DNA-MAP, A Knowledge-Based Decision Support System for Australian Defence Force Forensic Ancestry Prediction

Kyle Anthony James

Bachelor of Forensic Science

Hons (I)

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Research Methods Group

and

Centre for Genomics and Personalised Health

School of Biomedical Sciences

Faculty of Health

Queensland University of Technology

2021

Acknowledgements

I would like to begin by thanking my supervisor Janet Chaseling, an exemplary person who not only helped me rekindle my joy of statistics but also provided me with this opportunity to achieve a dream I never thought possible. Janet, you have helped me improve myself both professionally and personally, reshaping me into someone I can be proud of, and I promise one day to stop waffling on in my writings! I will always appreciate the times we have worked together over these past several years and reminisce about the fun we had.

Dimitrios Vagenas, you have been such a wonderful supervisor, stepping in to provide me with support and guidance before I was even your student! You always have a bright idea on how to overcome any limitations I stumbled across, and your kind words kept me going during the toughest times.

To my supervisor Kirsty Wright, thank you for including me in a cause that I know is so dear to you, this project was incredible to work on and I am extremely grateful. Thank you for encouraging me to improve myself and seeing something in me I could not see myself. Public speaking has become so much easier now, and that would not be possible without the support and practice I achieved with your help!

Thank you to my fellow PhD candidate Andrew Ghaiyed, who supplied the data for this project, and Alanah Cronin who became my coding wizard and tutor. Additional thanks to Jasmine Connell, Melinda Mitchell and Lee Jones who have provided both support and numerous brainstorming for ideas.

A massive thank you to the members of QUT and especially IHBI for providing me with a home for the latter half of my project, with special mentions to my supervisor Lyn Griffiths, Larisa Haupt and the other members of the GRC for their support and encouragement. To my supervisor Albert Gabric, thank you for your support and guidance during the first half of my PhD and helping me through my confirmation. I would also like to acknowledge and thank the Australian Government Research Training Program Scholarship, which allowed me to complete my research while maintaining full-time enrolment.

None of this would be possible without the love and support of my family and friends (who finally have an answer to the question "are you finished yet?"). Especially, my mother who has provided me with all the support possible from another human being during these stressful times, words cannot describe how grateful I am.

This thesis is written in honour of Cliff Simpson, whose personality and supporting words could brighten even the darkest room, you were truly a gift.

Keywords

Biogeographic ancestry prediction, Classification tool, Generic Bayesian, Logistic model tree, Parsimony, Rare event detection, Simulated admixture, STRUCTURE.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date: 02/08/2021

Abstract

"*One of the most important reasons to identify unknown persons is because non-identification may result in numerous issues at emotional and legal level for the surviving family members and friends.*" (Beauthier et al., 2009, p.54).

Reliable prediction of biogeographic ancestry (BGA) requires a complex biological and statistical process. BGA can be used for criminal case work, missing persons, counterterrorism, ancient remains, including historical military remains. In this thesis, a knowledge-based decision support system (KBDSS) is developed which can assist researchers and investigators perform BGA prediction without requiring the user to have a statistical background. This thesis presents a case study on the use of the KBDSS, named DNA-Military Ancestry Predictor (DNA-MAP), for unidentified WWII remains recovered from the Asia-Pacific by Unrecovered War Casualties-Army (UWC-A). Specifically, the KBDSS is used to estimate the probability of the remains being those of an Australian soldier, or those of a Japanese soldier, to facilitate decisions about the soldier's final resting place. To determine the requirements of a KBDSS for BGA prediction, two sets of literature reviews were performed. The first analysed (*i*) the literature surrounding the creation and validation of KBDSSs, with particular emphasis on a system's typical format. This literature review identified key components for a KBDSS such as, the user inputs to be provided to the system, the processing hub where statistical modelling is performed, and the format of the outputted report. Based on these components, a second literature review (*ii*) was performed on studies of BGA prediction to identify the factors that affect BGA prediction, while considering new factors not previously considered. This review resulted in the identification of ten factors required for reliable BGA prediction, namely: (1) admixture, (2) parsimony, (3) choice of classifier, (4) defining the relevant populations, (5) sample size in the relevant population, (6) selecting a sample from the relevant populations, (7) the possibility of rare events, (8) prior probability, (9) degradation and (10) the inclusion of a margin of error.

Subsequently, the literature was searched for each of the ten BGA factors, to determine how, if at all, these factors had been addressed previously in the forensic and non-forensic literature. Following a comparison of BGA classifiers previously been used in the literature, a new classifier was chosen for implementation, namely, the Parsimonious Logistic Model Tree (*p*LMT) classifier (Landwehr et al., 2005). The *p*LMT consists of iterative applications of the basic Logistic Model Tree to create successive models, each of which provides an independent estimate of the required probability of ancestry. These independent estimates are then combined using a geometric mean average which becomes the final value used to determine ancestry. The method is parsimonious since it uses only

the components of the data (markers in a DNA panel) which are deemed the most informative based on a normalised information gain (also referred to as the Kullback-Leiber divergence). An additional difference from the methods of BGA prediction found in the literature, is that the pLMT procedure does not necessarily use all markers available in the original dataset, rather, each analysis is performed only using the markers contained in the unknown sample. The unknown sample information drives the modelling process.

The effectiveness of the *p*LMT classifier for BGA prediction was compared to two commonly used classifiers in forensics and BGA prediction, the program STRUCTURE and the Generic Bayesian method. To compare these three classifiers, and address the other BGA factors identified, a case study of a current UWC-A operation was considered. The focus of the case study was remains from World War II Australian and Japanese soldiers in the South-East Asia Pacific. Discrimination between these two populations was achieved using the ancestry informative DNA marker panel, Ghaiyed Population Specific Panel (GPSP) (in collaboration with Ghaiyed (2020)), which contained 40 single nucleotide polymorphisms (SNPs) selected for their ability to distinguish between Australians and Japanese.

The available samples were: (*i*) WWII era Australian individuals (n = 108), (*ii*) Contemporary Japanese individuals found in the 1000 Genomes Project (n = 104), and (*iii*) degraded WWII era Australian samples (n = 80). Samples (*i*) and (*iii*) were collected in collaboration with Ghaiyed (2020). Due to the small sample sizes available, and the lack of publicly available data, additional samples were generated for testing. Individuals were simulated based on various admixed pedigrees were simulated using an admixture simulation tool, SimAdmixtR (available at https://github.com/danwkenn/SimAdmixtR, Kennedy (2019)). For a DNA sample with a complete GPSP profile, the *p*LMT created a classification model which consisted of five independent LMT models which involved 34 out of the 40 total SNPs.

Using simulated data, ten scenarios were created which consisted of varying degrees of admixture between Australian and Japanese ancestors across four generations. For each scenario, a sample of ten thousand individuals was simulated. Initially, simulated data were used to define classification thresholds of the predicted probability of Australian ancestry which would enable one of three outcomes of allocated ancestry: (*i*) Australian, (*ii*) Ambiguous and (*iii*) Japanese. A second, independent group of simulated data (using the same scenarios but each with ten thousand newly simulated individuals) were then used to validate these thresholds. The establishment and validation of thresholds were performed for all three classifiers. The following metrics were used to compare the classifiers: (*i*) the number of Australian individuals who were correctly classified as Australian,

(*ii*) the number of Australian individuals who were incorrectly classified as Japanese, and (*iii*) the number of Australian individuals who could not reliably be assigned to a population group. The results of the *p*LMT classifier's validation when unknown samples included the complete GSPS panel, showed that for individuals with all Australian ancestors (no admixture), 99.48% were successfully classified as Australian and none were incorrectly classified as Japanese; the remaining samples were Ambiguous. For simulated individuals with a pedigree consisting of 75% Australian ancestors, approximately 80.64% were classified as Australian, while the remainder were Ambiguous. The equivalent figures were 100% and 73.15%, respectively, for the Generic Bayesian classifier, and 100% and 73.84%, respectively, for STRUCTURE, with the remaining individuals again being classified as Ambiguous. For all three classifiers, the percentage of individuals who could be classified as Australian, as opposed to Ambiguous, declined as the admixture proportion increased, however, no cases of incorrect assignment were observed. The result of the comparison between all three classifiers showed that there were admixture scenarios where the percentage of correctly classified individuals with complete GPSP profiles was equal to or greater for *p*LMT than for both STRUCTURE and the Genetic Bayesian classifier.

A degradation experiment using simulated data and randomly removed sets of SNPs was performed using the *p*LMT classifier to determine the minimum number of SNPs needed for accurate classification, information which is needed for the development of guidelines. A minimum of ten SNPs, out of the GPSP's original 40, was recommended for accurate model creation. This recommendation was made after the *p*LMT was found to be unable to create accurate classification models with less than ten SNPs. Note that there were still instances where ten SNPs were available but the *p*LMT was unable to create an accurate classification model, suggesting that the recommended ten may be conservative.

The effectiveness of the three classifiers on degraded samples (incomplete GPSP profiles) was also tested using a subset of the WWII Australian samples (n = 80) which had incomplete GPSP profiles thus representing degraded samples. Five of the available 80 were excluded from the experiment as they had less than ten SNPs, based on the results of the previous degradation experiment. Both STRUCTURE and the *p*LMT classifiers, which classified all 75 samples as Australian (100%), outperformed the Generic Bayesian classifier which classified 40 individuals (53%) as Australian and 35 individuals (47%) as ambiguous. These results indicate that both STRUCTURE and the *p*LMT classifiers are more readily suited for classifying degraded samples, such as would be expected in forensic case work. To test how the number of SNPs available affected a classifier's performance, a regression analysis was performed between the number of missing SNPs and each of the classifier's

outputs, (*i*) the probability of Australian ancestry for the *p*LMT, (*ii*) the Australian Q value for STRUCTURE, and (*iii*) the natural log of the likelihood ratio of Australian ancestry for the Generic Bayesian. The results of the regression demonstrated that the *p*LMT's outputted probability of Australian ancestry was significantly affected by the number of missing SNPs, where a reduction of ≈ 0.001 in the Geometric Mean of Australian Membership Probability was observed for each SNP removal ($p = 0.0477$). Note that this reduction was slight, and all degraded samples ($n = 75$) were still correctly classified as Australian using the established thresholds. Based on the obtained R-squared value, approximately 20% of the variation observed was due to the number of missing SNPs. For STRUCTURE, there was no significant reduction observed in the Australian Q value ($p = 0.776$), demonstrating that the number of missing SNPs did not contribute to the calculated Q value. Finally, for the Generic Bayesian classifier, a significant reduction in the log likelihood ratio of 2.941 ($p \leq 0.001$) for each additional missing SNP, with the number of missing SNPs accounting for approximately 92% of the variation, based on R-squared. Note that this equates to an average reduction of approximately 19 in the likelihood ratio for each missing SNP.

The Generic Bayesian classifier suffered from several limitations, namely, reduced classification ability for degraded samples, assumptions which may lack a scientific basis in real-casework, difficult comprehension for reported statements and the inability to handle values of zeros (the likelihood ratio). While STRUCTURE is currently considered the 'gold-standard' for ancestry prediction, it suffers the limitations of lengthy run-times, silent crashes, potential bias when selecting a value for *K* (the number of assumed populations) and assumptions which may lack a scientific basis in real-casework. Given these limitations of these methods the *p*LMT classifier is a suitable alternative classifier which provides a shorter run-time (minutes as opposed to STRUCTURE's hours) and introduces a parsimonious nature to the aspect of classification not previously accessible in other classifiers.

For casework where historical knowledge is available (before the DNA is considered), it is possible to include an estimated weighting (a prior odds ratio), based on a believed difference in the present sizes of the two populations of interest. This prior odds ratio can subsequently be used to inform the conditional probability of a DNA profile being observed in either population (due to shared genes or the possibility of a rare event), thus providing an updated posterior probability. A sensitivity analysis was performed to evaluate how the posterior probability is affected by three factors: (*i*) the probability of Australian ancestry estimated using DNA evidence (the conditional probability), (*ii*) the sample size of the original data available for this estimation, and (*iii*) the prior odds ratio. It was found that the posterior probability is dramatically affected by the prior odds value, with an estimated

probability of 0.99 assuming an equal odds prior representation, being reduced to a posterior probability of 0.93 for a prior odds value of 0.5 and being further reduced to a posterior probability of 0.57 for a prior odds value of 0.05. In addition, the sample size of the data used for model formulation had a direct effect on the likelihood ratio, which is used to estimate the posterior probability. A posterior probability of 0.57 (Prior Odds = 0.05) obtained using a sample size of 100 individuals was increased to 0.87 when a sample size of 500 individuals was used.

