process would reduce the time spent to do the process manually. Furthermore, the use of location markers aids in determining where family branches which the user needs cluster geographically. Similarly, the use of Google Earth simplifies the process as it has the capability of having a timeline of when the family branches clustered during different years. This would aid in simplifying the process of human identification.

Genetic genealogy which is the use of DNA testing in combination with genealogical research documents and historical documents [6]. Genetic genealogical databases have been used in the present to identify historical human remains [5] , human trafficking, disaster victims, criminal activities and to identify unknown human individuals [7]. Genetic genealogy is used in these cases due to it having high amounts of information when compared with DNA databases which are currently available [7].



```
0 @F122@ FAM
1 WIFE @I375@
1 HUSB @I376@
1 MARR
2 DATE ABT 1790
1 CHIL @I377@
1 CHIL @I380@
1 CHIL @I381@
1 CHIL @I360@
```

Figure 0.1 The GEDCOM records of Amalie Dranchmann where the bold text shows the family she was born into and is listed as a child [9, Fig.2 , p.45]

The standard filer type for genealogical data is called GEDCOM [9]. These files were first created in 1984 by The Church of Latter-day Saints [8]. Before GEDCOM files were created people used techniques such as rekeying and direct import [8]. The rekeying technique was stopped as databases started increasing as it was a time consuming process and prone to errors where the applicants for genealogy were required to print data from their old application and rekey it onto a new one [8]. Hence, currently the GEDCOM file type is used which consists of a plain text and unique number system for each individual added to the database [9]. The following numbering system and the text format of GEDCOM files are and prone to errors where the applicants for genealogy were required to print data from their old application and rekey it onto a new one [8]. Hence, currently the GEDCOM file type is used which consists of a plain text and unique number system for each individual added to the database [9]. The following numbering system and the text format of GEDCOM files are further shown clearly in Figure 1.1 where Amanda is a child of a family of three which consists of a husband, wife and three children with unique numbers as shown.

1.2 Motivation

Visual depictions of geographic data have been regarded as a vastly superior medium for analysing geneologic history. The project was motivated by an industry need for a simple, self contained program capable of displaying the relevant geographic data associated with persons listed in a given GEDCOM file as no such programs currently exist.

1.3 Aims and Scope

The project aims to produce a user friendly program capable of taking in a GEDCOM file and producing an interactive map displaying the locations and information of each individual listed in the GEDCOM file.

1.4 Objectives

The project contains 4 core objectives. The completion of these objectives will act as milestones for the project. The objective (in order of importance) are as follows

1.4.1 - Objective 1: GEDCOM to Comma-Seperated Values (CSV) file Parser

The program created must include a means of a parsing GEDCOM file to produce a CSV file with the relevant information. This includes Name, Date, Location, GEDCOM ID and Event Type.

1.4.2 - Objective 2: Get Coordinates from Location

The program created must be capable of converting from written locations (eg. Adelaide, South Australia) to coordinates.

1.4.3 - Objective 3: Conversion to Keyhole Markup Language (KML)

The program must be capable of converting the above CSV file to a KML file for use in Google Earth.

1.4.4 - Display in Google Earth

The program must be capable of displaying an interactive map via a Google Earth type program, the details of which is given in the KML file.

1.5 Overview

This document reviews the progress the project team have made so far, having produced a basic but functional Google Earth display as required in Section 1.4.4 as well as the investigations into solutions for the objectives in Section 1.4.1 and 1.4.2. These possible solutions include altering existing GEDCOM parsers and off the shelf Geocoding solutions such as the Google Geocoding API. These methods will be further developed in the final report. The document also includes a completion plan giving the timeline and methods for the completion of the project.

2. Literature Review

The following chapter provides information that is related to the topic of genetic genealogy. As this is a fairly new technology there are not a lot of literature available on it. Hence, the following section consists of related material to the topic of genetics and genealogy and gives a brief overview about the existing technologies that are present.

2.1 Existing Technologies

Currently there are no programs available that are able to complete the full conversion from GEDCOM to a visual representation. However, there do exist products capable of doing individual steps. These existing technologies have been extremely helpful for the project, both as examples to pull information from to create better programs but also as potential options to incorporate into the project itself. The existing technologies that are useful to the project can broadly be separated into three categories, GEDCOM parsers, Geocoding solutions, and Google Earth projects. Each of these categories give information toward the completion of objectives 1 through 3 respectively.

Many GEDCOM parsers exist, for varying languages and to obtain various details. Github alone hosts hundreds of repositories of GEDCOM parsers in various languages and many formats for the parsed information [10]. Many of these parsers do not retain the information required for this project and are thus not suitable for the project. However several projects such as this [10]by joephayes create basic .csv files with the relevant information. A better example of an existing GEDCOM parser is the GEDxlate program from GEDmagic. This program takes a GEDCOM file and translates the data into a variety of formats, including .csv. The program runs quickly and obtains all information required. However this program is still only capable of doing the conversion and thus would not satisfy the aim of the project.

For Geocoding, there are many possible solutions. As with GEDCOM parsers many examples exist on repository sites such as Github but the broadest and most reliable solution found was the Google Geocoding API. Calls to this API convert addresses, both specific and general to geographic coordinates which can then be used to place markers on a map. Beyond the simple geocoding, there are options available for filtering by country, postal code and more, as well as the ability to bias toward certain regions. This API perfectly covers everything required for this objective and with its built in synergy with both Google Earth and Google Maps this is a potential solution for this objective in the project. With a billing cost of 5 USD per 1000 calls the program created would thus have to have some way of financing itself if this existing technology were to be implemented.

For the visualisation of the geographic information, Google Earth was recommended to be used by both the supervisor and collaborator of the project. The program Earth Point has been used by geneologists to display the geographic information obtained from GEDCOM files and is quite robust, being capable of cycling though the location markers as a timeline is moved and separating the markers into related clusters. However, this Google Earth Tool is only functional for the United States and therefore lacks some of the functionality desired from this project. The collaborator for this project, Colleen Fitzpatrick has said she is happy with the Earth Point program and stated she is ok with that being used as a separate program to complete the third objective of the project if needed.

While no programs currently exist capable of meeting all objectives of the project, a program was previously available that could cover the first three objectives, and combined with Earth Point was a viable solution for experts in the field such as Colleen Fitzpatrick. Unfortunately from speaking with Colleen the project team found this program, Gen Detective, was no longer functional in the way it previously had been and was thus no longer viable for her work. As such the project team began further investigations into existing literature in the field to better understand what was required for the project.

2.2 DNA and genealogy

In terms of genetic genealogy, the use of genetic genealogy aids in creating a hierarchy of family or biological relationships between individuals. Furthermore, genetic genealogy can also determine the type of relationship between individuals [11]. Genealogists have the option of testing three different types of information in DNA when looking for genetic connections between individuals which consists of Y-chromosomal DNA, mitochondrial DNA and autosomal DNA. Mitochondrial DNA consists information of both males and females [21].

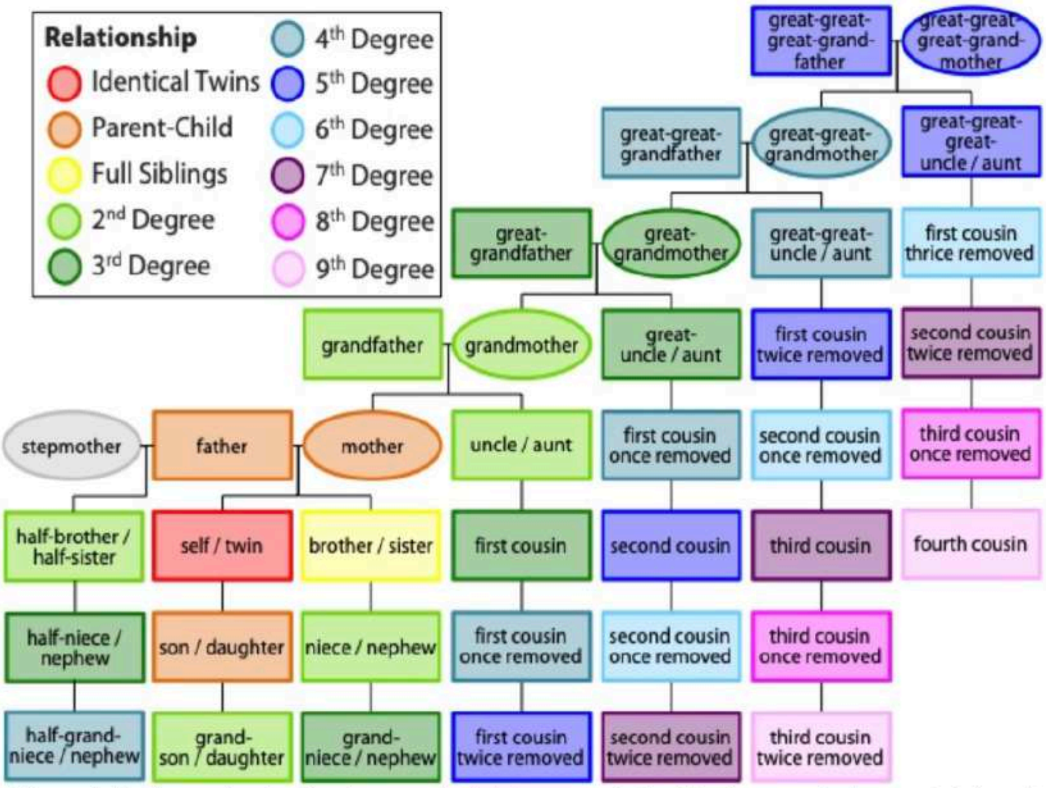


Figure 1: Pedigree showing the degrees of relatedness, as defined by the expected amount of shared DNA. Each relationship is defined with respect to the red "self / twin" box.

