SCHOOL OF
ELECTRICAL AND
ELECTRONIC ENGINEERING

THE UNIVERSITY
of ADELAIDE

Who Killed the Somerton Man

Zihe Peter Wang a1684198

ELEC ENG 4068 HONOURS PROJECT

B.E. in Computer Systems Engineering

Date submitted: June 7th 2019

Supervisor: Derek Abbott

Acknowledgments

Appreciate the help from the supervisor Dr. Derek Abbott who provide important suggestions and guidance to the project.

Executive Summary

Somerton Man case is most mysterious case in last century. A unknown man was murdered on Somerton Beach, and identifications of the killer and the victim are still mysteries nowaday. The project aims to investigate the identification of the Somerton man with his DNA data provided. Unfortunately, the DNA data is incomplete and has a high drop rate, therefore the team of the project would be required to use different strategies and techniques to recover the DNA and find out any possible characteristics of the Somerton man.

To approach the goals of the project, the team would have firstly recover the DNA data and conduct DNA analysis via online services. With different recovery algorithm, the team would have multiple sets of recovered DNA and compare the differences between them. Another key task is the degradation process of complete DNA data. By degrading a complete DNA file, the difference between a complete DNA and incomplete DNA can be observed.

Since the project is still in progress, there is no typical characteristic of the Somerton man being identified.

1 Introduction

1.1 Motivation

The main topic of the project is human identification via using software programming and genetic analysis techniques. The project conducts a study on investigating the identification of the victim in the Somerton man case which is one of the most mysterious cases in last century. On December 1st 1948, a well-dressed male was found dead on Somerton Beach in Adelaide [1]. He was clean-shaven, well dressed in a suit and no belongings could prove his identity [7]. Later the man was called as Somerton man. The figure below shows the look of the Somerton man.

Figure 1.1: The Somerton Man

After more than half century, the identification of the Somerton man is still unsolved. With the supply of Somerton man's DNA data extracted from his hair. Unfortunately, the DNA data is incomplete, but the team of the project would try the best to find who the Somerton man is with modern techniques.

As mentioned previously, human identification is the main topic of the project. In modern society, human identification techniques is useful in multiple aspects, such as criminal investigation or seeking relatives. Most current identification techniques would require high quality DNA samples, but the project focus on investigating identification techniques based on low quality genetic data. Also the project concentrate on using electrical engineering methods to improve the identification techniques.

1.2 Objectives

The aim of the project is to investigate the identification of the Somerton man. To be more specific, the group is aiming to identify any possible physical characteristics, genetic diseases or ethnicity of the Somerton man. To achieve these goals, the team would use software and genetic analysis techniques to work on the Somerton man's DNA data (eg. Recovering the Somerton man's DNA data).

2 Background

2.1 DNA

DNA is the hereditary material which stores the genetic information in humans [2]. There are two types of DNA in human beings, one is known as nuclear DNA which is located in cell nucleus and another type is mitochondrial DNA which is located in the mitochondria. This project only focuses on the analysis of nuclear DNA. DNA stores genetic information as a sequence built up with four types of nitrogen bases which are adenine (A), guanine (G), cytosine (C), and thymine (T) [2]. Also, a sugar molecule and a phosphate molecule are attached to each nitrogen base to form a molecule called nucleotide. The bases would pair up (A with T and C with G) and multiple nucleotides are placed in two strands to form a double helix which looks like a spiral [2]. In general, a DNA is a genetic sequence formed by multiple base pairs. The genetic instructions of building and maintaining an organism are obtained from the order of these base pairs [2]. There are about 3 billion bases in human DNA, in which more than 99% of the bases are common in all human beings, and the physiological differences among people depends on these 1% DNA.