The Delta method was used to estimate the variance of a function to allow for confidence intervals to be applied to the resulting probability of Australian ancestry. Providing a measure of the variation which is inevitable when the data are samples representing a much larger population, enables a measure of the margin of error, or reliability, to be placed on the results. Rather than outputting only a point estimate, information is provided to account for the presence of sampling error and its propagation across the modelling process.

The statistical models created from this research were then developed into a user-friendly KBDSS software application for BGA prediction, DNA-Military Ancestry Predictor (DNA-MAP). DNA-MAP provides the user with an estimate of ancestry in a clear, English statement which includes a measure of reliability, and provides various suggestions to the user for consideration to create a feedback loop. Suggestions include demonstrating how the results will vary if the user selected a different level of confidence or if a larger sample size had been used. DNA-MAP is the prototype of an ancestry prediction tool that is the first of its kind in forensic BGA prediction with key features including a parsimonious approach to marker selection, adaptability to any case relevant DNA panel, incorporation of rare event detection and prior information, provision of a measure of reliability, and a process which allows the information in the unknown sample to drive the classification models. These are features that have either not been addressed previously in the literature or not compiled into a single BGA prediction tool.

Research Outputs

R.1 Proposed Publication Strategy

James, K, Chaseling, J, Wright K, Griffiths L R, Vagenas, D, "Statistical Criteria for Accurate Biogeographic Ancestry Prediction".

James, K, Chaseling, J, Wright K, Ghaiyed, A, Griffiths L R, Vagenas, D, "Parsimonious Logistic Model Tree: a binary approach for inferring ancestry".

James, K, Chaseling, J, Wright K, Griffiths L R, Vagenas, D, "DNA-Military Ancestry Predictor: a knowledge-based decision support system for biogeographic ancestry prediction".

R.2 Oral and Poster Presentation Record

Australian and New Zealand Forensic Science Society International Symposium, Auckland, New Zealand, 2016.

- "*Sensitivity Analysis of Variables Involved in Ancestry Prediction of an Unknown Individual*". **Poster** Presentation.

Haploid Markers, Bydgoszcz, Poland, 2018.

- "*DNA-Military Ancestry Predictor (DNA-MAP): Software to Assist in Ancestry Prediction of Unidentified Historical Military Remains*", **Poster** Presentation;
- "*Rare Events – The Rare Topic in Forensics!*", **Poster** Presentation.

Australian and New Zealand Forensic Science Society International Symposium, Perth, Australia, 2018.

- "*DNA-Military Ancestry Predictor (DNA-MAP): Software to Assist in Ancestry Prediction of Unidentified Historical Military Remains*", **Oral** Presentation;
- "*Rare Events – The Rare Topic in Forensics!*", **Poster** Presentation.

Australasian Applied Statistics Conference, Rotorua, New Zealand, 2018.

- "*DNA-Military Ancestry Predictor (DNA-MAP): Software to Assist in Ancestry Prediction of Unidentified Historical Military Remains*", **Poster** Presentation;
    - Awarded **"Best Poster Presentation – Runner Up"**
- "*Rare Events – The Rare Topic in Forensics!*", **Oral** Presentation.

Table of Contents

List of Figures

List of Tables

Abbreviations

ABS	Australian Bureau of Statistics
ADF	Australian Defence Force
AIM	Ancestry Informative Marker
AP	Achievable Production
APSIM	Agricultural Production Systems Simulation
ASCII	American Standard Coding for Information Interchange
ASHG	American Society of Human Genetics
AUROC	Area Under the Receiver Operating Characteristic
BGA	Biogeographic Ancestry
δ	Absolute Allele Frequency Difference (*Delta*)
DNA	Deoxyribonucleic Acid
DNA-MAP	DNA-Military Ancestry Predictor
DVI	Disaster Victim Identification
ECDF	Empirical Cumulative Distribution Function
F_{st}	Fixation Index
GBR	Great Britain
GMAMP	Geometric Mean of Australian Membership Probabilities
GPSP	Ghaiyed Population Specific Panel
GUI	Graphic User Interface
HGDP	Human Genome Diversity Project
KBDSS	Knowledge-Based Decision Support System
LMT	Logistic Model Tree
LR	Likelihood Ratio

MAF	Minimum Allele Frequency
MCDA	Multiple-Criteria Decision Analysis
MCMC	Markov Chain Monte Carlo
NAS	National Academy of Science
NATA	National Association of Testing Authorities
NRC	National Research Council
ORCA	Optimal Rate of Correct Assignment
PCAST	President's Council of Advisors on Science and Technology
*p*LMT	Parsimonious Logistic Model Tree
PNG	Papua New Guinea
PPAA	Posterior Probability of Australian Ancestry
PVSEL	Pressure Vessel SELection
rCRS	Revised Cambridge Reference Sequence
RAP	Average Milk Production Across a Region
RMP	Random Match Probability
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
STR	Short Tandem Repeats
SWGDAM	Scientific Work Group of DNA Analysis Methods
TMB	Tetramethylbenzidine
UWC-A	Unrecovered War Casualties – Army
VNTR	Variable Number Tandem Repeat
WEKA	Waikato Environment for Knowledge Analysis
WWII	World War Two

Lexicon

Admixture Two types of admixture are discussed in the biogeographic ancestry, archaeology and ancient DNA literature, namely population-level admixture and family-level admixture. The former, is the result of two or more populations interbreeding in previous eras during settlements and geographic migration. The latter corresponds to ancestors from different population groups being introduced into a recent family pedigree resulting in offspring who carry a proportion of genetic information from each ancestral population. Note that only the latter is of interest in this thesis, namely, family-level admixture going back four generations (great-grandparents).

Population A homogenous group of individuals who have been selected based on one or more demographics/criteria using self-declared information. Therefore, the individual's self-declaration is assumed to be fact despite the possibility that their self-believed ancestry may differ from their true-genetic ancestry.

Validation "*Validation' is broadly defined as the process by which the scientific community acquires the necessary information to assess the ability of a procedure to obtain reliable results, determine the conditions under which such results can be obtained, and define the limitations of the procedure*" (Ogden et al., 2009, p.187). In the forensic community, the term validation is associated with a method being rigorously tested against standards which have been established by a governing accreditation body, such as the National Association of Testing Authorities (NATA) (Ogden et al., 2009). However, in this thesis, the term is used in the broader sense, that is, confirming that the outputs of a statistical model are correct in a situation where the answer is known.

Chapter 1 – Introduction

1.1 Project Overview

The prediction of biogeographic ancestry (BGA) is a complex process that requires extensive biological and statistical testing. Note that the focus of the research presented in this thesis is on the statistical aspect of BGA prediction; it will be assumed that the biological testing prior to statistical modelling is performed accurately.

An individual's BGA is the culmination of unique biological variations in their DNA that have occurred over numerous generations based on geographic location. For example, a human clade that through migration and geographic separation (such as the loss of land bridges over time) evolved independently, would genetically present as homogeneous, with little change (barring random mutation) over time. Compare this to a second human population which was situated in a "migratory hub" and became exposed to genetic mixture between multiple coinciding populations; for example, Europe. This second population would evolve to display a completely different range of genetic profiles compared to the first which only had access to the genes that were available in the original ancestors. To infer BGA for an unknown individual, numerous sections of the unknown individual's DNA, known to be highly variable between populations, are examined and comparisons are made to determine with which of the populations of interest, if any, the unknown individual shares similarity. This technique presents itself as a complex issue of BGA analysis; the initial search for these highly variable regions, requires extensive sampling from the populations of interest and thorough genetic testing of the human genome.

Previous BGA studies have determined that achieving accurate prediction is based on a combination of collecting representative samples from the true populations of interest, a DNA panel with high discrimination power, and a statistical classifier with high accuracy (Cheung et al., 2017; Phillips, 2015). To determine how these three factors, and any other factors which may be identified, have been previously addressed in the literature, a review of relevant BGA prediction methods will be performed. In instances where an approach has not been addressed in the relevant forensic literature, literature from other disciplines will be explored for possible alternative methods. These factors will be combined into a user-friendly knowledge-based decision support system (KBDSS), where information from various sources related to an individual's BGA can be uploaded and analysed in a single program. These sources include the genetic information provided by the utilised DNA panel, historic information (or relevant case information), the possibility of a rare event, and accounting for sampling error as determined by the size of the available sample.

To assist with the creation of this KBDSS, the current operations of Unrecovered War Casualties – Army (UWC-A) will be utilised as a case study. It is estimated that there are thousands of unaccounted for Australian soldiers whose remains have yet to be recovered after dying in past military conflict (Unrecovered War Casualties – Army, n.d.). UWC-A, a multidisciplinary team within the Australian Defence Force (ADF), was formed to investigate areas where there may be the remains of Australian solders. When a set of remains are discovered UWC-A attempts to determine their possible ancestry to decide where the soldier should be laid to rest, and where possible for Australian soldiers, identify who the soldier is by name. By incorporating a user-friendly KBDSS into UWC-A procedures, BGA prediction may be possible without the user requiring a statistical background or experience in the necessary computer coding languages needed to perform the analyses. It is important to note that the KBDSS will act only as a decision-supporting tool, and that the decision-making process of assigning ancestry will ultimately be performed by UWC-A. This thesis will detail the construction of the KBDSS, named "*DNA Military Ancestry Predictor*" (DNA-MAP), and outline the proposed statistical methodology for BGA prediction for an unknown individual while clearly stating and explaining the assumptions and limitations of the underlying processes. For this methodology it is assumed that the outcome variable of BGA is binary, that is, only two populations are considered for this case study, these populations are WWII-era Australian and Japanese soldiers. It is acknowledged that most individuals in the Australian forces in the WWII-era were individuals with a biogeographic origin related to ancestors who settled into Australia from a European background and therefore can be considered as having 'European-Australian' ancestry. For the remainder of this thesis the term Australian will refer to these individuals. Note that it is recognised that the remains of individuals from other nationalities may be recovered by UWC-A, such as American or local indigenous groups, however, in this thesis a binary approach was adopted to demonstrate proof of concept in the process.

An important aspect of this thesis is accounting for possible error, and where possible, providing processes which will mitigate error. There are two types of error that are critical for ancestry prediction: classification error and sampling error. Classification error relates to the possibility of the implemented method/model misclassifying an individual into an incorrect ancestry. This type of error can be mitigated by applying certain techniques to make the method/model more conservative and reduce the opportunity of under/overfitting. These techniques include options such as cross-validation and will be discussed further in this thesis. Sampling error is affected by how accurately the sample taken reflects the true greater population. Such error will be minimised by ensuring that

the individuals utilised in the sample are representative of the population of interest (which may vary from the population as a whole), and that an adequate size is taken.

Due to the standards of forensic casework aiming to maintain as low an error as possible, reducing the chance of misclassification is an important aspect of this thesis and DNA-MAP's operations. For misclassification to occur, an individual would need to deviate from their own population group sufficiently to enable them to be more closely associated with a different population group. If this deviation is not taken into account in into the classification system, a misclassification could occur. It is proposed by this thesis that the method applied must ensure the chance of such a deviation occurring would be extremely low, but at the same time acknowledge and quantify the remaining possibility. Methods of detecting a rare event are utilised in DNA-MAP to ensure that if a deviation as described did occur it would be detected and accounted for in the classification model.

1.2 Global Objective

The global objective of this research is to develop a user-friendly KBDSS that would provide UWC-A with a reliable BGA prediction estimate that mitigates the chance of misclassification of WWII soldiers' remains to almost zero and aids in their decision-making, reducing the proportion of 'ambiguous' classifications.

1.3 Specific Aims

Five specific aims are proposed in this project, namely, to:

1. Identify and review key components and factors from the relevant literature on KBDSSs and BGA prediction studies;
2. Determine methods of ensuring relevant populations are selected given the DNA panel and specific situation;
3. Develop and validate an optimal methodology for classifying individuals, and compare said method to other alternative methods from the literature;
4. Expand upon the currently used methodology to incorporate prior information, degradation of DNA (missing data) and measures of reliability and error;
5. Develop a user-friendly KBDSS using the developed methodology;

1.4 Significance

The primary gap that was identified during this research is the lack of guidelines in the forensic science community describing how accurate and precise BGA prediction should be performed. In this thesis, several statistical factors were identified that play a key role in BGA prediction that are

either overlooked in the literature or as of yet have no recommended universal approach. Of particular importance are the detection of a rare event, and the incorporation of prior historical information together with genetic data. Where available, methodology has been outlined in this thesis to demonstrate possible approaches for other applications of BGA prediction.

The creation of DNA-MAP provides a user-friendly software application for binary BGA prediction. While software such as STRUCTURE can be used as a predictor of BGA, such tools stop short of providing a clear statement with an interpretable measure of reliability together with a warning of their limitations.