Figure 2.1 The degree of relatedness compared with self/identical twin [6, p.4]

A common ancestor who has the similar genetic characteristics can be found with the use of mitochondrial DNA and Y-chromosomal DNA. Furthermore, both DNA types can also be used to identify migration [11]. The three types of information in DNA mentioned play a huge role in determining the locations in which an individual's ancestors lived in and the type of relationship the individual has with them. As seen in Figure 2.1, a hierarchy of family and their relation to one another can be produced as such with the use of genetic genealogical databases which is highly beneficial [6]. Hence, in the context of Figure 2.1, it can be seen that the great-great-great grandfather of the identical twin can be displayed as such and the relations between other members can also be displayed.

2.2.1 Y Chromosomal DNA

Y-chromosomal DNA only consists information on males. Hence, the information in Y chromosomal DNA can be used to determine a male individual's paternal descent [11]. Figure 2.2 shows an overview on how paternal descent can be found with the use of Y chromosomal DNA.

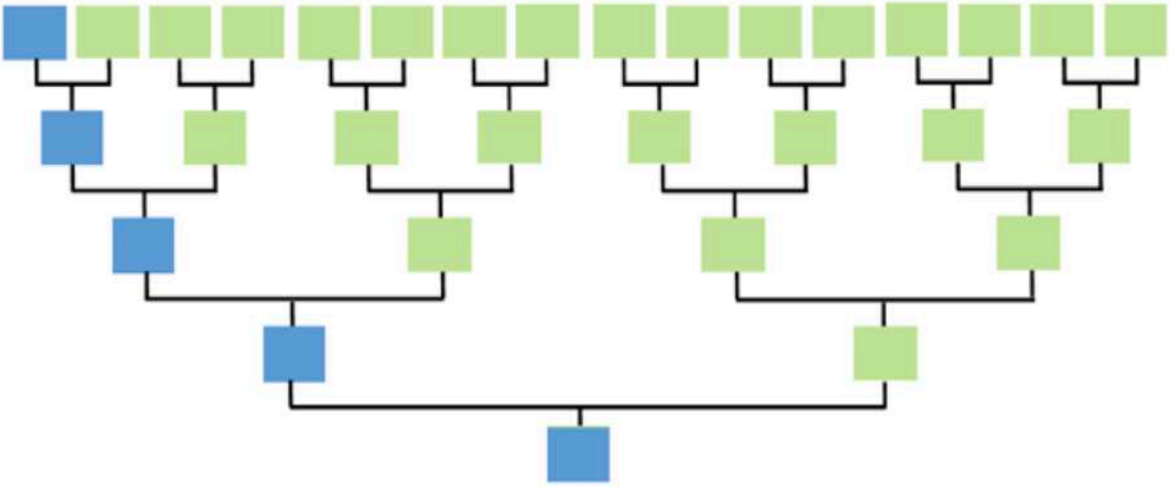
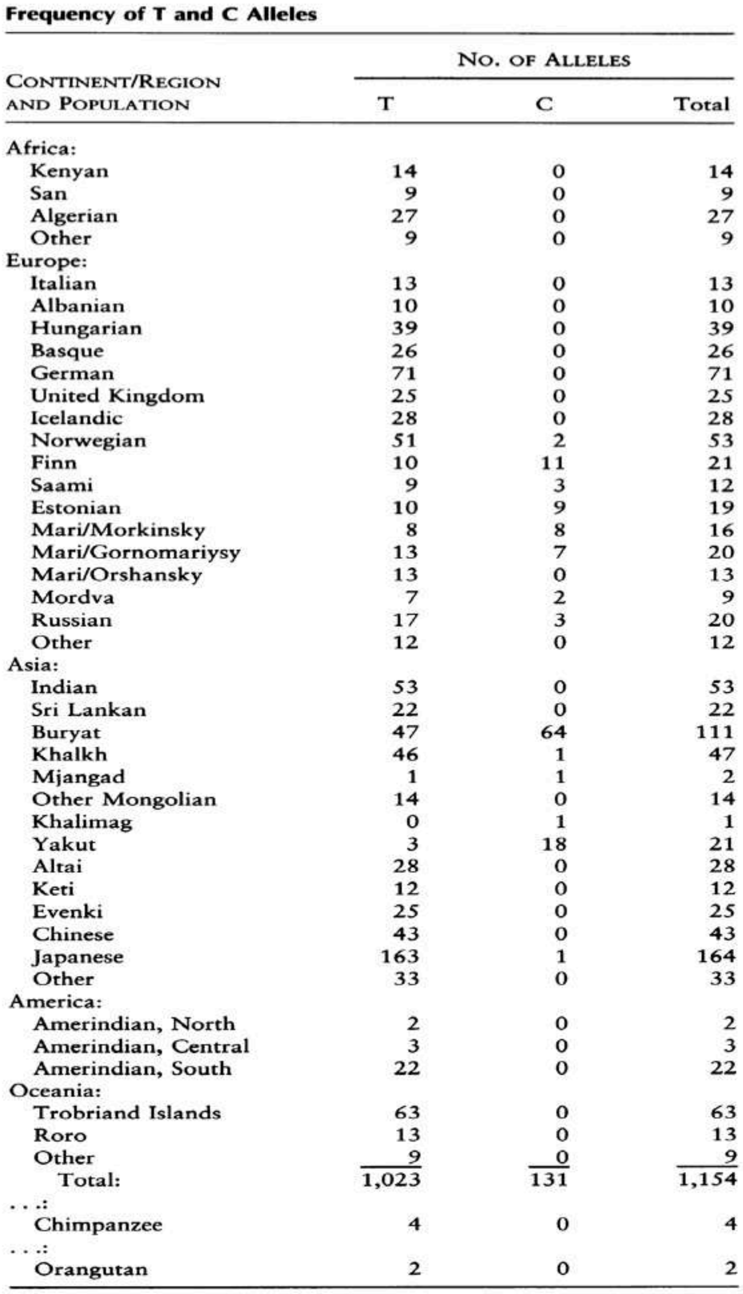


Figure 2.2 Overview Paternal descent which is shown in the colour Blue [17, p.1]

The information in the Y chromosomal DNA can also be used to study the spread of humans who live in the present in respect to humans lived in the past [12]. All individual gene also consists of allele which are considered as an alternate form of a given gene [14]. The Y chromosomal T allele to C allele transition are mostly restricted to individuals in Asia and Europe [15]. With the use of the following information in Y chromosomal DNA genealogists and archeologists suggest that parts in East and north Asia have been continuously inhabited during the past 35,000 years [15]. Moreover, the research done by the genealogists show that the use of Y chromosomal DNA aids in comparing the maternal and paternal information which aided in providing evidence of Asian paternal contribution to northern European populations [15].

The study done by T. Zerjal et al. [15] consisted of 1154 males and as seen in Table 2.1 it can be seen that there is a difference in T and C allele amounts in different populations in different countries. Furthermore, as seen in Table 2.1 populations in southern Asian, southern European, American, African and Oceanic only contained T allele. Both T and C alleles were found in Asian and northern European populations as seen in Table 2.1 [15]. Hence, the study provides evidence which is required to show Y chromosomal DNA aids in finding the distribution of the male population by comparing information from the paternal and maternal information and by comparing the results of the amount of C and T alleles.