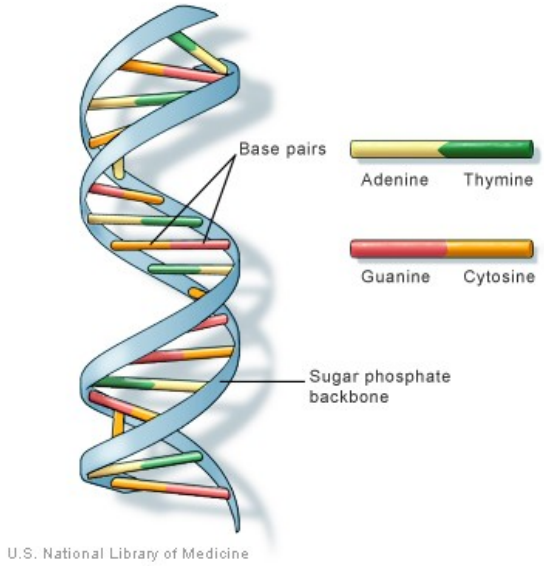
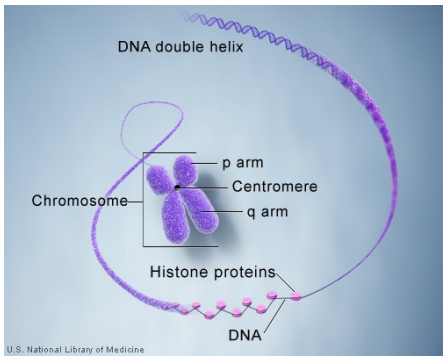
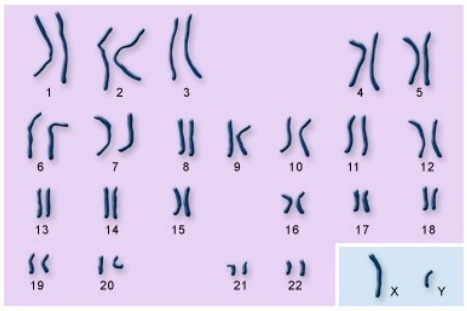


Figure 2.1: DNA structure

2.2 Chromosome

Chromosome is an integrated package of DNA molecules. It has thread-like structure, and DNA molecules are coiled up around hi stones proteins to form the structure [3]. There are 23 pairs of chromosomes in human body's cell, which is 46 chromosomes in total. 22 pairs are called autosomes which are common for both males and females and the last 23rd pair is sex chromosomes which differ males and females. In this project, the DNA data analysis would only focus on autosomes [4].





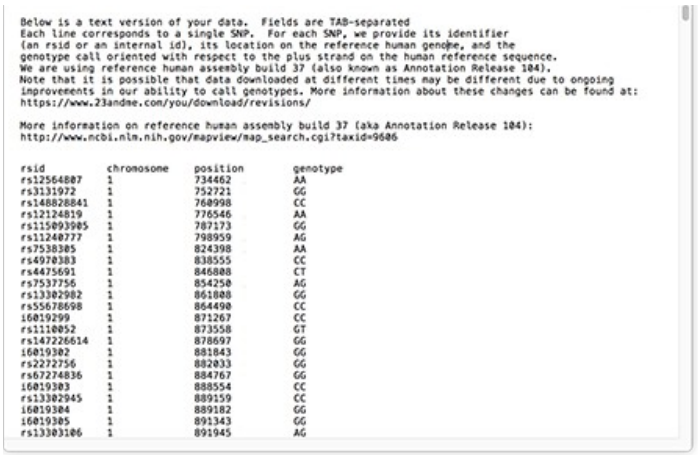
U.S. National Library of Medicine

2.3 SNP

Single nucleotide polymorphisms(SNPs) are most common type of genetic variation among human beings [5]. Each SNP represents a difference in a nucleotide which is a single DNA molecule [6]. For instance, a SNP may replace a nucleotide of base guanine (G) with cytosine (C). These SNPs can be found nearly once in every 1,000 nuceotides on average in a person's DNA. Most SNPs do not effect health of owner. However, some of these variations may associated with diseases.