DNA-MAP is adaptable for any binary classification method and can be utilised with different populations and DNA panels. By recording which, if any, SNPs are missing from the unknown test sample's profile and subsequently removing the same SNPs from the original training data, DNA-MAP ensures that the unknown profile drives the modelling process. The use of the unknown profile as the driving force has not been previously observed in the literature. DNA-MAP can readily create a new classification scheme that is appropriate for the specific case of interest. This SNP removal process allows DNA-MAP to adapt to samples with missing data, which may occur from DNA degradation.

Two important aspects of DNA-MAP's algorithm are the incorporation of a prior odds ratio and the provision of a given level of confidence in the obtained results. A posterior is calculated by incorporating empirical cumulative distribution functions to estimate the probability of observing a BGA profile in the populations of interest; said probability is then combined with a prior odds ratio chosen by the user. Providing a measure of the associated error of a resulting probability is paramount to DNA-MAP's reporting process. Each of DNA-MAP's two outputted statistics, the geometric mean of the probability of Australian ancestry and the posterior probability, have a method for estimating the associated variance which is then used to calculate a confidence interval. These methods utilise the Delta method, which allows DNA-MAP to estimate the variance of a function, ensuring that any propagation of error is accounted for during analysis, ensuring that the result is not simply stated as a point estimate.

1.5 Structure of Thesis

The structure of this thesis will be as follows:

Chapter 2 provides an overview of a typical KBDSS, exploring the relevant literature to determine the key components and functions of these systems by analysing four specific KBDSSs selected from the literature. *Aims addressed in this chapter: 1.*

Chapter 3 provides a literature review of BGA prediction studies, outlining a list of factors that may affect a classifier's effectiveness. This chapter compares previous approaches performed in the forensic literature and other disciplines. *Aims addressed in this chapter: 1.*

Chapter 4 details the materials and methodology used in this thesis, and outlines the case study, and how the two classifiers selected for testing were utilised. *Aims addressed in this chapter: 2, 3, 4.*

Chapter 5 provides the results obtained from the methodology outlined in Chapter 4. *Aims addressed in this chapter: 3, 4.*

Chapter 6 describes DNA-MAP's process, how the results from Chapter 5 were incorporated into the KBDSS software application to create a decision-supporting tool for UWC-A. *Aims addressed in this chapter: 5, 6.*

Chapter 7 is the discussion, suggesting where methods may have been limited, and proposing future directions for additional research with concluding remarks about the overall research.

Chapter 2 – Implementing a Knowledge-Based Decision Support System

2.1 Introduction – The Need for an Information Hub

Prediction of BGA for an unknown set of skeletal remains is a complex process, which involves a variety of biological and statistical methods and which can encompass unwieldy amounts of data. Benefit, therefore, lies in creating a "meeting point" system, through which the various types of data involved in BGA prediction can be accurately analysed and combined into a single output. Simply put, this "meeting point" system could be constructed in the form of a packaged algorithm contained within a programming language such as R or Python, however, a standalone algorithm requires the user to have some minimal understanding of the programming language. The primary intended users for a BGA prediction system are forensic scientists, the military, and police investigators, and it cannot be assumed that every user would have training or even an interest in the relevant programming language. Therefore, the system must have a user-friendly front-end that requires as few instructions and as little training as possible.

It was decided that a KBDSS would be a suitable format for a BGA prediction tool. The utility of a basic Decision Support System (DSS) becomes apparent when one considers the nature of ancestry prediction for forensic investigations. A DSS ensures the human element of the decision-making process is kept intact, by supplying the pertinent information in an accessible format that assists the user with their decision (Leni et al., 2013). Marin (2008) details that most DSS fall into one of five type-based categories, (*i*) communication, (*ii*) data, (*iii*) document, (*iv*) knowledge, and (*v*) model. The proposed KBDSS belongs to the knowledge-based category. A KBDSS is described as a specialised problem-solver which is designed for a specific task and can "…*suggest or provide action*…" to the user (Marin, 2008, p.2).

This chapter will use four KBDSSs selected from the literature to establish common themes seen in such tools; it will identify the primary components of any KBDSS and develop a proposed plan for the development of a KBDSS for BGA prediction.

2.2 What is a KBDSS?

Black and Stockton (2009, p.1) state that KBDSSs can be broadly defined as "*computational systems that provide access to a wealth of information pertaining to a specific problem*.". KBDSS have also been said to utilise knowledge from numerous sources to support an expert in problem-solving and decision-making (Workneh et al., 2019). The key term in this statement is "support", as discussed in Section 2.1 the objective of a KBDSS is to provide detailed reports summarising various inputs of data, but ultimately, the outputs are used to inform a human decision-maker. Since their conception, KBDSSs have been developed for a wide range of disciplines, with some examples from the literature shown in Table 2.1.

To date, a KBDSS has not been utilised for BGA prediction, making this thesis the first implementation of such a system in this discipline. However, in the broader forensic literature there have been instances of DSSs being utilised in areas such as entomology or policing (Morvan et al., 2007; Noor et al., 2014; Oatley et al., 2006; Shen et al., 2006).

Table 2.1: Examples of Applied KBDSS found in the Literature.
Knowledge-based decision support systems for a variety of disciplines which have been implemented to assist one or more users to improve quality/efficiency in the given area.

Study	Discipline	Scope	Primary User (Decision-Maker)
Kerr et al. (1999a), Kerr et al. (1999b)	Agriculture	Optimisation of factors used in dairy farming to improve overall milk production	Dairy farmers, bank managers, loan officers, farm consultants
Ritchie (1990)	Traffic Management	Addressing congestion in large or complex traffic networks	Control room staff
Uricchio et al. (2004)	Water Treatment Management	Evaluating relationships between human activities and environment conservation	Environmental resource managers
Workneh et al. (2019)	Clinical Research	Detection and diagnosis of acute abdominal pain	Physicians
González-Ferrer et al. (2018)	Healthcare	Assisting civilians and health professionals detect mistakes, reducing wasted resources, selecting health policies	Health Professionals
Yurdakul et al. (2020)	Manufacturing	Selection of materials for creating high-pressure components in machinery	Manufacturers
Zouri et al. (2019)	Clinical and Health Management	Improving patient quality-of-life based on various performance metrics	Managers and Physicians
Asad et al. (2019a)	Industrial Safety Management	Prevention of hazardous activities in oil and gas drilling operations	On-site industrial managers
Jung and Chung (2016)	Health Management	Providing recommendations for preventative management and health improvements for obese youths.	Dieticians and Obese Youths.
Jo et al. (2016)	Education	Identify key elements of a smart classroom (integrated with IT) to achieve positive effects on education.	Teachers and Administration

A typical KBDSS is comprised of three layers: *i*) inputs, *ii*) statistical modelling, and *iii*) reported output. Figure 2.1 illustrates this in a generalisation, note that more layers may be required for systems of greater complexity.

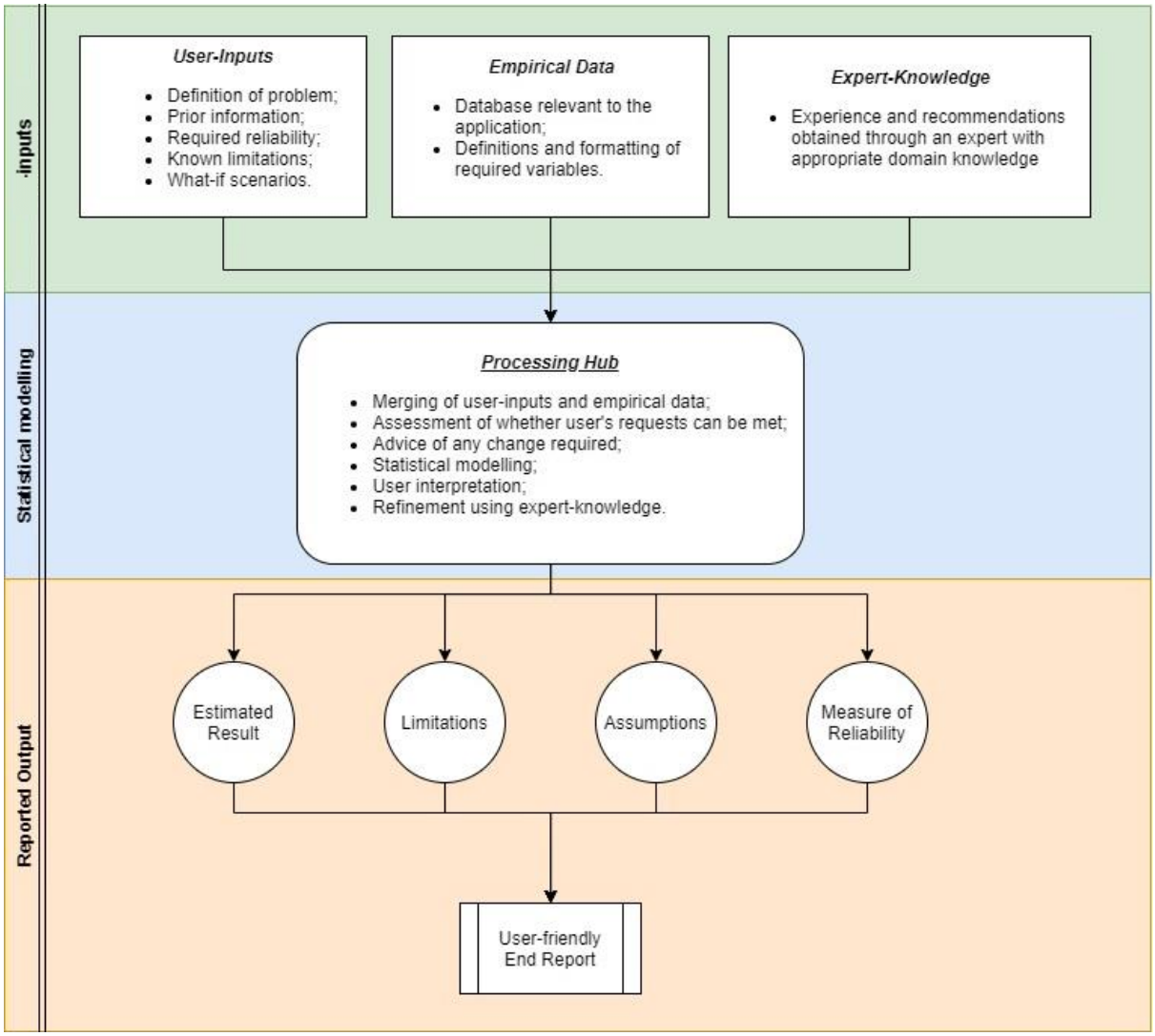


Figure 2.1: Generalised Flowchart of a KBDSS
The general KBDSS will consist of three primary phases i) inputs, ii) statistical modelling, and iii) reported output, each phase handles multiple functions within the system.

The input layer is where the KBDSS's user will spend most of the time when interacting with the system, often in the form of a user-friendly front-end interface. It is here that the user will provide the system with key information that is pertinent to the user's targeted scenario. In addition to user-uploaded data, a KBDSS will usually have input components of data and expert knowledge, the latter based on ideal circumstances provided by an expert in the relevant discipline. The availability of expert knowledge allows these systems to make comparisons between the user's observed inputs and an expert's ideal standard (Asad et al., 2019a); recommendations can be suggested to the user based

on any discrepancies. The expert selected will be a professional with a wealth of experience and knowledge in the relevant discipline which they have obtained over a long period of time.

With all relevant user-inputs uploaded, the system will then proceed to the statistical modelling phase, which is often invisible to the user. As previously discussed, the users of KBDSSs are not expected to have the required training in the relevant programming language or to be statisticians. Therefore, the statistical modelling phase of a KBDSS itself provides little to the user in most cases, and instead may overwhelm or confuse them. Selecting which modelling techniques to implement in a KBDSS is closely related to the problem at hand. As will be shown in this chapter when examining KBDSSs from the literature, a wide variety of statistical techniques are utilised in these systems.

Once the user's data has been successfully analysed, the final stage of the KBDSS is to provide the user with a report of recommended actions or suggestions. As with the statistical modelling phase, the format of the reporting phase of a KBDSS will be largely driven by the system's purpose. Reports may contain: (*i*) suggested actions for the user, (*ii*) sensitivity analyses in the form of what-if scenarios, and/or (*iii*) simplified statements detailing the outcomes of utilised statistical models. Section 2.3 examines four KBDSSs that were selected from the literature to identify any common themes that are utilised across different KBDSSs. These KBDSSs were selected due to having a clear, and concise overview of their KBDSS's construction, methodology, and validation phases.