Table 2.1 C and T allele distribution in different continents and regions [15, p.1176]



Frequency of T and C Alleles

Continent/Region and Population	No. of Alleles T	C	Total
Africa:			
Kenyan	14	0	14
San	9	0	9
Algerian	27	0	27
Other	9	0	9
Europe:			
Italian	13	0	13
Albanian	10	0	10
Hungarian	39	0	39
Basque	26	0	26
German	71	0	71
United Kingdom	25	0	25
Icelandic	28	0	28
Norwegian	51	2	53
Finn	10	11	21
Saami	9	3	12
Estonian	10	9	19
Mari/Morkinsky	8	8	16
Mari/Gornomariysy	13	7	20
Mari/Orshansky	13	0	13
Mordva	7	2	9
Russian	17	3	20
Other	12	0	12
Asia:			
Indian	53	0	53
Sri Lankan	22	0	22
Buryat	47	64	111
Khalkh	46	1	47
Mjangad	1	1	2
Other Mongolian	14	0	14
Khalimag	0	1	1
Yakut	3	18	21
Altai	28	0	28
Keti	12	0	12
Evenki	25	0	25
Chinese	43	0	43
Japanese	163	1	164
Other	33	0	33
America:			
Amerindian, North	2	0	2
Amerindian, Central	3	0	3
Amerindian, South	22	0	22
Oceania:			
Trobriand Islands	63	0	63
Roro	13	0	13
Other	9	0	9
Total:	1,023	131	1,154
. . .:			
Chimpanzee	4	0	4
. . .:			
Orangutan	2	0	2

2.2.2 Mitochondrial DNA

Mitochondrial DNA is only passed down to an individual by the mother and is present in both males and females [16]. Mitochondrial DNA can be used to determine the maternal descent of individuals as it is only passed down by the mother of an individual [11]. An overview of how maternal descent is found with the use of mitochondrial DNA is found in Figure 2.3 below.

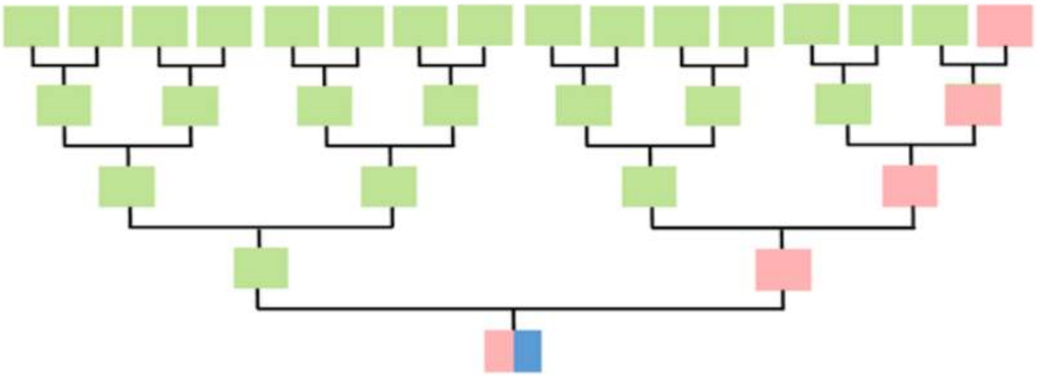


Figure 2.3 Overview of Maternal descent of an individual shown with the colour pink [17, p.1]

Mitochondrial DNA has been used and currently used to determine population genetics studies and human evolutionary studies. One of the main reasons mitochondrial DNA is beneficial is due to the fact that it lacks recombination which is highly beneficial in terms of exploring genealogical relationships with individuals at a regional and continental scale [18]. The lack of recombination in mitochondrial DNA means that the sequence of DNA does not change and remains the same for many generations. Furthermore, mitochondrial DNA mutates much faster when compared with the DNA from the nucleus [18]. A mutation which is highly beneficial in tracing maternal descents is called single nucleotide polymorphism(SNP) which is shown in Figure 2.4 [19]. DNA consists of four chemical bases which store information specific to each individual which are Adenine (A), cytosine(C), thymine(T) and guanine(G) [13].

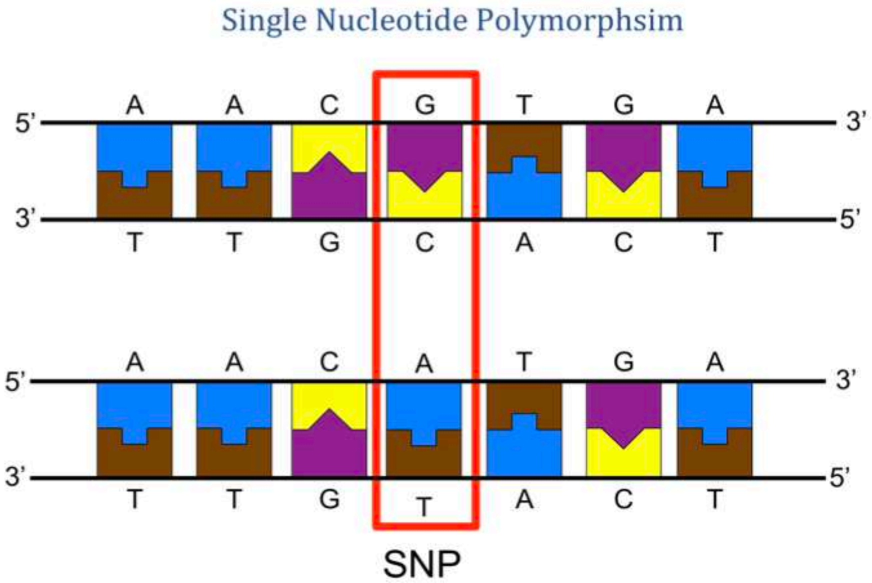


Figure 2.4 SNP mutation in a DNA sequence where in the fourth nucleotide is different in the DNA sequences in the two people [19, p.1]

The single nucleotide polymorphism mutation causes a single base pair in the DNA sequence to be swapped out for a different nucleotide. Hence, in the sense of Figure 2.4, it can be seen that the base pair of guanine and cytosine of person one is different to the base pair of adenine and thymine of person two. Hence, mitochondrial DNA can be used find maternal descendants by comparing patterns of SNP with other individuals. Individuals who have a similar patterned SNP will be considered as to having maternal lineage [19].

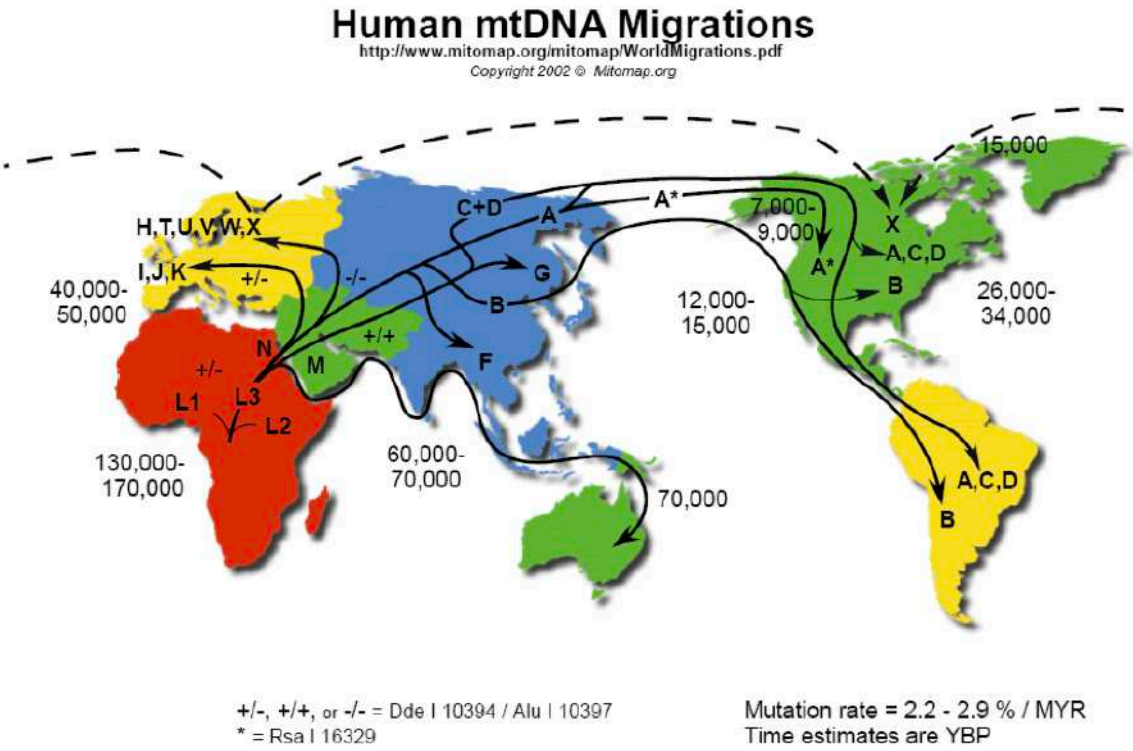


Figure 2.5 The use of mitochondrial DNA to determine haplogroup distribution around the world between 170,000 years ago until 2002 [19, p.2]

Haplogroups are defined as sequence of mitochondrial DNA which have had polymorphism variations which have occurred over than 150,000 years and correlate to the geographic locations of populations through maternal lineage [20]. When considering Figure 2.5, it can be observed that the most population in the map is shown in Africa and two haplogroups consisting of group M and group N branching out to other parts of the world where the haplogroup M branching out to Asia and haplogroup N branching out to Europe. Furthermore, Figure 2.5 shows that some populations turned in the direction of northern Asia which is shown by haplogroup A in the map. Hence, the use of mitochondrial DNA aids in finding the distribution of populations with the use of different haplogroups with their maternal information.