2.4 DNA reference file

A DNA reference file stores a group of SNPs data of owner's DNA. The format of DNA reference files using in this project is the same format of 23andMe company's file. A screen shot of a sample file is shown below.



```
Below is a text version of your data.  Fields are TAB-separated
Each line corresponds to a single SNP.  For each SNP, we provide its identifier
(an rsid or an internal id), its location on the reference human genome, and the
genotype call oriented with respect to the plus strand on the human reference sequence.
We are using reference human assembly build 37 (also known as Annotation Release 104).
Note that it is possible that data downloaded at different times may be different due to ongoing
improvements in our ability to call genotypes. More information about these changes can be found at:
https://www.23andme.com/you/download/revisions/

More information on reference human assembly build 37 (aka Annotation Release 104):
http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606

rsid            chromosome      position        genotype
rs12564807      1               734462          AA
rs3131972       1               752721          GG
rs148828841     1               760998          CC
rs12124819      1               776546          AA
rs115093905     1               787173          GG
rs11240777      1               798959          AG
rs7538305       1               824398          AA
rs4970383       1               838555          CC
rs4475691       1               846808          CT
rs7537756       1               854250          AG
rs13302982      1               861808          GG
rs556678698     1               864490          CC
i6019299        1               871267          CC
rs1110052       1               873558          GT
rs147226614     1               878697          GG
i6019302        1               881843          GG
rs2272756       1               882033          GG
rs67274836      1               884767          GG
i6019303        1               888554          CC
rs13302945      1               889159          CC
i6019304        1               889182          GG
i6019305        1               891343          GG
rs13303106      1               891945          AG
```

As shown in the figure, there are 4 columns rsid, chromosone, position and genotype in the DNA reference file. The rsid is a unique id used to identify a specific SNP [9]. The format of rsid starts with "rs" and followed by a number (eg. rs123456). These rsids are commonly used by researchers and databases. There is another special format of rsid that starts with "i" and followed by a number (eg. i123456). This "i" format is used internally by 23andMe to identify the unknown SNP and can not be used in public database. The second column chromosone identify which chromosome the SNP belongs to. Then the third column position indicates positions of SNPs in owner's DNA sequence. Last column genotype represent the base pairs of variants(A, T, G, or C). Note that there are some cases, the genotype result for some SNPs are not able be provided and "--" would be displays in genotype column [9].

3 Task 1: Testing with Somerton Man's DNA reference file

3.1 Aims

The aim of this task is to have a basic understanding of the DNA reference file and DNA analysis techniques. The project provide a DNA reference file of the Somerton

rs12132100	1	1023788	--
rs116334314	1	1026428	--
rs115662838	1	1026913	--
rs77334480	1	1027888	--
rs12731175	1	1030374	--
rs9651273 1	1031540	--	
rs6671356 1	1040026	--	
rs147606383	1	1045331	--
rs12080505	1	1045606	--
rs61766344	1	1054091	--
rs9442373 1	1062638	--	
rs4072537 1	1065296	CC	
rs11260598	1	1065726	--
rs61766346	1	1068883	--
rs139475585	1	1070467	--
rs141230226	1	1072181	--
rs11260603	1	1079198	CC
rs116661896	1	1079261	--
rs74045142	1	1092071	--
rs77791262	1	1092205	--
rs57527288	1	1092563	--
rs61768477	1	1095130	--
rs9442385 1	1097335	--	

Figure 3.1: Screenshot of Somerton Man's DNA reference file

man which is not a complete data. A screen shot of the file is shown below.

The first goal of this task is to evaluate the quality of the file including counting the total amount of SNPs and the amount of available SNPs. After this, the team should try to conduct some DNA analysis on the DNA reference file.

3.2 Methods

To approach the first goal of this task, the team has developed a program which provide functions for counting total amount of SNPs, amount of available SNPs (SNPs that do not have genotype of "--") and determine the percentage of available SNPs for each 22 chromosones of input DNA raw data. Program was developed with C++ language.

Then a website called GEDmatch has been used for conducting DNA analysis. GEDmatch is a website that has an open data personal genomics database and provide tools for DNA and genealogy research. The site become well known after law enforcement in California use it to the Golden State Killer case and are commonly used by all law enforcement in United State [10]. Somerton Man's DNA reference file was uploaded to the website and tried to conduct further DNA analysis.

3.3 Results

The counting outputs of Somerton man's DNA data is presented in figure 3.2. As the figure shown, there are more than 0.6 million SNPs in the files, but only about 2% of them have determined base pairs. In DNA analysis,

Counting results of SNPs			
Chromosome	Total amount	Exist amount	Percentage
1	49510	1014	2.05%
2	51771	978	1.89%
3	43023	658	1.53%
4	39473	621	1.57%
5	37028	661	1.79%
6	44021	880	2.00%
7	34356	655	1.91%
8	31681	601	1.90%
9	26445	519	1.96%
10	30522	705	2.31%
11	30943	705	2.28%
12	29432	596	2.03%
13	22080	393	1.78%
14	19961	441	2.21%
15	19006	440	2.32%
16	20396	558	2.74%
17	19401	519	2.68%
18	17674	372	2.10%
19	14879	514	3.45%
20	14781	375	2.54%
21	8607	245	2.85%
22	8915	303	3.40%
Total	613905	12753	2.08%

Figure 3.2: SNPs counting results of Somerton man DNA file

only the SNPs with available base pairs can be used and large amount available SNPs would be required.