2.3 Real-World KBDSS

2.3.1 HAZFO Expert 1.0

Created by Asad et al. (2019a), Hazard Free-Operation (HAZFO) Expert 1.0 was designed to improve the industrial safety management for onshore and offshore drilling sites. It was estimated that within the period of 2014 – 2019 an average of 70 individuals died and 2,500 suffered major injuries annually as the result of insufficient hazard preventions on oil and gas drilling sites (Asad et al., 2018a; Asad et al., 2018b; Asad et al., 2019a). Previous DSS in the discipline either lacked a sufficient database of potential hazards resulting in poor performance or were designed for a different purpose such as environmental and climate change prediction (Asad et al., 2019b). The first stage of HAZFO's construction, was collecting and creating the system's internal knowledge base, namely, what are the expected hazards and preventative measures that occur on the drilling site. Seven onshore and nine offshore drilling sites were examined throughout Malaysia, Saudi Arabia, and Pakistan, where a total of 150 possible potential hazards and 510 hazard preventative measures were found. Both the hazards and preventative measures were discovered through a combination of quantitative (using descriptive statistics) and qualitative (what-if scenarios) techniques.

The base model of HAZFO's system can be described as a comparative loop between the user's current safety measures and knowledge obtained from experts in the field on an ideal state of safety. By using IF and THEN conditional rules, HAZFO can perform operations on the user's data to recommend where, if any, improvements can be made to hazardous safety measures (Figure 2.2).

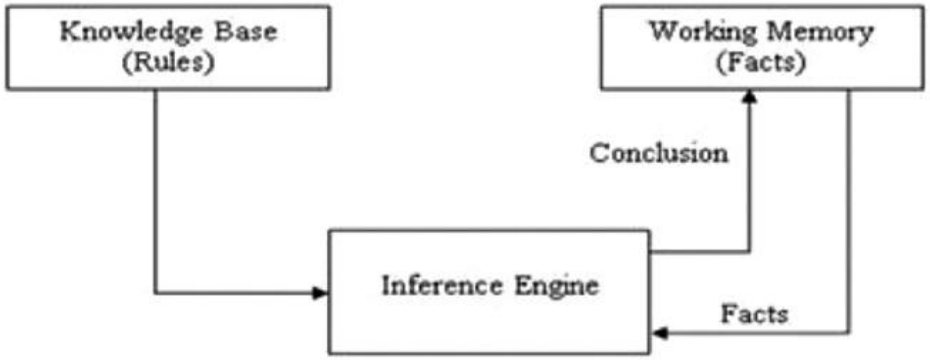


Figure 2.2: HAZFO Expert 1.0's Base Decision Model
By performing IF and THEN statements (Knowledge Base) on the user's observed data, the system can make inferred suggestions to safety measures based on an integrated ideal scenario provided by experts in the discipline (Working Memory) (Obtained from Asad et al. (2019a, Figure 1, p. 707).

The statistical modelling utilised by HAZFO involves Structured Query Language (SQL) interrogations of a predesigned database. Care was taken to ensure the collective sample size of expert knowledge and onsite inspections depicted complete management of safety with the possibility of injury as low as possible. The original intent of HAZFO was to improve the safety measures taken for on-site locations, however, additional applications were also discovered for the KBDSS. These were the conducting of risk assessments and job safety analysis pre- and post-drilling operation, and the implementation in education institutes to train the workforce prior to field work (Asad et al., 2019a).

The development of the KBDSS HAZFO can be summarised as follows:

What problem is the KBDSS assisting with: *Improvement of on-site safety management of oil and gas drilling.*

Who are the intended users: *Oil and gas drilling site managers, construction officials, educators, health and safety officials;*

What user-inputs are required: *Details regarding the current safety measures taken on the drilling site;*

What is the primary statistical modelling used: *What-if scenarios through comparison of descriptive statistics;*

What output is provided to the user: *Hazardousness preventative actions that should be taken to reduce the possibility of death or major injury on-site.*

Key theme/s identified from examining this KBDSS: *The utility of an extensively researched knowledge base as the major building block for the software providing adequate decision-support without the need for complex statistical modelling.*

2.3.2 DAIRYPRO

DAIRYPRO was constructed to assist dairy farmers with the optimisation of resources to improve milk production for the given circumstances of their farm (Kerr et al., 1999a; Kerr et al., 1999b). The dairy farmer would input the following variables into the software, based on their farm's particular circumstances:

1) Farm details such as annual milk production (in litres), number and breed of cows, and farm area (in hectares);
2) Area in hectares dedicated to the pasture species;
3) Amount and type of concentrate (in tonnes) fed to the milking herd (yearly) in both regular feed and through wet matter;
4) Amount and type of fertiliser (in tonnes) applied to the farm (yearly);
5) Daily milk production for the average cow on specified months, including the amount of feed given to the cow during the same month;
6) Amount of nitrogen fertilizer (in tonnes) applied to the relevant pasture or crop;
7) Average percent of butterfat in the milk for the herd.

To determine which dairy farm factors, if any, could be altered to improve average milk production, two statistical models are utilised by DAIRYPRO, a rule-base using expert knowledge, and multiple regression modelling. Once the user has inputted the aforementioned variables for their given dairy farm, the following descriptive statistics are estimated by DAIRYPRO:

1) the average milk production across a region (RAP), which is calculated using multiple linear regression, and;
2) the achievable production (AP) for an individual farm (using expert rules of thumb).

Both descriptive statistics (i.e. RAP and AP) serve to act as an estimation of the expected milk production for a given farm under its current conditions; a prediction in the case of RAP, and a

comparison of pre-defined idealistic scenarios from the AP. Note that these rules of thumb were created based on a combination of extensive interviews with a dairy expert, examination of dairy management guidelines and numerous discussions with dairy farmers and dairy advisors. Both statistics are compared to the farm's actual average milk production and a series of what-if scenarios are presented to the user. Each scenario hypothetically alters one or more of the dairy farm's input variables to observe any changes in estimated milk production; these changes are provided to the user in profit or loss margins. The user can then determine which variables should be adjusted, and by how much, to achieve an optimal level of milk production. For example, if DAIRYPRO records the user is "…*feeding too much concentrate*" a suggestion to resolve the issue is made, "*Excessive pasture substitution is occurring*…" (Kerr et al., 1999a, p.253).

DAIRYPRO consists of two modules, the first – referred to as FARMPROD – is where the estimation of optimal dairy production (RAP and AP) and suggestions for the user occur. The second module, FARMDIAG, extends beyond general milk production, and assists the user with determining efficiency with specified feeding programs (for the dairy cows), such as a winter or summer program. Unlike the first module, FARMDIAG relies on expert rules of thumb alone to provide suggestions to the user. Figure 2.3 describes DAIRYPRO's base model, demonstrating the roles performed by each module. A concept shared by both DAIRYPRO (Kerr et al., 1999a; Kerr et al., 1999b) and HAZFO (Asad et al., 2019a), is the acknowledgement that a well-structured expert knowledge base can be crucial when constructing a KBDSS. The inclusion of expert knowledge serves to improve the software's decision-supporting ability. Also important is input from the end-users. As stated in Kerr et al. (1999a, p.254) "*The consultation process was successful with major changes to DAIRYPRO being suggested by farmers*.". These changes greatly enhanced the final product (the KBDSS), outlining the critical role experts and end-users can have during construction.

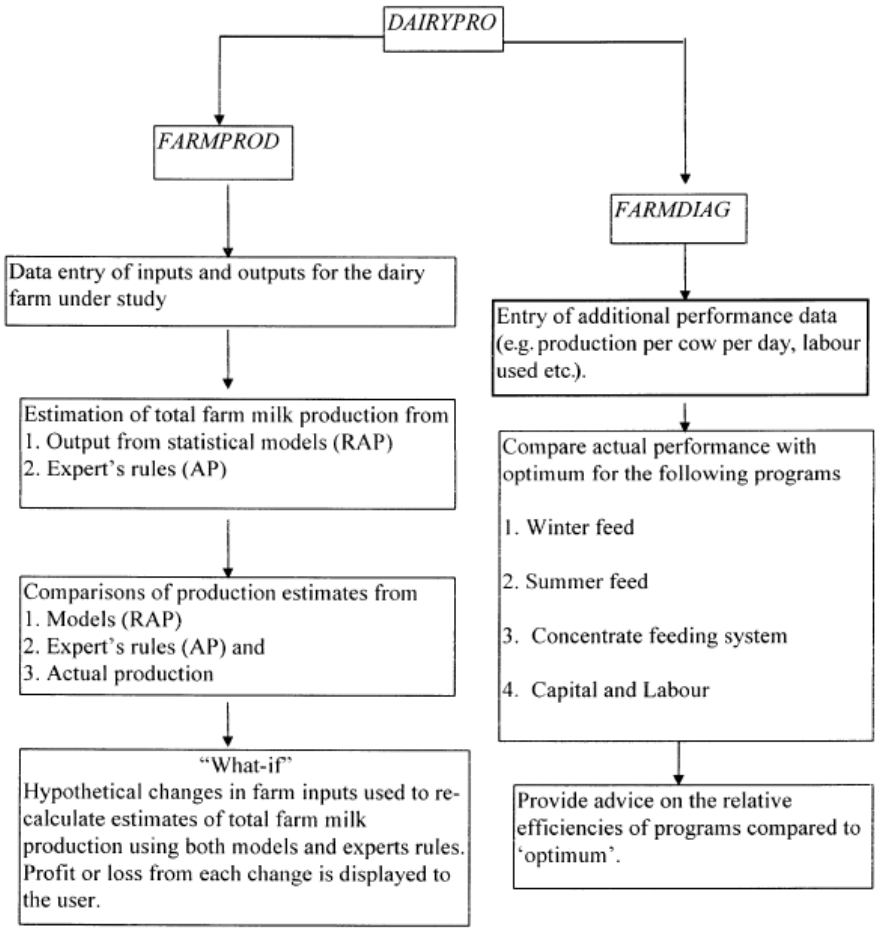


Figure 2.3: DAIRYPRO's Base Model
The DAIRYPRO KBDSS is comprised of two decision assisting modules: FARMPROD, which provides suggestions to the user on how farm conditions can be altered to improve the current milk production rate, and FARMDIAG, which provides suggestions on how the user can improve various feeding programs (taken from Kerr et al. (1999a, Figure 1, p.246)).

The development of the KBDSS DAIRYPRO can be summarised as follows:

What problem is the KBDSS assisting with: *Improvement of average milk production on dairy farms;*

Who are the intended users: *Dairy farmers, bank managers, loans officers, and farm consultants;*

What user-inputs are required: *Information regarding the conditions on the dairy farm, such as quality and quantity of feed and fertilizer;*

What is the primary statistical modelling used: *A combination of what-if scenarios created with expert rules of thumb and multiple linear regression modelling;*

What output is provided to the user: *Profit and loss margins to suggest where certain farm conditions can be altered to improve milk production.*

Key theme/s identified from examining this KBDSS: *Combining an extensive expert knowledge base with statistical modelling of available databases to provide decision-support based on both empirical experience and predictive algorithms. How the inclusion of a discipline expert in the early stages of a KBDSS's construction can identify new components to greatly improve the system's structure.*

2.3.3 PVSEL

During the construction of a pressure vessel (a large container designed to hold gases or liquids), appropriate material for its construction must be selected based on various criteria such as: strength of the material, temperature and pressure of intended use, corrosion resistance, hardness, and weld ability (Yurdakul et al., 2020). To assist with the decision-making process of material selection, the Pressure Vessel SELection (PVSEL) KBDSS draws from a database containing all feasible materials (Yurdakul et al., 2020).

The user first defines the required criteria needed for the selected material through a list of extensive user-inputs reflecting the desired specifications. Unlike the input phase of DAIRYPRO and HAZFO, PVSEL has an additional step where the user can weight each input based on the importance that this criterion be met. Following the user defining the required material specifications, PVSEL compiles an initial list of suitable materials that are contained within the software's internal database. Feasibility of materials to be included in this list is determined based on pre-defined specifications for each material that are known to PVSEL. For example, the user may outline that a required specification for their desired material is a working temperature range between –195°C and 360°C. Any materials within PVSEL's database that cannot adequately function within the user's specified temperature range are excluded from the list. The remaining materials are then rank ordered using three alternative multiple-criteria decision-analysis (MCDA) methods (TOPSIS, VIKOR, and ELECTRE – see Yurdakul et al. (2020, Appendix 1) for details). The independent rankings are then compared using Spearman's correlation and a decision for the best material is made. If the rankings disagree, an option to obtain a combined sum across the three rankings is used. The output of PVSEL is a comprehensive report outlining: (*i*) the user's inputted material specifications, (*ii*) the weighted criteria as defined by the user, (*iii*) a compiled list of feasible materials with their respective MCDA results, and (*iv*) a resulting table with the final materials recommended to the user.