2.2.3 Autosomal DNA

Autosomal DNA consists information on both males and females. This means that with the use of autosomal DNA information on both maternal and paternal descents can be obtained [11]. Autosomal DNA provides information on most of an individual’s DNA [21]. The downside of using autosomal DNA is that it is unable to trace ancestors who are beyond approximately 300 years. Hence, autosomal DNA can be used to find ancestors in recent generations. Furthermore, autosomal DNA can be used to identify an ethnicity of an individual by comparing the results with different people in different ethnicities [21]. Presently, the combination of both mitochondrial DNA with autosomal DNA is being used to determine an individual’s ancestors, the spread of population and their ethnicity [22].

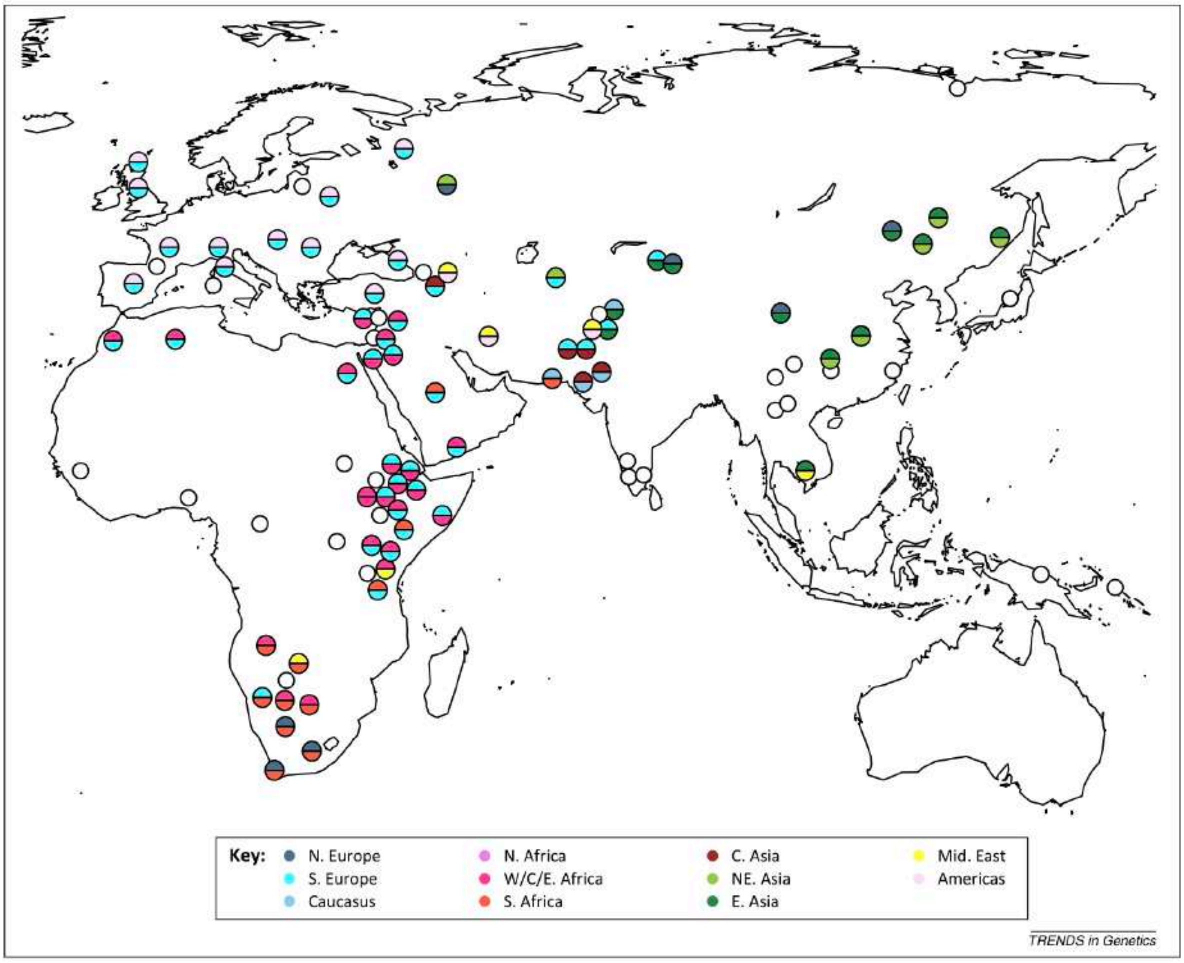


Figure 2.6 Distribution of population where empty circles mean there were no significant evidence on the population[22, p.381]

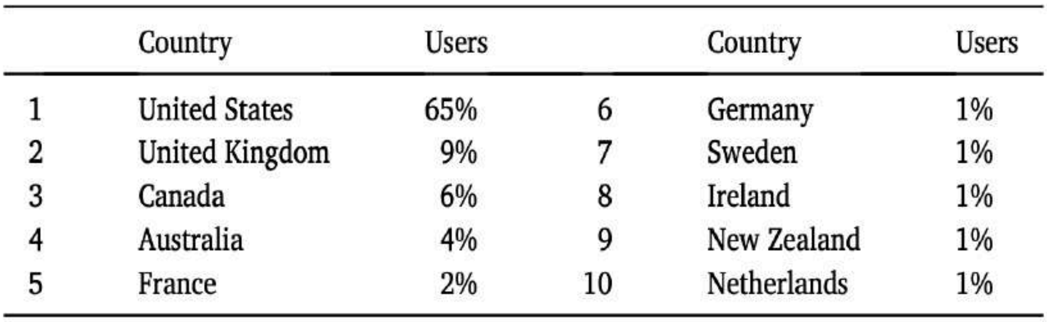
The use of mitochondrial DNA with autosomal DNA has given results on how the population of people have been distributed among different countries. The distribution of people has been shown in Figure 2.6 where it can be seen that different haplogroups have combined together and developed populations [22]. Hence, autosomal DNA can be used to determine and trace the genealogy of an individual. Furthermore, with the combination of autosomal DNA and mitochondrial DNA the distribution of population of an individual's ancestors can be found.

2.3 Current methods of testing genetic genealogy

The demand for genetic genealogy currently is on the rise in the field of forensic science as this is a fairly new technology. The tests which have been used currently mostly consists of comparison of an individual with a standard STR- based DNA profile [23]. Furthermore, high resolution commercial direct-to-consumer tests analysed under high-density microarrays are also being used presently as they give information of distant relatives compared to standard STR tests [1].

The amount of SNP genotype data has been increasing exponentially as the cost of testing has significantly reduced compared to the past. Similarly, the SNP databases available in the present is high and increasing exponentially [24]. This is due to the fact that many companies such as GEDmatch, deCODE genetics, AncestryDNA, MyHeritage and 23andMe have entered the field of genetic genealogy [25].

Table 2.2 The top ten countries with most frequent GEDmatch uploads [1, p.2]



	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%

The high amount of SNP databases have aided in the genetic genealogy technology to move forward. As seen in Table 2.2, users in countries such as United States, United Kingdom, Canada and Australia are using companies such as GEDmatch upload their SNP data which is open for public which aids in comparing related pairs with the use of the large scale SNP data available currently [1].

2.4 Literature Summary

In summary, it can be observed in the following chapter that the types of information found in DNA such as autosomal DNA, mitochondrial DNA and Y chromosomal DNA aid in finding relationships which an individual has with their ancestors, their ethnicity and the movements into different regions of the world in the past generations. These key aspects of DNA aid in determining an individual's background and aid in analysing the information found in the results during the progress of the project.

3. Project Development

The following chapter discusses the project progress and what each member of the group has contributed to the development of the project. Furthermore, the results found have also been discussed in the following chapter.

3.1 Project progress

The progress of the project so far can be separated into four sections, three based around the objectives and the tasks related to them, and one section for wider study of the subject matter and needs of the project.

As this project was started fresh this year with no prior projects to continue from, the first thing the project group did was discuss the explicit requirements with the supervisor and, when available, the collaborator Colleen Fitzpatrick. These discussions along with study of genealogy gave a greater understanding of what is needed from the project. It was from these discussions that the list of objectives was formed and the priority of them was determined.