Then the team upload Somerton man's DNA reference file to GEDmatch, but the website reject to process the data due to the file did not meet the minimum requirements of 2000 SNPs for each chromosome. The team was failed to conduct DNA analysis on Somerton man's DNA data.

In order to satisfy the minimum requirements of GEDmatch, a data recovery work would be required.

4 Task 2: Artificially complete DNA file

4.1 Aims

First aim is to recover Somerton man's DNA file to have more than 2000 available SNPs for each chromosome. Different recovery algorithms can be implemented to produced multiple synthetic DNA reference files. Then the second aim is using tools provided on GEDmatch to conduct DNA analysis. Analysis could including searching relatives or checking ethnicity. Also, the team should compare the analysis outcomes of different artificial DNA files.

4.2 Methods

The recovery works would be done by developing multiple programs with C++. In general, the recovery work is to replace fixed amount of empty SNPs with available SNPs. Several algorithms implemented would be introduced. First algorithm called random algorithm is to replace empty SNPs with random base pairs in genotype. Replacing empty genotype with homozygous pairs (AA, GG, TT, CC) can be considered as another algorithm called identical algorithm.

After creating several analyzable DNA files, upload them to GEDmatch and using one-to-many tool to check is there any relative person can be found in the database. In addition, ethnicity check tool is used to obtain the ethnicity proportion of each artificial DNA kit.

4.3 Results

With the developed program, multiple artificial DNA kits were created. Unfortunately, all of these DNA kits have 0 matches with other DNA in the public database which means these artificial DNA have no relative can be found in the GEDmatch database.

Kit: [redacted] [23andMe]

Kit	1:1	Name	Email	Largest Seg	Total cM	Gen	Overlap	Date Compared	Testing Company

0 is number of matches reported

Figure 4.1: match results of artificial DNA(replace all empty SNPs with random pairs)

Figures 4.2 to 4.5 are ethnicity proportions of artificial DNA kits implementing random algorithm with different amounts of SNPs. Figures 4.6 presents the ethnicity proportions of artificial DNA file implementing identical algorithm of AA with 3500 SNPs for each chromosome.

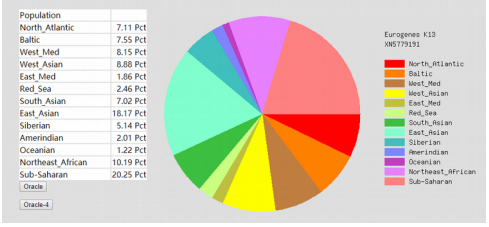


Figure 4.2: Ethnicity proportions (replace all empty SNPs with random pairs)

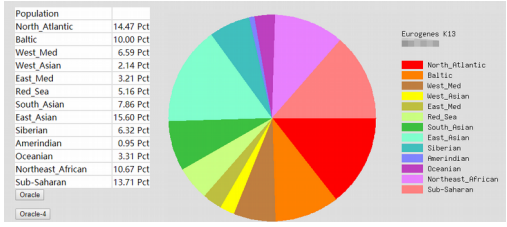


Figure 4.3: Ethnicity proportions (replace empty SNPs with random pairs, 2500 available SNPs each chromosone)

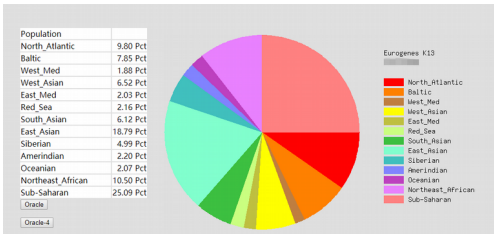


Figure 4.4: Ethnicity proportions (replace empty SNPs with random pairs, 5000 available SNPs each chromosone)