The development of the KBDSS PVSEL can be summarised as follows:

What problem is the KBDSS assisting with: *Selecting an appropriate material during construction of pressure vessels;*

Who are the intended users: *Pressure vessel manufacturers;*

What user-inputs are required: *Criteria regarding the user's desired specifications of the environmental conditions to which the material will be exposed;*

What is the primary statistical modelling used: *Three MCDA analyses whose resulting rankings are compared using a Spearman rank correlation test;*

What output is provided to the user: *A defined list of feasible materials that fit the user's specifications (ranked by suitability);*

Key theme/s identified from examining this KBDSS: A*llowing the user to weigh which inputs have case-specific importance, providing the KBDSS with additional information that can allow the system to further clarify advice for a given scenario.*

2.3.4 APSIM

Developed to aid in improving crop production, the Agricultural Production Systems Simulation (APSIM) (see Holzworth et al. (2014) for the reference of its current iteration), is one of the most prominent KBDSSs in the agriculture industry within Australia, and since its creation has been utilised in New Zealand and the USA. APSIM's primary goal is to assist users in determining which farming strategies are optimal for improving crop production (Keating et al., 2003; McCown et al., 1995). Areas that gain the most benefit from APSIM are those with uncertain degrees of rainfall, or with no rainfall at all, and those where soil erosion and/or infertility threaten crop production.

APSIM contains an expansive internal database of ideal crop conditions for commonly grown crops, pastures and forests (and their interactions with the soil) in tropical and temperate areas throughout Australia. For every crop-type provided in the APSIM database, the following data is available: phenology, biomass, canopy, root system type, senescence pools, water, nitrogen and phosphorus levels. The user can then select the crop-type that will be grown and input details regarding the soil used. Crop ontogeny is then simulated based on the relationships between the crop-type database and soil details. These relationships are tested using several agriculture-based modelling techniques (Keating et al., 2003, Section 3). There are four stages of APSIM's process, which are presented as modules (Keating et al., 2003). The first is a biophysical module that simulates both the biological and physical processes that occur in farming. The second module is a management module which allows the user to specify their intended management rules which then characterise the scenario being simulated - these rules are implemented as "What-if" scenarios for subsequent simulations. Module

three provides various options to facilitate the data input and output from APSIM's simulations. The final module is a simulation engine which drives the simulation process.

After APSIM has analysed the inputs and simulated the various "What-if" scenarios, it provides one or more possible outputted suggestions to the user on how to improve crop production. An example of possible suggestions from APSIM include (Keating et al., 2003, p.276):

- Resetting individual module values.
- Reinitialising all data in modules to a given state.
- Sowing/harvesting/killing crops.
- Applications of fertiliser, irrigation or tillage to soil.
- Calculation of additional variables to track system state.
- Reporting of system state in response to events and/or conditional logic

The development of the KBDSS APSIM can be summarised as follows:

What problem is the KBDSS assisting with: *Improving crop production in areas with subpar rainfall and soil fertility conditions.*

Who are the intended users: *Farmers, farm developers and managers, agricultural department employees.*

What user-inputs are required: *Desired crop-type to be grown as well as multiple conditions related to the soil being utilised.*

What is the primary statistical modelling used: *Through simulation, multiple agriculture-based modelling techniques are utilised to test the relationships between the user's desired scenario and the internal database of observed data.*

What output is provided to the user: *Possible suggestions regarding what action the user should take to improve crop production, such as application of various treatments or the sowing of a crop altogether.*

Key theme/s identified from examining this KBDSS: *The importance of a strong database to represent ideal conditions when performing heavy simulation through "What-if" scenarios.*

2.4 Development and Evaluation of DNA-MAP, a KBDSS, for BGA Prediction

Based on an examination of the four KBDSSs examples described above, and a literature review of other systems (provided in Table 2.1), a generalised framework for DNA-MAP has been developed, a KBDSS designed to predict BGA for an unidentified person.

The initial stage of creating a KBDSS is identifying the complex problem within the user's discipline and recognising that it could be resolved through a decision-making framework. Identification of the problem was a common theme observed in the literature of KBDSSs (Table 2.1). For example, in the construction of HAZFO, Asad et al. (2019a) demonstrate through census data that the oil and gas drilling workforce suffer from hundreds of cases of serious injuries and deaths annually. To reduce the number of cases, safety measures needed to be improved on these drilling sites, and the authors proposed that one approach of doing this was to create a KBDSS which could assist site managers in determining where improvements could be made. In the context of BGA prediction, investigators are faced with the issue of having an unidentified person who could belong to one of several populations of interest.

As stated by Asad et al. (2019a), Kerr et al. (1999a), and Kerr et al. (1999b) having an expert from the discipline involved during the KBDSS's construction can greatly improve the system's efficiency. This involvement can occur through the utilisation of an extensive knowledge-base obtained from the expert to be used in the system's modelling phase, or through direct interaction with the user regarding key features of the system itself. The KBDSS's developers act as the bridge between the expert and other users, providing both the statistical modelling and graphical user interface (GUI). For DNA-MAP, the intended users are UWC-A, though its utility could extend to other forensic and military investigations. During the development of DNA-MAP Dr Kirsty Wright, a UWC-A forensic scientist with an extensive background in forensic biology and experience in identifying and assigning BGA to unknown remains, was consulted. These consultations took place in the form of informal and unstructured interviews regarding key features – such as inputs, outputs and error thresholds – that would be required for DNA-MAP. Evolutionary prototyping was demonstrated to numerous potential users in the discipline through presentations at conferences and informal discussions.

2.4.1 User Inputs

DNA-MAP's primary question of interest is "*what is the probability that an unidentified individual belongs to a given population based on their observed genetic profile and any prior information?*". To determine what user inputs would be relevant to DNA-MAP, it is necessary to first ascertain what

information is available to the user when performing BGA prediction. The genetic sources of information required when inferring ancestry are the observed genetic profile for an unknown person, together with estimates of the relative frequencies of seeing that profile for any specified populations of interest. Note that the genetic profile is obtained through a specifically created DNA panel containing multiple genetic markers chosen for their highly discriminating power at distinguishing between populations of interest. These two inputs act as the main source of information for BGA prediction, as ancestry is inferred based on which population the unknown person is most closely aligned to genetically. Several criteria regarding the genetic inputs need to be considered. These are: (*i*) which genetic marker should be included in the DNA panel and how many, (*ii*) the sample size available from each population of interest, (*iii*) the possibility of DNA degradation, and therefore, missing ancestry profile information, (*iv*) dependency between utilised markers, and (*vi*) the possibility of missing data from a variety of sources including stochastic processes. Phillips (2015) lists these criteria as required when creating an ancestry informative marker (AIM) panel, that is, a biological test that is created for the sole purpose of discriminating between two or more populations using highly discriminating sections of the human genome. Secondary to the genetic information, for certain cases the user will also have access to information regarding the probability of an unknown person belonging to a certain population prior to any genetic testing. Such information could be available through historical records or census data and will form a pertinent user input in conjunction with the genetic information. Lastly, certain inputs may become apparent during the establishment of a KBDSS's statistical modelling, such as the user's desired measure of achievable confidence. For the UWC-A (who rely almost solely on DNA to predict ancestry) an incorrect ancestry decision will result in an Australian soldier being laid to rest in a Japanese War shrine or a Japanese soldier being laid to rest with an Australian headstone in a Commonwealth War Cemetery, such errors need a zero threshold. How these various user inputs and criteria are addressed for DNA-MAP is discussed in Chapter 3, while Chapter 6 provides a detailed overview of DNA-MAP.

In addition to the user inputs, a key process that is required in the early stages of DNA-MAP's algorithm is the use of checkpoints, where multiple error checks and data cleansing functions take place. Within these checkpoints various internal checks and functions occur which aim to reduce the possibility of clerical errors in uploaded data files and examine whether the user has provided sufficient information required for accurate classification. These checks and functions are a typical feature of most KBDSSs.

2.4.2 Statistical Modelling

Methods for predicting BGA for a set of remains are not new to the literature, and previous studies have proposed various techniques from simplistic conditional multiplicative estimates to more complex clustering algorithms (Cheung et al., 2017; Cheung et al., 2018a; McNevin et al., 2013; Phillips 2015; Phillips et al., 2009). Ultimately, BGA prediction is a classification problem, that is, taking an unidentified individual and assigning them to a single population based on observed relationships and trends in the utilised populations. For this thesis DNA-MAP will only be concerned with binary classification, that is, inferring BGA for a set of remains where only two populations are possible, Japanese and Australian WWII soldier populations. Creating a KBDSS and predicting BGA are two complex tasks alone and starting with the simpler binary classification for the initial development is seen as a logical starting point to develop a proof of concept. Chapter 3 will examine and compare the various BGA prediction methods that have been utilised previously in the literature and propose alternative methods which may have successful application.

2.4.3 Desired Output

When DNA-MAP is used, there are several outputs that are pertinent to the user. These are: (*i*) the probability that the unknown person belongs to a given population of interest based on the user inputs, (*ii*) any assumptions and/or limitations made by DNA-MAP and the statistical modelling used, (*iii*) the level of confidence the user can have in the outputted probability of BGA, and (*iv*) relevant information that is obtained from the DNA-MAP's algorithm, for example, any error prompts or further suggestions to the user. These outputs are compiled into a BGA prediction report which is presented using clear, concise, English statements to assist the user with subsequent decision-making.

2.4.4 Evaluating a KBDSS

Evaluating DNA-MAP, namely, ensuring the system is working as intended within clearly identified software limitations is an important aspect of KBDSS development. Mysiak et al. (2005) understood the importance of evaluating a KBDSS and provided a list of priority considerations to be addressed when creating a system together with suggestion to how they can be addressed. Table 2.2 summarises the considerations suggested by Mysiak et al. (2005, Table 1, p. 205).

Table 2.2: Priority Considerations for Evaluating a KBDSS

A list of measurements to be considered when evaluating a KBDSS, adapted from Mysiak et al. (2005, Table 1, p. 205)

Subject of Validation	*Examples of Measurement*
KBDSS Development Process	1. Involvement of future users in early development phases; 2. Appropriately defined system requirements; 3. Evolutionary system development; 4. Clear definition of beneficiaries
KBDSS Components	1. Precision of models; 2. Quality of data; 3. User interface; 4. Reporting system to choose a suitable technology and management of data; 5. Complexity of DSS and data inputs.
Decision Process	1. Appropriateness of logical process followed when using DSS; 2. Number of alternatives explored by DSS; 3. Internal communication; 4. Correspondence to and appropriateness for decision organisation.
Decision output	1. Quantification profit/loss from DSS usage; 2. Consensus achieved among decision-makers; 3. Savings of time or other resources through DSS usage; 4. Contribution to organisational efficiency; 5. Consistency of solution
User satisfaction	1. Degree of confidence in results derived by DSS; 2. Acceptance (willingness to change current management methods); 3. Improvement of personal efficiency; 4. Correspondence of DSS output with decision-making style; 5. Users' understanding of implemented models

Integrating information from various sources and providing different pathways depending on which inputs are available (and on occasion, based on input values themselves) allows the decision support system to have more applications within its discipline of interest. Other benefits of utilising a KBDSS include *i*) being able to summarise unwieldly amounts of data into a single report tailored to the scenario of interest, *ii*) combining multiple statistical models and feedback checks into a single application, and *iii*) the automation of a complex and resource-consuming process (Pick, 2008; Pick and Weatherholt, 2013). As Pick (2008, p.719) states "…*automation of tedious tasks allows a*

decision maker to explore a problem more thoroughly than would be possible without the DSS". The limitations of a KBDSS are less distinct, namely, the limitations rarely focus on the concept of a KBDSS, but rather on poor implementation of the system itself. As González-Ferrer et al. (2018) and Zouri et al. (2019) discuss, the quality of a KBDSS is only as good as the quality of the data being inputted. The use of incorrect data or illogical justification of statistical models are just some key examples of how a KBDSS can be limited. Close collaboration with an expert in the discipline (preferably one who acts as an intended end-user) can improve the KBDSS's performance, as "*most DSS development problems result from poor identification of end users' needs*" (Pick and Weatherholt, 2013, p.9). Additionally, poor selection of statistical modelling can lead to poor performance if the techniques utilised by the KBDSS are inappropriate for the actual question of interest.

2.5 Concluding Statement

This chapter has drawn on several KBDSSs from the literature, to outline the key factors that should be generally considered during development of a KBDSS. In addition, a generalised overview of DNA-MAP was presented which described initial reasonings behind the system's inputs, modelling, and outputs. A forensic scientist considered to be an expert in the discipline representing the end-user (UWC-A) was consulted during these early stages of DNA-MAP's development and was able to provide feedback and suggestions based on their experience and knowledge.

Chapter 3 analyses the literature surrounding BGA prediction to determine how key factors in the methodology have been previously addressed. Chapter 6 provides a detailed discussion of DNA-MAP's algorithm, describing both the GUI and the underlying procedures.