Initially the project team began work on a program written in JavaScript using the Google Earth Engine API to produce markers on a map. The basic form of this program was completed on April 14th. The program was able to take a list of names, descriptions and coordinates and produce markers on a Google Earth map. The markers were able to be colour coded based off of the description and interacting with the markers would bring up the name of the individual associated with the marker. The information in the list was taken from an array that was imported from a .csv file. The outcome of this project can be seen in Figure 3.1. This program fulfilled most of the requirements of Objective 3 but due to the way basic markers are handled in Google Earth the program would take several seconds per marker to produce the map as each marker is added as its own layer to the map. This time delay was not a deal breaker for Colleen, however whenever the zoom level of the map was adjusted then markers would have to reload and this was too inefficient to be useable. As such, a second program was created, this time using the Google Maps API instead of the Google Earth API. The result from this program can be seen in Figure 3.2. This allowed better interactivity with the markers, as info windows could be created housing additional information about the associated individual and did not face the same loading and reloading issues. Being much faster and with greater customizability than the Google Earth API program, this was decided as the better solution. After a Zoom meeting with Colleen to discuss existing programs and how this was done in the past, the program Earth Point was discussed. This program was capable of everything our program could do, with a cleaner look and as such Colleen has said she is comfortable working with this program in the future and directed us toward focusing our work toward completing the programs to convert from GEDCOM files to .csv as this is the functionality that is most needed.

Prior to this Shaun Fernando had already begun investigating methods for converting GEDCOM files into useable data for the rest of the program. Furthermore, he worked on finding more relevant literature which discussed more on how genetic genealogy worked. Hence, can be used in the theory and methods section of the literature review. After the discussion with Colleen, Harrison Boyce began to assist Shaun in looking into various parsers such as the GEDxlate parser and formed the basic plan for the parser functionality. Furthermore, GEDCOM-parser has also been looked at to see what functionality it possesses compared to GEDxlate to determine which approach is more suitable and efficient to parse

GEDCOM files. In terms of the parser, a few approaches have been taken to start the program. However, they have not been successful in fully implementing the parsing of GEDCOM files. Hence, currently an implementation where the combination of different methods is being tested.

Along with the work done on the parser, Geocoding solutions were investigated to find suitable options for use in the project. Initial plans were to implement free geographic lookup tables to convert location data from address to coordinate but with the discovery of the Google Geocoding API, that became the decided path for the project provided sufficient funding could reasonably be generated by the project to cover the $5 per 1000 calls fees involved with using the API. The idea of a monthly fee for use of the program was floated by Colleen as a fair way to get funding from those in the genealogic community to cover the API fees. An average GEDCOM file contains around 200-300 individual's events and investigations into unidentified persons will generaly involve around 5-10 GEDCOM files. These add up to approximately 1000-3000 API calls per investigation, resulting in an average cost per investigation of 2000 times $0.005, or $10. This can fluctuate highly between investigations as some will have much more information than others. An example provided by Colleen Fitzpatrick included 2498 events across 8 GEDCOM files. With this estimation a pricing structure of $10 USD per month would cover one full scale investigations a month per user. The number of API calls will be further reduced by saving the details of a map after its creation so that the coordinate data can be retained without the need for repeated API calls on the same individuals. In addition a share feature could be introduced to allow collaborating investigators to access the same information without having to run the same API calls. If this pricing structure proves to be non viable, the option exists to pivot to a pay-as-you-use structure wherein a user would pay for the required number of API calls. This method could be more unintuitive and cumbersome to users and as such the monthly price system is preferred.

3.2 Results

The code used to get the markers according to the geographic location on Google Earth editor is shown in Appendix B. Moreover the result of using the following code has been shown in Figure 3.1 which shows the markers for five individuals. A screen capture of the result from the Google Maps program can be found in Figure 3.2. The code of this program can be found in Appendix C.

The discussion with colleen changed the timeline for the deliverables as initially the markers and parser were going to be finished by week eight of semester one which shows that we were ahead of our schedule of finishing getting markers on maps. The new date to finishing the parser has been pushed to second week of semester two which is outlined in the plan shown in Figure 3.3. Currently, as mentioned both Harrison and Shaun are working on the best way to implement a GEDCOM parser which shows that both team members are on track in terms of getting the project objectives done on time.

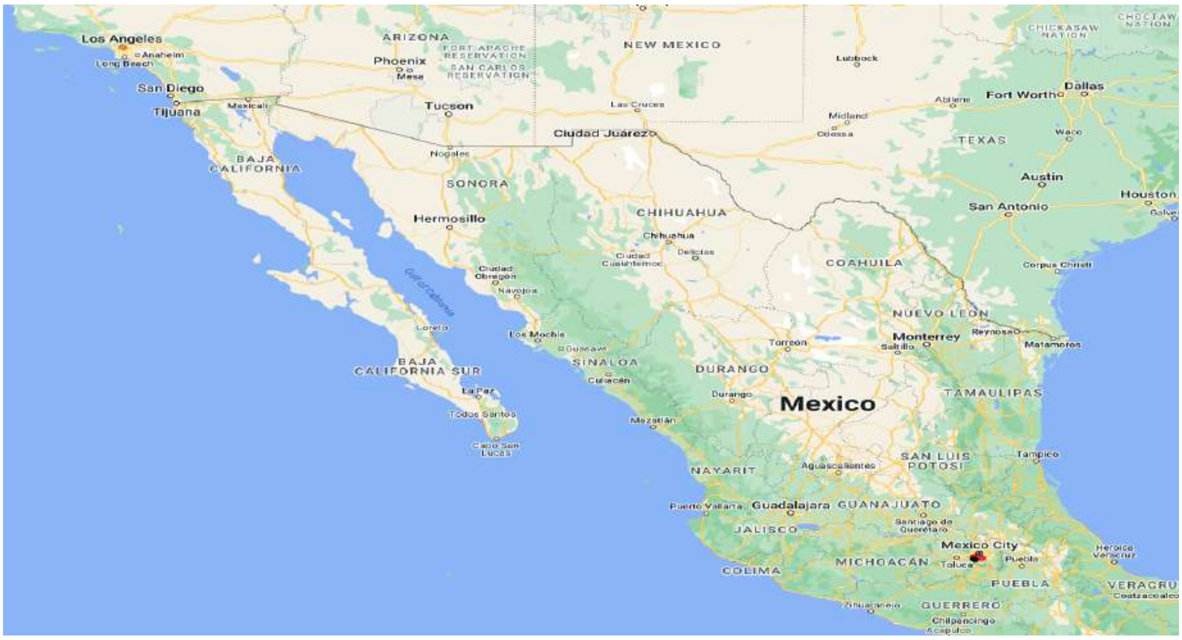


Figure 3.1 Result of markers on Google Earth code showing four events near Mexico City and one in Los Angeles, coloured by date.

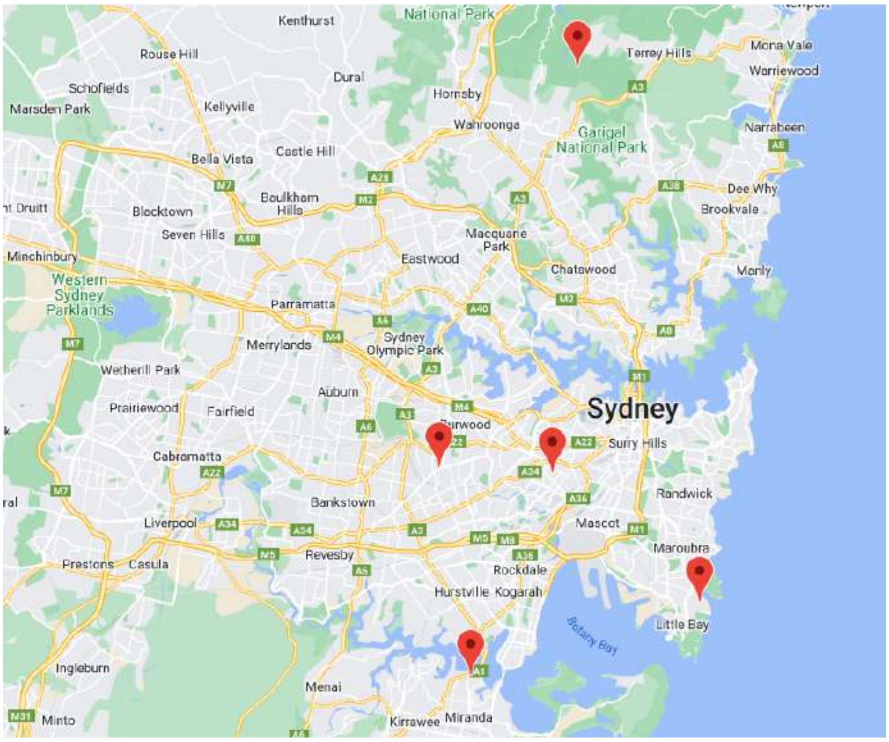


Figure 3.2 Result of markers on google maps showing five events throughout Sydney

3.3 Progress Summary

The project team have so far created two prototypes for fulfilling objectives three and four, both with only simple functionality but approved by the project supervisor as acceptable proof of concepts. Solutions to the remaining two objectives have been investigated and the desired system designs planned. These include a GEDCOM parser for creating .csv files with the relevant information written in Python and a system utilising the Google Geocoding API to obtain co-ordinate information of locations as needed for mapping in Google Earth. The following chapter goes further in depth into the plans for completing these systems and the project as a whole.