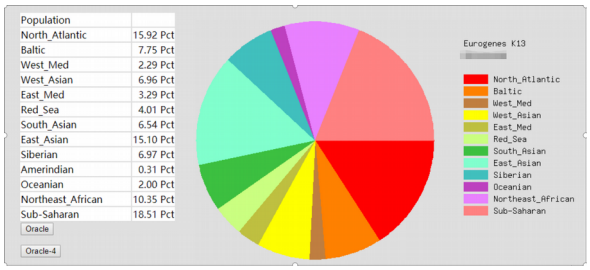


Figure 4.5: Ethnicity proportions (replace empty SNPs with random pairs, 3500 available SNPs each chromosone)

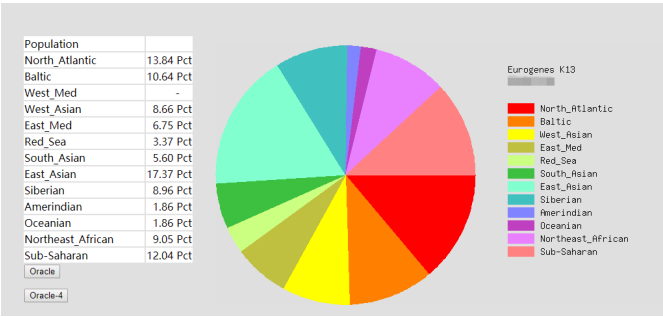


Figure 4.6: Ethnicity proportions (replace empty SNPs with AA pairs, 3500 available SNPs each chromosone)

5 Task 3: Research on SNP

5.1 Aims

During this task, the team would focus on searching clinical effects of each available SNP and identify any possible genetic disease or physical characteristics that Somerton man has.

5.2 Methods

To search the clinical effects of SNPs, the team would develop a data mining program to collecting information in SNP database. The SNP database currently used is dbSNP which is the largest database for nucleotide variations in the world, and is managed by the National Center for Biotechnology Information (NCBI) [11]. Figure 5.1 shows the information provided by dbSNP, and the team would collect the clinical significance refers to each rsid.

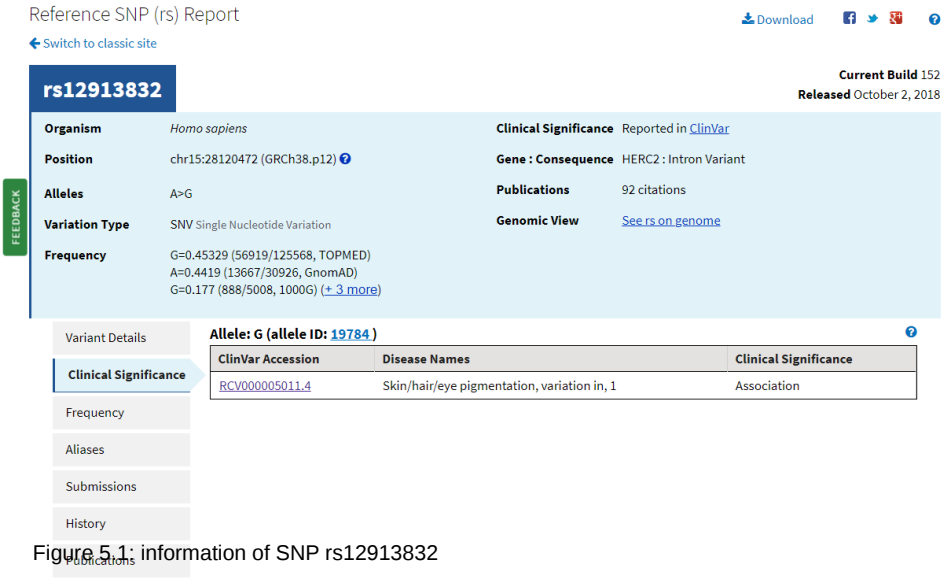


Figure 5.1: information of SNP rs12913832

5.3 Result

This task is currently in progress. So far the team has finished the demo of data mining program. A sample output is presented in figure 5.2. As the figures shown, the program do record the clinical effects, but with insufficient information. The demo only collect disease name and a part of these names are not provided or not specified.