Chapter 3 – Predicting BGA

3.1 What is Current Practice?

It is always paramount to have a rounded view of the literature to identify knowledge gaps and subsequently make a contribution towards filling these gaps. Previous studies have commented on three key factors to consider when attempting BGA prediction, these are: (*i*) adequate genetic representation of the populations of interest, (*ii*) sufficient genetic markers that provide informativeness (that is, clear genetic separation between the populations of interest), and (*iii*) an appropriate prediction algorithm (Cheung, et al., 2017, 2018a, 2018b; Phillips, 2015). However, as this chapter will discuss, there are additional factors beyond these that need to be considered to ensure accurate BGA prediction. Note that this thesis is not concerned with the development of a BGA prediction panel, rather, it assumes that all panels discussed and utilised in this thesis have already undergone extensive research. For a detailed guide on panel development, see examples such as Ghaiyed (2020) and Phillips (2015).

This chapter will be structured as follows: (*i*) a comparison of how BGA prediction has been utilised to date by commercial groups and for assisting forensic investigation, (*ii*) outlining the primary issues associated with BGA prediction observed in the literature and (*iii*) discussing how these issues have been previously addressed.

3.1.1 Commercial Groups

Ancestry analysis has become popular in recent years due to the commercialisation of ‘at-home’ ancestry kits; the popularity can be attributed to TV shows surrounding genealogical discoveries and massive TV/online marketing. It is reported that by the beginning of 2019, more than 26 million individuals had their DNA profile added to the database of the four leading commercial ancestry companies (Regalado, 2019). Different commercial kits offer a variety of results including: (*i*) information regarding heritage, (*ii*) connections to extended family, and (*iii*) susceptibility to ailments/disease which are known to be hereditary (Royal et al., 2010). With commercial companies boasting such large global databases, it suggests that the tests offered are accompanied with high precision and accuracy, but how true is this statement? When an individual obtains a result from a commercial test, no mention is made regarding any margin of error or variability. Furthermore, there have been instances of individuals who have submitted an identical DNA sample to several different commercial ancestry companies, only to receive different results from each test (Letzer, 2018). In 2018, one commercial ancestry company, “Orig3n”, faced allegations of falsifying genetic results after failing to detect that one sample submitted for testing belonged to a golden retriever (Griffith,

2019). The commercial company reported that the "subject" had a higher-than-average muscle mass and a cardio output that would be suitable for high-intensity activities such as boxing and cycling. The same canine sample was submitted to multiple other commercial ancestry companies, all of which correctly rejected the case, recognising the sample to belong to non-primate DNA.

The reason an individual may obtain a different result between two commercial tests is that the kits and analytical processes used differ between companies, being treated as "trade-secrets". Rarely is the development and validation process made public, making it difficult for investigators and scientists to determine the accuracy of such panels. Thus, the public is unaware of factors such as how these tests were designed and validated, and which algorithms are utilised in their analysis. In addition, the consumer is not always aware of the level of confidence that can be assigned to obtained results. While some commercial companies do provide a measure of confidence, for example, in the case of 23andMe the user can adjust this measure to observe the effect it has on the results, this feature is not guaranteed. As the answers to these questions are not readily available, these tests are unsuitable for use in forensic casework. Despite each commercial company having access to their own large global database, there is no sharing of data between companies. For example, one commercial company may have an extensive sub-database of individuals from the Middle East but lacks a comprehensive sample of individuals from Oceania; therefore, certain company kits have a reduced accuracy for certain global regions. These aspects of the kits are unavailable to the consumer.

The type of output that the consumer receives when using a commercial kit depends on what tests the company offers. These outputs may include autosomal, mitochondrial and Y-chromosomal ancestry testing and assigned percentages of ancestry contributions, phenotypic information, possible proneness to medical ailments and other traits. However, the consumer should be cautious as rarely is the error rate associated with these outputs provided.

3.1.2 Forensic Case Work

BGA prediction has multiple applications within the forensic discipline, with one example being to supplement eyewitness reports. Eyewitness reports can form the preliminary stage in criminal investigations, either leading investigators towards suspects, or excluding suspects (Marano and Fridman, 2019; Phillips, 2015). However, these reports are subject to several limitations including lighting conditions, cognitive bias, personal trauma, and memory distortion (Cheung et al., 2018b). The ability to predict BGA, and other externally visible characteristics such as hair and eye colour "*…provides opportunities to strengthen eyewitness accounts or in their absence, gain information about a suspect*" (Phillips, 2015, p. 49). Other primary applications of BGA prediction include

counter-terrorism, disaster victim identification, cold case investigation, missing persons (Phillips, 2015), archaeology, ancient DNA analysis (Bongers et al., 2020; Harvard Medical School, 2019; Slatkin, 2016; Wright et al., 2018), and historical military remains.

Unlike the commercial application of BGA prediction, information regarding the data and techniques used in forensic investigation is more readily available. To ensure that BGA prediction is as close to the standard of "evidence" as possible, relevant studies should typically provide: (*i*) a clear description of utilised data (often providing the data itself as a supplementary file), (*ii*) details of the classification algorithm used (where possible, links are provided for utilised software/packages), and (*iii*) results from any validation experiments. Despite all the information being accessible, there is still an apparent lack of unification across BGA prediction studies regarding statistical standards (such as sample size) and the methods utilised for inferring a population of origin.

The literature surrounding BGA prediction can be classified into two primary categories: (*i*) those concerned with the creation and validation of panels consisting of ancestry informative markers for discriminating populations (group-level classification), and (*ii*) validating a classifier's accuracy through the comparison of the predicted ancestry to an individual's declared ancestry (individual-level classification). Research related to the former is concerned with determining which sections of the human genome are suitable for distinguishing between two or more populations of interest. When selecting genetic markers, criteria of interest typically include differential variant frequencies between populations, independence between selected markers, the number of necessary markers, and other assumptions about genetic structure (Phillips, 2015). In the latter research category, interest lies in utilising patterns observed in the genetic markers to create statistical models which can infer a population of origin to an unknown individual with an associated measure of accuracy. Criteria typically of interest should include achievable accuracy, precision, ease of calculation/performance, and validity of assumptions. The focus of this thesis, and of the accompanying KBDSS, falls into the latter of these two research categories.

A similar area to BGA prediction is the research performed in archaeology and ancient DNA analysis (Bongers et al., 2020; Harvard Medical School, 2019; Slatkin, 2016; Wright et al., 2018). Extra components in these studies are the concepts of time and evolution. As opposed to BGA prediction, where interest lies in distinguishing between two or more divergent populations, the archaeologic and ancient DNA analysis studies trace the various evolutionary populations through history that culminated into a given convergent population. These ancient DNA analysis studies will often employ group-level classification methods that are also utilised in the BGA prediction studies, such

as Principle Component Analysis and the *F*-statistic. For a compiled list of classifiers that are employed in ancient DNA studies, the reader is referred to Slatkin (2016). As these studies are concerned with group-level classification, they will not be examined in further detail in this thesis, with exceptions being made where applicable methodology was observed.

Various statistical classifiers that have previously been reported in the literature will be examined to determine which, if any, are suitable to integrate into the KBDSS. Based on an analysis of the relevant literature, the following is a list of key factors that need consideration when selecting a BGA prediction method.

1) Admixture – Naturally occurring mixture between populations in recent generations, can occur at a population-level of a family-level;
2) Parsimony – Estimating the minimum amount of information needed;
3) Classifiers – Determining which classification model should be utilised;
4) Prior probability – Accounting for the probability of an individual having a higher chance of belonging to a specific population prior to any genotyping;
5) Relevant populations – Ensuring relevant and accurate populations are selected for a given scenario;
6) Sample size & Rare Event – Determining the appropriate sample size to ensure an accurate representation of the population and including measures to allow for the possibility of a rare but unseen event;
7) Degraded samples/partial profiles – How well the utilised classifier can handle missing data while still outputting accurate classification;
8) Margin of error – Applying a measure of reliability to the estimated probabilities of ancestry.

Many factors that have been previously discussed in the literature (Cheung et al., 2017; Phillips, 2015), but currently no standardised method of BGA prediction has been adopted by the forensic science community. This is a prominent gap in the global forensic science literature, as BGA prediction is not included in prominent forensic DNA analysis and interpretation guidelines or reports by groups such as:

- President's Council of Advisors on Science and Technology (PCAST) Report (President's Council of Advisors on Science and Technology, 2016);
- Scientific Working Group on DNA Analysis Methods (SWGDAM);
- European Network of Forensic Science Institutes (ENFSI) (Willis et al., 2015).

Each factor will be explored in detail in following sections. Note that while BGA prediction can be performed using a variety of genetic measures (including mitochondrial DNA analysis), the measure of interest in this thesis is autosomal single nucleotide polymorphisms (SNPs)

3.2 Factors to be Addressed for Ancestry Prediction

3.2.1 Admixture

A common limitation that can arise during ancestry analysis is the concept of "admixture". In this thesis, only recent family-level admixture is of interest to UWC-A casework, that is the mixture of two populations in recent generations resulting in an offspring with non-homogenous ancestry. For example, consider a child that is the offspring of an individual from Hungary and an individual from Poland. An admixed individual, this child would likely have segments of DNA which could be linked to each parental population. For an investigator to go a step further and assign BGA to the admixed child, certain complications will arise. Depending on which segments of DNA are analysed, the investigator may only observe ancestry indicative of one parental population, while the other population goes unobserved. By chance, an individual containing admixture from Poland and Hungary may be genetically similar to a typical individual from Slovakia (located between the two countries), which could lead to a potential misclassification (Cheung et al., 2018a).

When admixture occurs, there is the possibility of individuals occurring in a population who may deviate from what is considered "genetically-typical" of that population. This can lead to misclassification. It is important to note there is also the possibility that random mutations may occur, resulting in an individual's DNA being indicative of the wrong BGA. In current available data, there is little opportunity to obtain data with known levels of admixture, and there are no readily available datasets for admixed Australians and Japanese individuals. Therefore, in situations where no admixed data is available, one approach is to simulate individuals with known levels of admixture by using known non-admixed individuals as ancestors. This approach was previously utilised in Cheung et al. (2018a). These admixed samples are then classified using the employed classification model, to determine how well the current practice can estimate the true percentages of original contributing populations. The value of using such a simulation is that "*it is worthwhile to gain knowledge of the admixture profile of a population sample, even though this is highly variable*" (Phillips, 2015, p.60). The ability of current BGA prediction models to correctly detect and resolve admixture will be discussed in Section 3.2.3.

3.2.2 Parsimony

The principle of parsimony in statistics refers to a statistical model or theory that utilises as few parameters as possible, makes use of linear models as opposed to non-linear models, relies on as few assumptions as possible, and provides simple explanations (Crawley, 2012). There is a belief that when performing ancestry predictions, the greater the number of markers used, the higher the achievable accuracy (Pardo-Seco et al., 2014). Cheung et al. (2019), however, advocate that this has yet to be proven, and that it may be advantageous to utilise smaller, refined panels consisting of highly efficient markers to reduce cost. Note that with the advancements in current SNP panel analyses the costing factor is becoming less of an issue, with current technology providing the ability to sequence up to a million markers at a time (LaFramboise, 2009). Therefore, the primary objective when selecting which SNPs to include and exclude from a panel should be based on discrimination power.

To demonstrate the utility of parsimony, consider a set of *n* ancestry informative markers which have been selected to distinguish between two populations and have been ranked according to discrimination power. In this context, a high discrimination power is defined as a genetic marker with a variant that is observed in a high proportion in one population, while either low or absent in the other. Using the marker with the highest discrimination power will yield discrimination power, dp_1; the addition of each marker will increase that power by some measure resulting in $dp_1, dp_2, \ldots, dp_n$. Eventually, the discrimination power will plateau, and a point will be reached where effectively complete discrimination is achieved; the addition of further markers will become unnecessary, that is, no additional information will be gained (Tal and Tran, 2018). However, while no information may be gained, there is the possibility that the addition of further markers may introduce increased noise. This noise, in the context of BGA prediction, refers to the two associated errors that may be present for each marker, these are:

i) sampling error: the probability that an error has occurred because the sample size collected was not large enough to accurately represent the true population;

ii) classification error: the probability that the classification algorithm has misclassified an individual for that given marker.