4. Completion plan

The plan for the future of the project can be broken down into four main sections. First, the plans for the completion of each of the remaining objectives (1.4.1 through 1.4.3) and finally the plan for the combination and testing of the program as a whole and combined system, from GEDCOM files to interactive map. This plan only covers the work intended to be completed by the 2022 project team in detail and will not fully cover potential additions in future years as the project is iterated on and improved. These will be covered in detail in the final report at the end of the project. Furthermore, a timeline of how the project will be completed is given in the Gantt chart in Appendix A.

4.1 Objective 1

The first system to be developed will be the parser from GEDCOM to .csv. This component has already been heavily investigated and existing options assessed. The team came to the decision to create a custom GEDCOM parser as the existing options are either too isolated to be used in a combined program, or unsuited to the projects needs. With a solid fundamental understanding of how the existing parsers work the team will first create the input system for taking in the GEDCOM file and reading the data into strings in Python. Once that is functional the separation of information into arrays will be done via checking the types of tag (FAM vs INDI, etc) and location of whitespace in the GEDCOM file. These separated values can then be written into a .csv file using a writer object in Python. This will be done using the writerow function, using the array from each individual event (birth, death etc.) for each row. A basic form of this subsystem of the final program is intended to be completed by the end of week two of semester two so testing can begin. A final, working version is intended to be completed by end of week four of semester two. This subsystem will be worked on by both team members with Shaun taking lead. Following the completion of this subsystem, the geocoding subsystem will be created.

4.2 Objective 2

Of the three remaining subsystems to be produced, the geocoding subsystem requires the least work. This is due to the planned solution being the use of the Google Geocoding API. With a robust system already completed and easily integrated into the other systems in the program, getting basic functionality of the geocoding subsystem will be quick to complete. The main work to be done is the implementation of a system to detect and avoid potential errors of the API that can occur through the API assuming a location from minimal information. For example, given the input “Perth, United Kingdom” the geocoding API would successfully give the co-ordinates for the city of Perth in Scotland. However if just given “Perth” the API would instead give the co-ordinates of Perth, Western Australia as this is the larger city. As such a system will be created to give the user the option of selecting potential options or simply removing the event from the list if not enough information is detected. In general, two levels of location information (city and state, city and country, etc.) will be considered acceptable while single level information (city only) will be tested. The co-ordinate output will be measured against recent outputs and if sufficient disparity is found the user will be prompted to discard or accept. This system is planned to be completed by the end of week five of semester two. and testing to be completed by end of week seven of semester two. The final subsystem to be produced will be the most challenging and as such will be worked on by Harrison while Shaun completes the final work on the geocoding subsystem.

4.3 Objectives 3 and 4

This subsystem will be the solution to objective 1.4.3, the .csv to .kml converter. This subsystem is similar to that of the GEDCOM to .csv parser but where that subsystem only requires a conversion of file types, this subsystem must create a map from scratch using information given. This means much greater customisation and adjusting of data presentation is required. To begin, the simplest form of markers will be produced as shown in the example code in the appendices. These markers will initially be solely based off the co-ordinate data from the .csv files. Once simple markers are produced, additional information will be added and the markers made interactive to show the additional information. This section will require the most work and currently has the least research done into it as the project collaborator, Colleen Fitzpatrick has deemed it the least important due to the existence of Earth Point which does a sufficiently good job as a .csv to .kml converter. The planned completion of this subsystem is end of week eight of semester two but due to the lack of need for this subsystem it may end up being that a simple program is created, capable of displaying much of the required information and fulfilling objectives 1.4.3 and 1.4.4 but not sufficient to outdo the existing option of Earth Point. As such it may fall to future groups to iterate on this component to create a truely superior option with full integration into the other subsystems for smooth operation by the end user.

4.4 Combined System

Following the completion of the second subsystem, the two completed subsystems will be integrated into each other. The result of this will be a program capable of taking a GEDCOM file and creating a .csv file with co-ordinate data instead of location names. This should be simple to do as each subsystem program has a defined input and output and the output of the GEDCOM parser is exactly what is needed for the input of the geocoding system. This is intentional design and should save time for the project team to allow further work to be done on the third subsystem, despite it being deemed as of lesser importance. The integration of the first two subsystems is planned to be completed by the end of week five of semester two and testing finished by the end of week seven of semester two. Should the third subsystem be completed, its integration will be as simple as that of the first two, again, due to the choice of input and outputs for the subsystems matching exactly. This integration is planned for the end of the project, around 23rd of September 2022.

4.5 Plan Summary

In summary, thorough plans have been made for each of the core subsystems of the program. A GEDCOM parser being written in Python, utilising the writer object, is the plan for subsystem one. Subsystem two will utilise the Google Geocoding API and sanity checks via the user for potentially inaccurate location data to fulfill objective two and subsystem three having been deemed of the lowest priority has been positioned as the last to be completed. Further investigations into upgrades of the basic prototypes created will be completed in the week s to come. Plans have been made to integrate all three subsystems into a single program for ease of use by the user. With these plans having been made and a timeline set for the completion of the project, the project shall be completed within all time constraints and within budget.

5. Summary conclusion

In summary, the project aims to produce a single program capable of taking several GEDCOM files and producing an interactive map displaying the location of the events described in the GEDCOM files. This will be done through the creation of three subsystems to complete the core objectives of the project. Firstly a system to produce a .csv file from the event information in the GEDCOM files. Secondly, a system to convert from location names to coordinate data, making sure to sanity check locations via the user where necessary. And finally, a system to generate an interactive map via Google Earth, placing interactive markers at the locations specified by the co-ordinates in the .csv file. These systems will be integrated together into a single program to make the user experience much better than the currently existing methods which require separate programs and conversions for each step. The final program will be useable by all people interested in genealogical invesitgations but is primarily aimed at helping with investigations into unidentified persons by Forensic Scientists such as the collaborator for the project, Colleen Fitzpatrick. The project team have created a simple system for creating the final map when given coordinates and related information and have created thorough plans for the completion of the other two subsystems based off extensive research into existing technologies.

References

[1] D. Kling, C. Phillips, D. Kennett and A. Tillmar, "Investigative genetic genealogy: Current methods, knowledge and practice", *Science Direct*, p. 1, 2021. Available: http://Investigative genetic genealogy: Current methods, knowledge and practice.

[2]Blau, S, Graham, J, Smythe, L & Rowbotham, S 2020, 'Human identification : a review of methods employed within an Australian coronial death investigation system', *International Journal of Legal Medicine,* vol. 135, no. 1, pp. 375-385.

[3]Hassan, O, Abu, N, Abdin, Z, 2019,' HUMAN IDENTIFICATION SYSTEM:REVIEW', I*nternational Journal of Computing And Business Research,* vol.9 , pp. 1-7.

[4]Ward, J 2019, *How do we identify human remains?*, The Conversation.

[5]Harding, T, Milot, E, Moreau, C, Lefebvre, J, Bournival, J, Vézina, H, … Labuda, D 2020, 'Historical human remains identification through maternal and paternal genetic signatures in a founder population with extensive genealogical record', *American Journal of Physical Anthropology*, vol. 171, no. 4, pp. 645–658.

[6]E. Greytak, C. Moore and S. Armentrout, "Genetic genealogy for cold case and active investigations", *Forensic Science International*, vol. 299, pp. 103-113, 2019. Available: 10.1016/j.forsciint.2019.03.039.

[7]C. Alho, M. Dorn and A. Eduardo, "Would GENEALOMICS be an appropriate term to designate family tree research based on genome-wide data?", *Journal of Genetic Genalogy*, vol. 9, pp. 3-4, 2021. Available: https://jogg.info/wp-content/uploads/2021/12/91-Issue.pdf.

[8] Jones, Tamura. 2010. 'A Gentle Introduction to GEDCOM'.Modern Software Experience.August 24. http://www.tamurajones.net/AGentleIntroductionToGEDCOM.xhtml.

[9] Roued-Cunliffe, Henriette. 2017. "Visualising Historical Networks: Family Trees and Wikipedia". *Academic Quarter | Akademisk Kvarter*, nr. 15 (oktober):40-53. https://doi.org/10.5278/ojs.ak.v0i15.2684.

[10]"Build software better, together", *GitHub*, 2022. [Online]. Available: https://github.com/search?q=gedcom+parser&type=Repositories.

[11] E. Greytak, C. Moore and S. Armentrout, "Genetic genealogy for cold case and active investigations", *Forensic Science International*, vol. 299, pp. 103-113, 2019. Available: 10.1016/j.forsciint.2019.03.039.

[12] Hammer MF, ElisNA (1995) Appendix 1.Y Chromosom ConsortiumNewslett 2:8-9

[13]"What is DNA?: MedlinePlus Genetics", *Medlineplus.gov*, 2021. [Online]. Available: https://medlineplus.gov/genetics/understanding/basics/dna/.