In addition, dbSNP only provide a brief description of the clinical effects. More

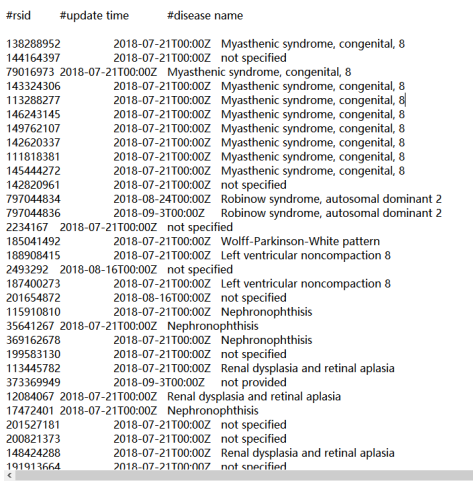


Figure 5.2: sample outputs of data mining program

details are linked to another database called ClinVar. Clinvar is a freely accessible, public database that provide medical reports of the relationships among human variants and phenotypes [12]. The information provided by Clinvar requires a high level of knowledge in genetics areas, therefore further research would be required.

The next version of the program should be able to collect the information of allele and Clinvar accession ID which links to the Clinvar database.

6 Task 4: Degradation of complete DNA

6.1 Aims

In previous tasks, all works are done by using incomplete DNA raw data. In this tasks, the team aims to work out how DNA analysis works on complete DNA reference files, and what could happen to the DNA analysis results if these complete DNA data are degraded to the same level of the Somerton man's DNA data.

6.2 Methods

Firstly, the team would obtained at least 2 complete DNA reference files from volunteers to conduct the tests. Once the complete DNA data is received, analysis done in task 2 should be executed again with the complete files. Observe the outcomes.

Then develop a program to degrade the complete DNA data to the same level of Somerton man's DNA data. Rerun the replacing program developed in task 2 to ensure the degraded file meet the minimum requirements of the GEDmatch. Therefore, the complete DNA reference files should be firstly degraded, then replace the empty SNPs with AA base pairs until the files have 2000 available SNPs for each chromosome. Finally, upload these modified DNA data to GEDmatch and rerun the tests for DNA analysis.

6.3 Methods

The team ordered 2 complete DNA reference files. But unfortunately, due to some unexpected issues, only one DNA reference file has been received. Another one should be received within a month, and at the start of next semester.

By uploading the complete DNA reference file to GEDmatch, a set of relative DNA has been found in the database. A part of these relative DNA kits are presented in figure 6.1.

Kit	1:1	Name	Email	Largest Seg	Total cM	Gen	Overlap	Date Compared	Testing Company
	A			14.0	154.1	3.3	48501	2019-05-08	WeGene
	A			14.1	153.6	3.3	48943	2019-05-08	-
	A			10.7	152.4	3.3	48370	2019-05-30	-
	A			10.7	152.4	3.3	48370	2019-05-30	-
	A			16.5	140.4	3.3	48519	2019-05-08	23mofang
	A			16.5	140.4	3.3	48519	2019-05-08	23mofang
	A			15.2	131.3	3.4	48378	2019-05-03	23mofang
	A			19.1	120.4	3.4	48144	2019-05-03	23Mofang
	A			11.9	120.3	3.4	48738	2019-05-08	23mofang
	A			10.8	115.6	3.5	48759	2019-05-08	-
	A			12.6	114.4	3.5	48918	2019-05-03	-
	A			12.2	97.9	3.6	48863	2019-05-08	-
	A			12.5	44.7	4.2	54132	2019-05-03	23andMe
	A			12.1	42.8	4.2	53953	2019-05-03	23andMe
	A			11.0	36.8	4.3	81879	2019-05-03	gesedna
	A			10.5	35.7	4.3	53677	2019-05-03	

Figure 6.1: parts of relative DNA to the complete DNA reference file

Figure 6.2 and figure 6.3 represent the ethnicity proportion of the complete DNA reference file and the degraded DNA file. According to the figures, the major proportions of ethnicity of complete DNA are east Asian (83.13%) and Siberian(14.82%). After the degradation, the proportion of other ethnicity increased, but the 2 major proportion east Asian and Siberian still occupied large area in the pie charts. This can be an evidence proves that the major proportion of ethnicity for an incomplete DNA could be the major ethnicity proportion of the complete DNA.