Unlike discrimination power, these errors do not plateau and will continually increase with the addition of more markers, that is, the error will propagate through the system. However, the concept of noise and its role in panel and model development is complex. If the inclusion of every marker subsequently increased the total noise of a panel, conclusions could be made that the use of a single

SNP is the best solution as it introduces the least amount of noise. Even if a SNP were truly fixed in one population and absent in another, the use of a single SNP is not practical due to the possibilities of random mutation and degradation. Therefore, multiple markers are required to reduce the impact of these possibilities and a balance must be achieved between the total number of markers and their collective noise. By assuming the noise introduced by each marker in not equal, one approach could be to categorically weight the markers, for example, homozygote markers with greater disparity in allele probabilities between populations would theoretically introduce less noise in comparison to heterozygote markers. Alternatively, the concept of noise could be treated as the by-product of sample size, where if a large enough sample size was collected the noise would be diminished by the strong data. Therefore, noise can be treated analogously to discrimination power in conjunction with the perceived confidence of the available data. To demonstrate this relationship, consider a SNP marker panel consisting of 1000 SNPs, whose allele frequencies – based on a sample of $n = 20$ from each population – are situated around a 60%/40% disparity. Based on these frequencies having a low discrimination power, further substantiated by the small sample used to obtain them, the possibility of misclassification is likely. Replace this panel with one consisting of only 20 SNPs, but whose frequencies, now based off a sample of $n = 300$ from each population, are close to fixed ($\approx 95\%/5\%$). There is greater confidence in this smaller panel as the discrimination power per SNP is higher and is based on a larger sample size, despite there being significantly fewer SNPs available. Therefore, to combat the "noise" of a panel, it is important to only include SNPs that have an acceptable level of discrimination power, noting that sample size also play an important role in a SNP's discrimination power.

A parsimonious classifier will utilise only the number of markers required for the best classification that set of variables can achieve, while minimising the possible error that occurs with the inclusion of additional markers. One approach to developing a parsimonious model is to utilise a concept known as information theory. Information theory can be described as the theory describing the process of "…*reproducing at one point exactly or approximately a message selected at another point*" (Shannon, 1948, p.379), namely, recreating a sequence of symbols which may otherwise be obscured due to interference. The earliest application of what would eventually be known as information theory can be traced back to the introduction of Morse code (Beechey, 1876); where frequently used letters, such as "E", were formatted to be transmitted more quickly than uncommonly used letters, such as "J". Information theory has since been utilised in numerous disciplines, with a prominent example being the work of Alan Turing during WWII, who used information theory in cryptanalysis to decode the complex German "Enigma" cipher (Good, 1979), providing key

intelligence to Allied forces. In BGA the aim is to find the minimum number of genetic markers required to accurately convey the underlying ancestry, that is, to match an individual's perceived BGA with their true genetic BGA.

The pivotal moment in information theory's history was its official establishment as a discipline in Claude Shannon's publication (Shannon, 1948). Shannon's information theory of communication was invented during the creation of the communication network, as people began attempting to send messages between continents. The limitation faced at the time was that the message could not travel big distances without so much distortion and weakening that it was not discernible by the receiver. The issue was "*noise*", a stochastic, natural phenomena which destroys/masks parts of a message. To demonstrate the logic behind Shannon's theory, and why noise is causing problems, consider a generalised communication system (Figure 3.1).

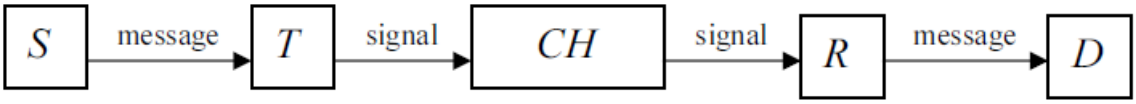


Figure 3.1: Communication System
Flowchart diagram of generalised communication system, taken from Lombardi et al. (2016, p.3).

Where:

i. *S* is the source, which generates the initial message;
ii. *T* is the transmitter, which converts the message into a format (signal) to be transmitted, if encoding is required in the system, it occurs here;
iii. *CH* is the channel, namely, a medium/device where the transmitted signal is carried from sender to receiver;
iv. *R* is the receiver, if encoding was performed at the transmitter then the message is decoded;
v. *D* is the destination where the message is received.

At the source a message is constructed, *S*, which consists of a set of individual states, $s_1, \ldots, s_n$, usually termed *letters*. An example of this form of message are the codons, the array of bases (A, C, G, U), which encode and define the amino acids, as shown in Figure 3.2. The message (the amino acid itself) is the totality of the codon, for example, "Valine" (Val, Column 1, Row 4) can be encoded either by "*GUU*", "*GUC*", "*GUA*" or "*GUG*".

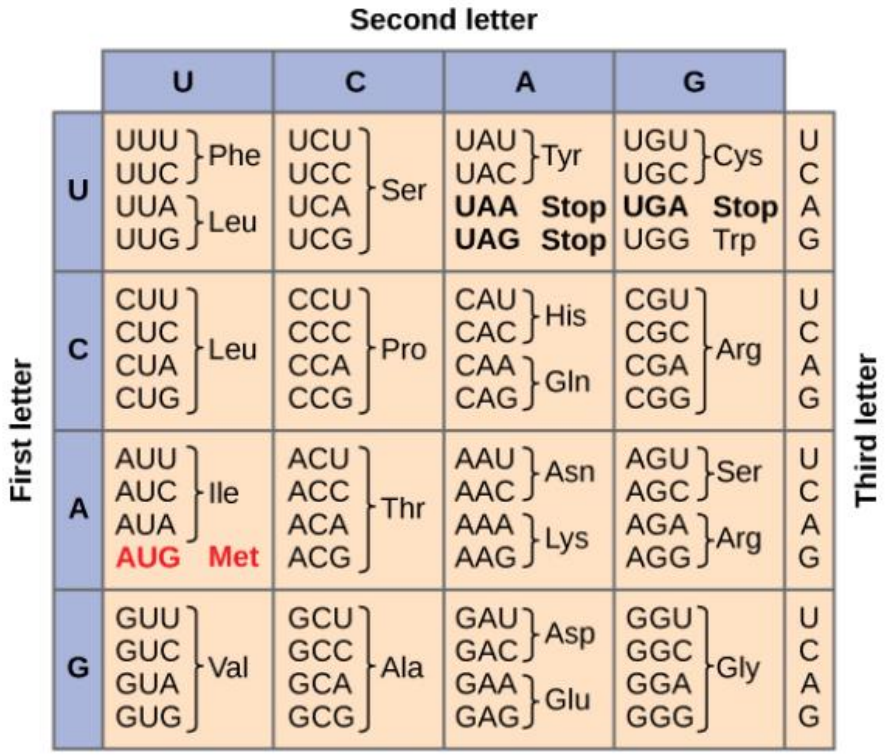


Figure 3.2: List of Amino Acids
Codon chart for amino acid classification, taken from Openstax (n.d).

Another example that people experience every day unknowingly, and the initial problem that Shannon faced back in the 1900s, is the sending of an online message through a communication system, such as a voice message of a phone conversation or the text message in an email. In these examples, every letter on an electrical keyboard has an underlying binary code unique to itself, as dictated by the American Standard Coding for Information Interchange (ASCII). For example, the message "*A*" is encoded by the binary sequence "*1000001*". During transmission through the channel the message can be degraded due to noise, which may cause the received message to be missing parts of the original message, for example, "*1?00?0?*" where "?" is the missing information. To apply the same ASCII classification system to the degraded message, there are now multiple possible outcomes as the receiver is uncertain what information was lost due to noise. (Figure 3.3).

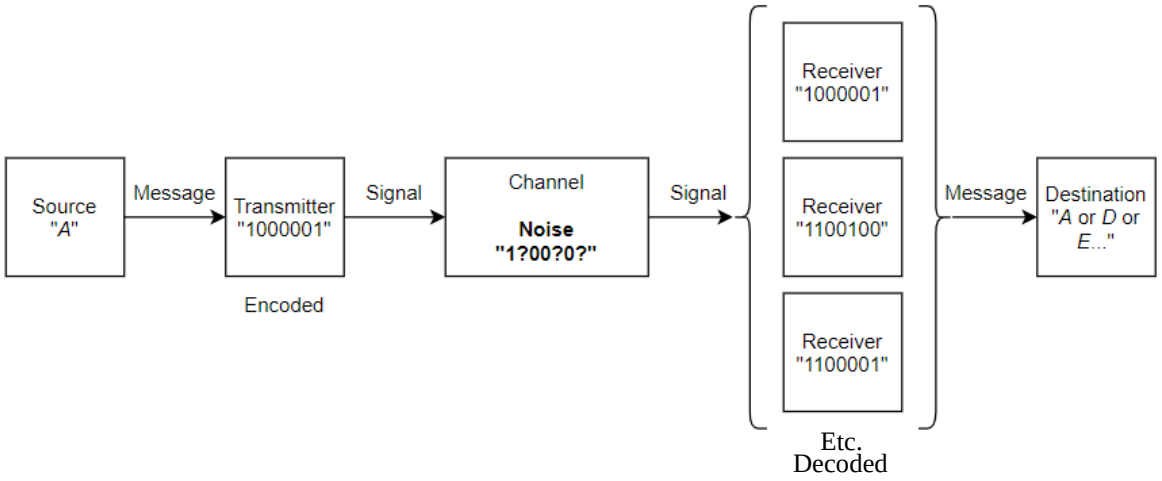


Figure 3.3: Updated Communication System
Example of the effect noise has on message communication, note that the three decoded examples shown are only some of the possibilities but there are many more that are feasible.

Noise can be considered as uncertainty, where the message prior to transmission had a high degree of *certainty* and a low degree of *uncertainty* (as the message was known); but the message received had a low degree of *certainty* and a high degree of *uncertainty* (there were numerous possible solutions). Shannon conceived this ratio of certainty versus uncertainty as information where the more information a message had, the greater degree of certainty one could have in the received message. If p_i is the probability of the *i*th letter/state, then under Shannon's definition, the amount of information this letter provides to the message is equal to:

$$\log\left(\frac{1}{p_i}\right)$$

Which can be simplified to:

$$-\log\left(p_i\right) \qquad (3.1)$$

For equiprobable states, information is equal to $1/p$. As Shannon explains, the use of the logarithmic scale has the useful properties: *i*) scaling large, unwieldy numbers to a manageable scale, and *ii*) practicality, since parameters of interest in communication (time and bandwidth) scale linearly with the logarithmic scale (Shanon, 1948). Selecting which logarithmic base to use should be based on the unit of measurement, for example, a binary unit of two likely alternatives should utilise $\log_2 X$ (termed by Shannon, a *bit*). Since the source will typically produce a sequence of letters/states to comprise a

message, the average amount of information produced at the source, $H(S)$, termed Shannon's entropy (due to relations to the terminology used in thermodynamics (Lombardi et al., 2016)), equals:

$$H(S) = -\sum_{i=1}^{n} p(s_i) \times \log\left(p(s_i)\right) \tag{3.2}$$

Where $p(s_i)$ is the probability of the s_ith letter/state, and s refers to the individual symbols which construct the complete message set S.

To demonstrate how information theory is applicable for BGA prediction, consider the original source message as the complete set of SNPs used in the panel, S, which will be used to estimate an individual's ancestry. This message is encoded by the series of genotypes, acting as the s_ith state, observed from the utilised SNPs. Sections of the message may be omitted due to degradation or stochastic errors, causing some SNPs to be unavailable for analysis. The question of interest, therefore, is how many of the available SNPs need to be analysed to accurately decode the message, which is required to predict the individual's ancestry.

An important aspect of information theory is determining the minimum amount of information required for accurate message decryption. In any typical communications system, there is the possibility of error occurring at each given letter/state within a message. Note, that "communications system" here is any given system in which a message comprised of individual symbols is encrypted and decrypted across some form of channel/medium (such as the amino acid cypher or the SNP panel). It is possible that a given symbol within a message is misread, or some alternative stochastic influence occurs, which introduces error(s) into the final message. Assuming, the chance of such an error occurring is independent for each symbol, then the addition of each symbol would increase the possibility of error in a linear manner infinitely. Conversely, the addition of symbols to increase information rarely occurs linearly. The balance of information gain makes information theory a valuable tool for ancestry analysis, and thus for finding the minimum number of SNPs required to predict ancestry. DNA analysis (generally speaking) is an expensive task, with one of the criteria affecting the total cost being the number of SNPs being analysed. Determining the minimum number of SNPs required to accurately predict ancestry provides a cost-benefit option to the forensic scientist, outlining which SNPs available are providing the highest amount of information. Determining the minimum number of SNPs required is also beneficial for degraded samples, it is highly unlikely for a degraded sample to have all SNPs in a panel available.