[14]"Allele", *Genome.gov*, 2022. [Online]. Available: https://www.genome.gov/genetics-glossary/Allele.

References

[15] T. Zerjal et al., "Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis.", *PubMed Central (PMC)*, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1712423/.

[16] S. Luo et al., "Biparental Inheritance of Mitochondrial DNA in Humans", *Proceedings of the National Academy of Sciences*, vol. 115, no. 51, pp. 13039-13044, 2018. Available: 10.1073/pnas.1810946115.

[17]"AncestrySupport", *Support.ancestry.com*, 2022. [Online]. Available: https://support.ancestry.com/s/article/Y-DNA-mtDNA-and-Autosomal-DNA-Tests?language=en_US.

[18] C. Moraes, "Sorting mtDNA Species—the Role of nDNA-mtDNA Co-evolution", *Science Direct*, vol. 30, no. 6, p. 1002, 2019. Available: https://reader.elsevier.com/reader/sd/pii/S1550413119306126?token=2636893951A708F0C0DEA248B5B92DE0A8907F58B0A9D9EAAB73E87D1DD4698E39023715A62B1883ED0511B6F8630084&originRegion=us-east-1&originCreation=20220529074357.

[19] *Ase.tufts.edu*. [Online]. Available: https://ase.tufts.edu/chemistry/hhmi/documents/Protocols/Maternal%20Ancestry_Introduction_Reworked_Aug_25_2011.pdf

[20] C. Kenney, D. Ferrington and N. Udar, *Retina- Chapter 32 - Mitochondrial Genetics of Retinal Disease*. 2013, p. 635.

[21]"Understanding genetic ancestry testing", *UCL Division of Biosciences*, 2022. [Online]. Available: https://www.ucl.ac.uk/biosciences/departments/genetics-evolution-and-environment/research/molecular-and-cultural-evolution-lab/debunking-genetic-astrology/understanding-genetic-ancestry-testing.

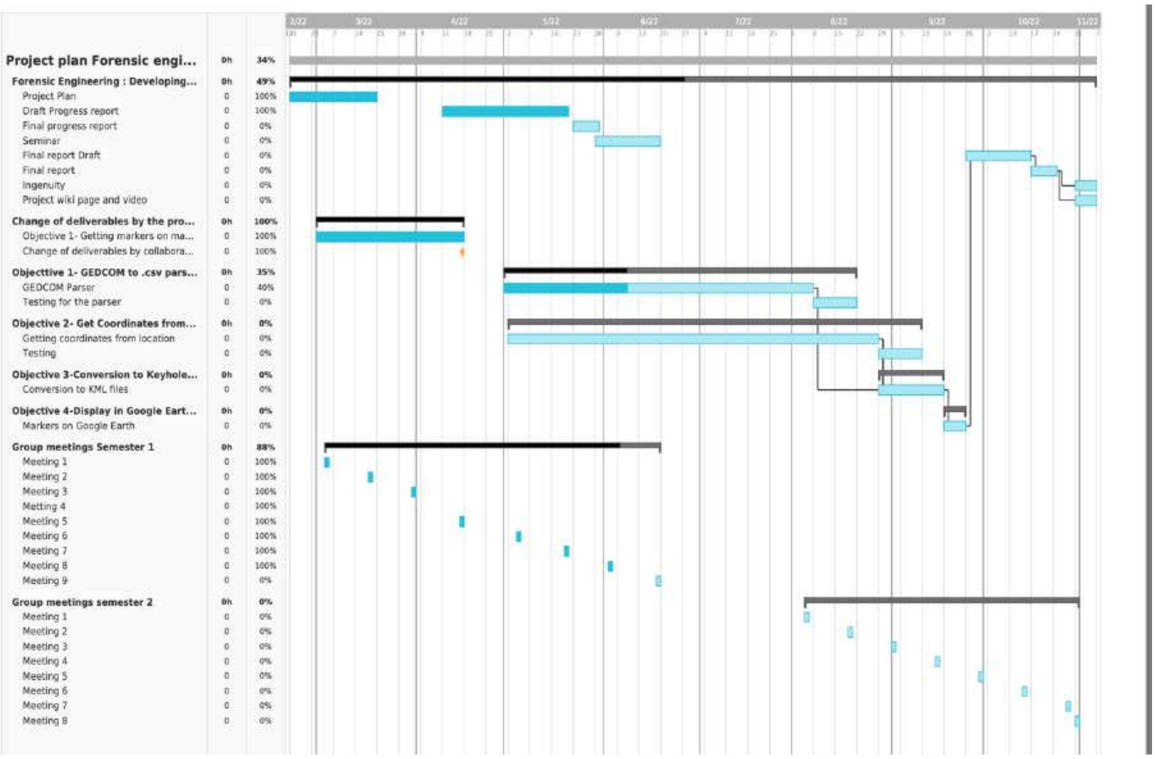
[22] J. Pickrell and D. Reich, "Toward a new history and geography of human genes informed by ancient DNA", *Science Direct*, vol. 30, no. 9, p. 381, 2022. Available: https://www.sciencedirect.com/science/article/pii/S0168952514001206.

[23] N. Wyner, M. Barash and D. McNevin, "Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype", *Frontiers in Genetics*, vol. 11, p. 1, 2020. Available: 10.3389/fgene.2020.00884.

[24] *Should we be making use of genetic genealogy to assist in solving crime?*. The Biometrics and Forensics Ethics Group, 2020, p. 10.

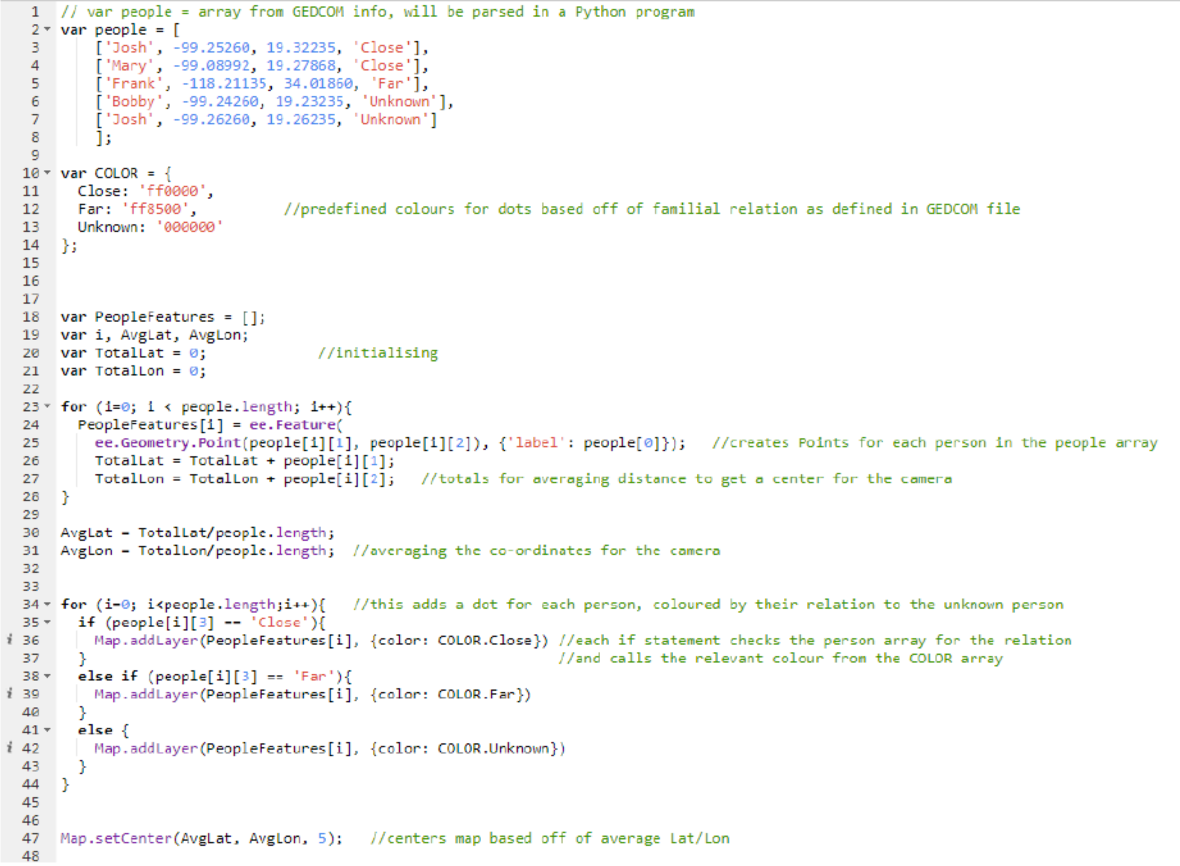
[25] "Autosomal DNA testing comparison chart", *International Society of Genetic Genealogy Wiki*, 2022. [Online]. Available: https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart.