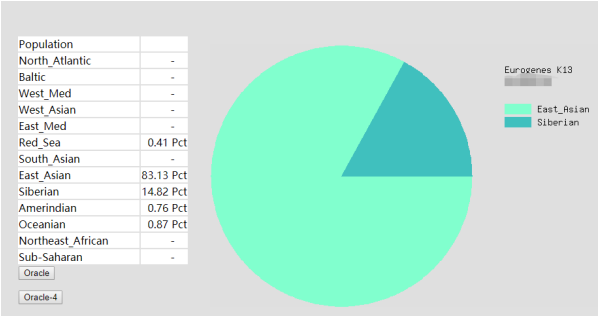


Figure 6.2: ethnicity proportion of complete DNA

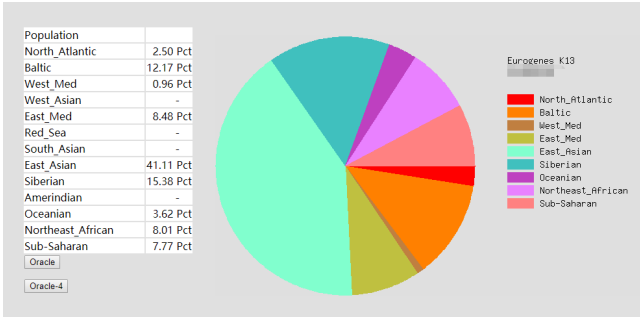


Figure 6.3: ethnicity proportion of degraded complete DNA

Figure 6.4 is the ethnicity proportion of a DNA kit that modified from Somerton man's DNA by replacing empty SNPs with AA base pairs until each chromosome has 2000 SNPs. In this figure, the greatest portion is North Atlantic. This can somehow lead to a clue that the ethnicity of Somerton Man is North Atlantic. But since the team only has one complete DNA file test results to support this theory which is clearly not a strong evidence. Further research or test would be required to study such theory.

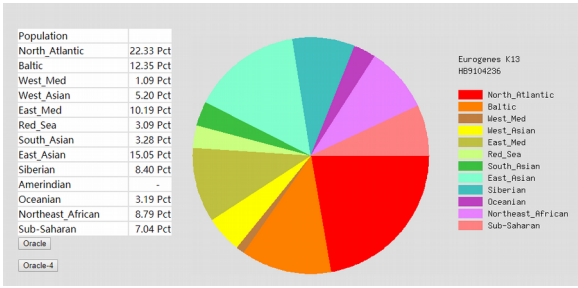


Figure 6.4: ethnicity proportion of Somerton man's DNA

7 Project Management

7.1 Budget

There are $250 budgets assigned to each member in the project, in which is $500 budgets in total for the project. Most budgets are spent on ordering 2 DNA kits from 23andme company for DNA testing. The details are shown in the table below. There is a plan on spending the rest of budgets on purchasing the advance services provided on GEDmatch. But the team is still evaluating demand of using these services.

Item	Quantity	Cost
23andme DNA kit (including shipping fees)	2	$200 each
	Total Cost	$400
	Remaining Budget	$100

7.2 Risk Management

The risk assessment table are listed below. Several risks occurred during the progress. One of the group member was absent in the meeting several times due to time clash. But there is always at least one member attend the meeting with the supervisor. Members sometimes misunderstand assigned task, but issues were always fixed in the meeting in the following week.

Risk	Likelihood	Consequences	Risk Estimation
Absence of meeting	**Unlikely**	**Minor**	**Low**
Miscommunication of members	**Unlikely**	**Moderate**	**Medium**

Loss of data	**Unlikely**	**Severe**	**High**
Delay of task completion	**Likely**	**Major**	**High**
Bugs in codes	**Likely**	**Minor**	**Medium**
Out of budget	**Rare**	**Severe**	**Medium**
Misunderstanding of tasks	**Unlikely**	**Moderate**	**Medium**
Unethical works	**Unlikely**	**Major**	**Medium**
Member drop the course	**Rare**	**Severe**	**Medium**
Bad quality of purchased items	**Unlikely**	**Major**	**Medium**

8 Conclusions

So far, the team has finished first 2 tasks, and have create several artificial DNA kits for testing. Also, the team has working on searching clinical effects of each SNP exist in Somertan Man's DNA raw data. And the last tasks degradation of complete DNA files has already started and several test results have been presented.

Unfortunately, there is no much information being found to identify any characteristic of Somerton man. The only clue is that the Somerton man might have part of North Atlantic ethnicity according to the results discussed in task 4 which can not be recognize as a strong evidence to this clue.

Although the current outcomes can not provide identify the Somerton man, the 2 major task (task 3 and 4) have already started, and could allow the team to develop more genetic information about the Somerton man in next semester.

8.1 Future Work

The major effort would be done in semester B are finishing task 3 and 4. Several suggestion are provided for future works:

- In task 3, the information provided by dbSNP and clinvar database require high level understandings in genetics. In order to save times on doing research, it suggest that to find a professional in genetics area to provide some advices for the team.
- In task 4, the team could degrade the files from higher levels to lower levels (remaining 80% SNPs to remaining 50% SNPs). Conduct the same DNA analysis in task 4 results, and observe how the results changes.

Appendix A: Source Code

A.1 Risk matrix

Risk matrix

Likelihood	Consequences				
	Negligible	Minor	Moderate	Major	Severe
Almost Certain	Medium	High	Very High	Very High	Very High
Likely	Medium	Medium	High	Very High	Very High
Slight	Low	Medium	High	High	Very High
Unlikely	Low	Low	Medium	Medium	High
Rare	Low	Low	Low	Medium	Medium

References

[1] Bineth, J, "Somerton Man: One of Australia's most baffling cold cases could be a step closer to being solved" *This Is About,* 13 December 2017. [online] Available at: https://www.abc.net.au/news/2017-12-14/somerton-man-cold-case-could-be-one-step-closer-to-solved/9245512 [Accessed 1 Jun. 2019].

[2] U.S. National Library of Medicine, "What is DNA?",*U.S. National Library of Medicine,* May. 28, 2019. [online] Available at: https://ghr.nlm.nih.gov/primer/basics/dna [Accessed 2 Jun. 2019].

[3] U.S. National Library of Medicine, "What is a chromosome?",*U.S. National Library of Medicine,* May. 28, 2019. [online] Available at: https://ghr.nlm.nih.gov/primer/basics/chromosome [Accessed 2 Jun. 2019]

[4] U.S. National Library of Medicine, "How many chromosomes do people have?",*U.S. National Library of Medicine,* May. 28, 2019. [Online]. Available: https://ghr.nlm.nih.gov/primer/basics/howmanychromosomes. [Accessed: 02- Jun- 2019].

[5] U.S. National Library of Medicine, "What are single nucleotide polymorphisms (SNPs)?",*U.S. National Library of Medicine,* May. 28, 2019. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. [Accessed: 02- Jun- 2019].

[6] G. Shaw. “Polymorphism and Single nucleotide polymorphisms (SNPs)” Science Made Simple, Vol. 112, pp.664-665 2013.

[7] “DEAD MAN FOUND LYING ON SOMERTON BEACH” *The News,* December 1, 1948, p. 1 [online]. Available: https://trove.nla.gov.au/newspaper/article/129897161. [Accessed: 03- Jun- 2019]

[8] “Cryptic Note On Body” *The News,* June 6, 1949, p. 1 [online]. Available: https://trove.nla.gov.au/newspaper/article/36371152. [Accessed: 03- Jun- 2019]

[9]"Raw Data Technical Details", 23andMe, 2019. [Online]. Available: https://customercare.23andme.com/hc/en-us/articles/115004459928-Raw-Data-Technical-Details. [Accessed: 04- Jun- 2019].

[10]S. Zhang, "The Coming Wave of Murders Solved by Genealogy", The Atlantic, 2019. [Online]. Available: https://www.theatlantic.com/science/archive/2018/05/the-coming-wave-of-murders-solved-by-genealogy/560750/. [Accessed: 04- Jun- 2019].

[11]"General Information about dbSNP as a Database Resource", Center for Biotechnology Information (US), 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK44469/. [Accessed: 06- Jun- 2019].

[12]"ClinVar: public archive of relationships among sequence variation and human phenotype", National Center for Biotechnology Information (US),

November 14, 2013 [Online]. Available:
https://www.ncbi.nlm.nih.gov/books/NBK44469/. [Accessed: 06- Jun- 2019].

Glossary and Symbols

DNA: Deoxyribonucleic acid

SNP: Single-nucleotide polymorphism