Appendix A. Project Plan (Gantt Chart)



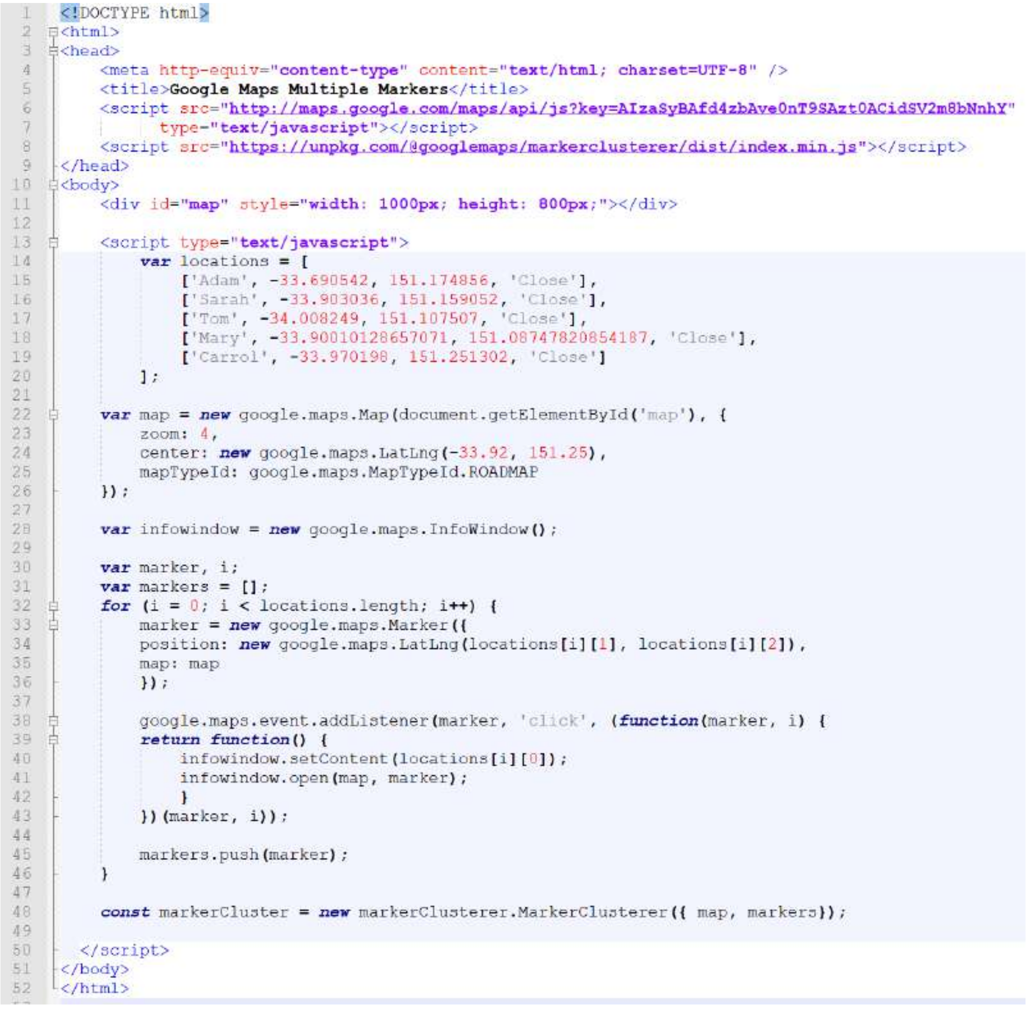
Project plan Forensic engi...	**0h**	**34%**
Forensic Engineering : Developing...	**0h**	**49%**
Project Plan	0	100%
Draft Progress report	0	100%
Final progress report	0	0%
Seminar	0	0%
Final report Draft	0	0%
Final report	0	0%
Ingenuity	0	0%
Project wiki page and video	0	0%
Change of deliverables by the pro...	**0h**	**100%**
Objective 1- Getting markers on ma...	0	100%
Change of deliverables by collabora...	0	100%
Objecttive 1- GEDCOM to .csv pars...	**0h**	**35%**
GEDCOM Parser	0	40%
Testing for the parser	0	0%
Objective 2- Get Coordinates from...	**0h**	**0%**
Getting coordinates from location	0	0%
Testing	0	0%
Objective 3-Conversion to Keyhole...	**0h**	**0%**
Conversion to KML files	0	0%
Objective 4-Display in Google Eart...	**0h**	**0%**
Markers on Google Earth	0	0%
Group meetings Semester 1	**0h**	**88%**
Meeting 1	0	100%
Meeting 2	0	100%
Meeting 3	0	100%
Metting 4	0	100%
Meeting 5	0	100%
Meeting 6	0	100%
Meeting 7	0	100%
Meeting 8	0	100%
Meeting 9	0	0%
Group meetings semester 2	**0h**	**0%**
Meeting 1	0	0%
Meeting 2	0	0%
Meeting 3	0	0%
Meeting 4	0	0%
Meeting 5	0	0%
Meeting 6	0	0%
Meeting 7	0	0%
Meeting 8	0	0%

Appendix B. code used to get the markers according to the geographic location on Google Earth editor



```
// var people = array from GEDCOM info, will be parsed in a Python program
var people = [
    ['Josh', -99.25260, 19.32235, 'Close'],
    ['Mary', -99.08992, 19.27868, 'Close'],
    ['Frank', -118.21135, 34.01860, 'Far'],
    ['Bobby', -99.24260, 19.23235, 'Unknown'],
    ['Josh', -99.26260, 19.26235, 'Unknown']
    ];

var COLOR = {
  Close: 'ff0000',
  Far: 'ff8500',          //predefined colours for dots based off of familial relation as defined in GEDCOM file
  Unknown: '000000'
};



var PeopleFeatures = [];
var i, AvgLat, AvgLon;
var TotalLat = 0;             //initialising
var TotalLon = 0;

for (i=0; i < people.length; i++){
  PeopleFeatures[i] = ee.Feature(
    ee.Geometry.Point(people[i][1], people[i][2]), {'label': people[0]});   //creates Points for each person in the people array
    TotalLat = TotalLat + people[i][1];
    TotalLon = TotalLon + people[i][2];   //totals for averaging distance to get a center for the camera
}

AvgLat = TotalLat/people.length;
AvgLon = TotalLon/people.length;  //averaging the co-ordinates for the camera


for (i=0; i<people.length;i++){   //this adds a dot for each person, coloured by their relation to the unknown person
  if (people[i][3] == 'Close'){
    Map.addLayer(PeopleFeatures[i], {color: COLOR.Close}) //each if statement checks the person array for the relation
  }                                                       //and calls the relevant colour from the COLOR array
  else if (people[i][3] == 'Far'){
    Map.addLayer(PeopleFeatures[i], {color: COLOR.Far})
  }
  else {
    Map.addLayer(PeopleFeatures[i], {color: COLOR.Unknown})
  }
}


Map.setCenter(AvgLat, AvgLon, 5);   //centers map based off of average Lat/Lon

```

Appendix C. Google Maps program code



```
<!DOCTYPE html>
<html>
<head>
    <meta http-equiv="content-type" content="text/html; charset=UTF-8" />
    <title>Google Maps Multiple Markers</title>
    <script src="http://maps.google.com/maps/api/js?key=AIzaSyBAfd4zbAve0nT9SAzt0ACidSV2m8bNnhY"
          type="text/javascript"></script>
    <script src="https://unpkg.com/@googlemaps/markerclusterer/dist/index.min.js"></script>
</head>
<body>
    <div id="map" style="width: 1000px; height: 800px;"></div>

    <script type="text/javascript">
        var locations = [
            ['Adam', -33.690542, 151.174856, 'Close'],
            ['Sarah', -33.903036, 151.159052, 'Close'],
            ['Tom', -34.008249, 151.107507, 'Close'],
            ['Mary', -33.90010128657071, 151.08747820854187, 'Close'],
            ['Carrol', -33.970198, 151.251302, 'Close']
        ];

    var map = new google.maps.Map(document.getElementById('map'), {
        zoom: 4,
        center: new google.maps.LatLng(-33.92, 151.25),
        mapTypeId: google.maps.MapTypeId.ROADMAP
    });

    var infowindow = new google.maps.InfoWindow();

    var marker, i;
    var markers = [];
    for (i = 0; i < locations.length; i++) {
        marker = new google.maps.Marker({
        position: new google.maps.LatLng(locations[i][1], locations[i][2]),
        map: map
        });

        google.maps.event.addListener(marker, 'click', (function(marker, i) {
        return function() {
            infowindow.setContent(locations[i][0]);
            infowindow.open(map, marker);
            }
        })(marker, i));

        markers.push(marker);
    }

    const markerCluster = new markerClusterer.MarkerClusterer({ map, markers});

  </script>
</body>
</html>
```