Information theory has been utilised previously in ancestry analysis, where the theory has been applied when creating an ancestry panel and estimating the minimum number of SNPs needed for accurate distinction between populations (Rosenberg et al., 2003; Tal and Tran, 2018). Tal and Tran (2018) present their approach for measuring informativeness for a set of DNA markers derived from Shannon's information theory and formulate it into a Bayes classifier (that is, incorporating population priors into the calculations). Rosenberg et al. (2003) made use of Shannon's theory to create a measure of an ancestry panel's informativeness based on two models: *i*) the no-admixture model where individuals were assumed to only originate from one of *K* populations (where *K* is the number of populations of interest), and *ii*) an admixture model, where coefficients were estimated for an individual's proportion of association to each population of interest. Alternatives to informatic-theoretic methods are also utilised for measuring a panel's "informativeness". These tend to be simplistic algebraic methods based on comparing allele frequencies without considering interactions between DNA markers and include the absolute allele frequency difference (δ) between two populations of interest and the Fixation Index (F_{st}) (where $F_{st} \approx \delta/(2-\delta)$) (Phillips, 2015; Rosenberg et al., 2003). The focus of this thesis is on creating a classification system using an already predefined set of SNPs, these information-theoretic and algebraic techniques which aim at estimating the number of SNPs needed in a panel will not be discussed in further detail in this thesis. However, the methodology introduced will draw on information theory and will be discussed as presented.

3.2.3 Classifiers

As Cheung et al. (2017, p.902) state "*a good classifier should be able to accurately predict BGA under a number of conditions*". These conditions are that *i*) non-admixed individuals should ideally be assigned wholly to a given population, and *ii*) admixed individuals should have BGA proportions that reflect the relative contributions of the appropriate populations. This section will compare and discuss four classifiers that have previously been utilised for BGA prediction, together with a classifier proposed in this thesis, outlining the benefits and limitations of each method. The classifiers included here are: STRUCTURE, Generic Bayesian, Genetic Distance Algorithms (GDA), Multinomial Logistic Regression (MLR), and Logistic Model Tree (LMT).

STRUCTURE. This is a program which utilises a Bayesian cluster algorithm to infer population structures within a dataset using observed patterns in genotype data (Pritchard et al., 2009, available at https://web.stanford.edu/group/pritchardlab/structure.html). The program was originally developed by Pritchard et al. (2000) as a technique for identifying and separating populations based on genetic structures, with later extensions made by Falush et al. (2003) and Falush et al. (2007). Current applications for STRUCTURE include: (*i*) demonstrating the presence of population

structure, (*ii*) identifying distinct populations and sub-populations, (*iii*) assigning individuals to possible population origins, and (*iv*) the identification of admixture (Pritchard et al. 2009, p.3). To date, STRUCTURE has been considered the gold standard of BGA prediction tools (Cheung et al., 2018b; Phillips, 2015) and has been utilised in several BGA prediction studies. A key example of STRUCTURE being applied to BGA prediction is the investigation following the 11-M Madrid commuter train bombing as documented in Phillips et al. (2009). Where STRUCTURE's performance is compared with an alternative classifier, the Generic Bayesian approach. In the investigation, interest lay in determining whether several biological samples, believed to belong to the perpetrators, had a Spanish or Moroccan origin; these populations were selected based on case information. Figure 3.4 (sourced from Phillips et al. (2009, Figure 1, p.4)) shows a standard STRUCTURE output, and demonstrates how the clustering method can be a means to visualise the distinction between the two populations, Moroccan (n = 48) and Spanish (n = 48). In addition to these two groups, seven case samples are shown on the right, and, based on metrics utilised by STRUCTURE, it can be determined which of the two groups the samples are more closely related towards. An important factor when using STRUCTURE is the need for the user to provide a value for *K*, the number of population groups STRUCTURE is to use in its calculations. Ensuring *K* accurately reflects the number of groups present in the dataset is important, as STRUCTURE will attempt to create *K* clusters, therefore, if the number of groups in the dataset differs from the assumed *K*, the model will be inaccurate (Porras-Hurtado et al., 2013; Pritchard et al., 2009). Providing a value for *K* can be a limitation in certain scenarios, as the exact number of population groups in a dataset is not always clear, requiring the user to run multiple iterations of the analysis under several values of *K*, if *K* is unknown. The *K* value which fits the user's beliefs the best is then selected, introducing a level of bias.

Figure 3.4: Phillips et al. (2009) STRUCTURE Output
Each vertical strip represents a single individual within each of the relevant sets (Moroccan, n = 48; Spanish, n = 48), where K = 2 (Sourced from Phillips et al. (2009, Figure 1, p.4)).

To measure the extent of how *K* affects the outputted models, Kalinowski (2011) used STRUCTURE on simulated populations under several values of *K*. Kalinowski (2011) observed that STRUCTURE would (*i*) frequently create incorrect clusters when *K* was not representative of the true populations

present and (*ii*) be highly influenced by variation in sample size (Kalinowski, 2011; Kidd et al., 2011). This is supported by a review by Cheung et al. (2017), who found inferring a value of *K* to be subjective. To avoid the possibility of incorrect clusters based on subjective *K* values, both Cheung et al. (2018b) and Kalinowski (2011) suggest the use of repeated STRUCTURE runs using several values of *K*. However, this approach introduces high run-times and is not robust. In addition to the subjectivity of *K*, other limitations associated with using STRUCTURE include (Cheung et al., 2017, 2018b; Kalinowski, 2011; McNevin et al., 2013):

i) High run-time;
ii) Silent crashes;
iii) Strict formatted input files;
iv) Model assumptions of independence regarding genetic structure.

Of particular importance is the observation made by Cheung et al. (2017) and McNevin et al. (2013) that STRUCTURE assumes Hardy-Weinberg Equilibrium when inferring population clusters. These authors then comment that ancestry markers which have been subjected to selection criteria, that is, handpicked for their ability to discriminate populations, are less likely to be in equilibrium; caution should be exercised as the predictions may be inappropriate when this assumption is invalid.

An additional aspect of STRUCTURE is that the program provides the user the option to use an admixed model or a non-admixed model. The user can select one of the two models based on whether the user believes that the individuals originate purely from one of the *K* populations (non-admixed) or that the individuals may have a mixed ancestry (admixed) (Pritchard et al., 2009). The STRUCTURE user manual recommends starting with the admixture model as it is a "*reasonably flexible model for dealing with many of the complexities of a real population*" (Pritchard et al., 2009, p.7), and that admixture is a common feature of real data which, is unlikely to be detected in the no-admixture model.

While STRUCTURE may have several drawbacks that make the program complex, requiring significant time to both learn and use the application, it is not without benefits. Comparisons of STRUCTURE and a number of other classifiers (Generic Bayesian, GDA, MLR), are provided by Cheung et al. (2017) and Cheung et al. (2018a), for situations involving non-admixed and admixed individuals, respectively. For the non-admixture situation, Cheung et al. (2017) used a training dataset of 1093 individuals from four populations collected from the 1000 Genomes Project (Genomes Project Consortium, 2015), namely, Africa (n = 246), Europe (n = 380), East Asia (n = 286), and America (n = 181). They used the results to classify a test dataset of 516 individuals collected from

the CEPH Human Genome Diversity Panel (Li et al., 2008), namely, Africa (n = 95), Europe (n = 150), East Asia (n = 210), and America (n = 61).

Figure 3.5 (Cheung et al., 2017, Online Resource 8) shows the output from their STRUCTURE analysis, where K = 4. While the European and East Asian training sets are highly homogenous, both the African and American training sets exhibit noticeable admixture. Regardless, STRUCTURE was able to assign all test subjects with the highest classification accuracy compared to the other classifiers. Classification accuracy was inferred by comparing the test individual's predicted ancestry to the original self-declared accuracy and utilising Area Under the Receiver Operating Characteristic (AUROC) Curve values.

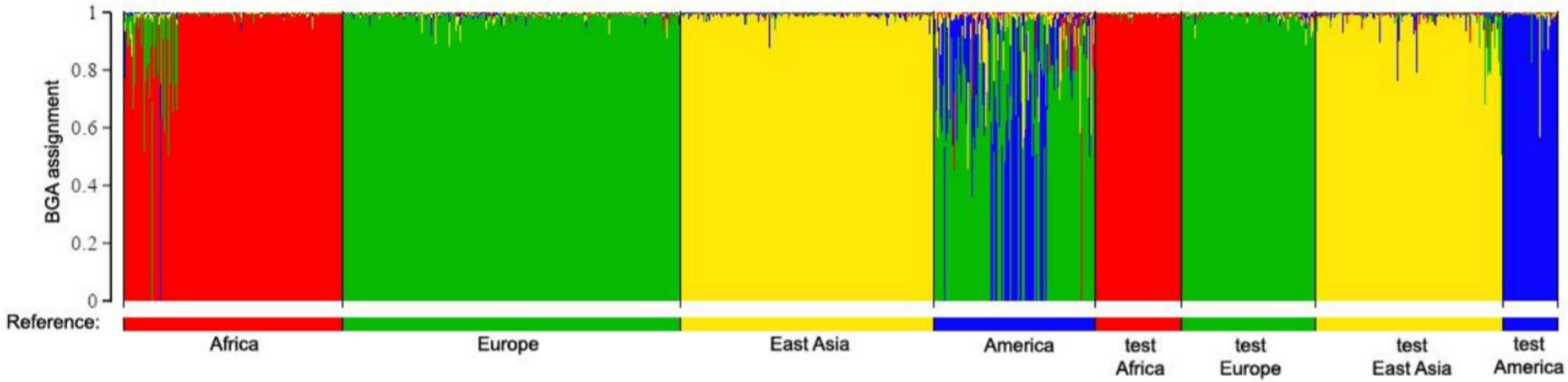


Figure 3.5: Cheung et al. (2017) STRUCTURE Output
Each vertical strip represents a single individual within each of the relevant sets where K = 4 (Sourced from Cheung et al. (2017, Online Resource 8)).

To test STRUCTURE's ability to handle admixed samples, Cheung et al. (2018a) simulated individuals by selecting samples from the aforementioned four training sets that were unambiguously non-admixed based on self-reporting, resulting in four 'non-admixed' samples of African (n = 66), European (n = 64), East Asian (n = 23), and Amerindian (American) (n = 5). Using these non-admixed individuals as initial ancestors, third generation offspring were produced under a total of 35 scenarios as given in Table 3.1 (Cheung et al., 2018a, Table 1, p.106)

Table 3.1: Cheung et al. (2018a) Admixture Scenarios
35 scenarios that Cheung et al. (2018a, Table 1, p.106) used to simulate individuals of varying admixture.

Combination	African	European	East Asian	Amerindian	Number of simulated individuals
1	100%				23
2		100%			23
3			100%		23
4				100%	5
5	50%	50%			23
6	50%		50%		23
7	50%			50%	10
8		50%	50%		23
9		50%		50%	10
10			50%	50%	10
11	75%	25%			23
12	75%		25%		23
13	75%			25%	10
14	25%	75%			23
15		75%	25%		23
16		75%		25%	10
17	25%		75%		23
18		25%	75%		23
19			75%	25%	10
20	25%			75%	10
21		25%		75%	10
22			25%	75%	10
23	50%	25%	25%		23
24	50%	25%		25%	10
25	50%		25%	25%	10
26	25%	50%	25%		23
27	25%	50%		25%	10
28		50%	25%	25%	10
29	25%	25%	50%		23
30	25%		50%	25%	10
31		25%	50%	25%	10
32	25%	25%		50%	10
33	25%		25%	50%	10
34		25%	25%	50%	10
35	25%	25%	25%	25%	30

Classification accuracy for admixed individuals was determined by comparing the predicted relative contributions for each population to the known scenario. For example, ideally an individual created from scenario 5 (50% African and 50% European, Table 3.1) should have a STRUCTURE output with all individuals having approximately the same proportions of membership to the two

populations. Cheung et al. (2018a) observed that STRUCTURE performed well for "simple" admixed scenarios, that is, those comprised of a mixture of two populations. However, the introduction of a third population caused a noticeable reduction in STRUCTURE's ability to accurately infer contributing populations. The validity of these comparisons is, however, questionable, as the sample sizes used to simulate individuals were small. Consider, especially, the Amerindian individuals who were simulated based on allele frequencies from a sample size of five, which was sub-sampled from a previous sample of 61. There is a possibility that these five individuals, or even the original sample of 61, is of inadequate size to accurately reflect the true population. In conjunction with the sample size used to create the simulations, the outputted number of simulated individuals was also small, with no scenario shown in Table 3.1 exceeding 30 individuals.

Generic Bayesian. The Generic Bayesian approach combines information regarding genotype/allele frequencies for reference populations to update a prior probability of belonging to a given population. Note that the Generic Bayesian is also referred to as the Naïve Bayesian. In the statistical literature, the generalised version of this technique is referred to as Bayes' theorem of conditional probabilities. There are two components to the Generic Bayesian method as used for forensic DNA situations: (*i*) the likelihood ratio (LR) comparing an individual's observed genotype/allele frequencies in one population to those in a different population, and (*ii*) the prior odds ratio which is the ratio of the probabilities of an individual belonging to a population before any genetic testing has occurred. Methodology for calculating the Generic Bayesian is shown and discussed in Section 4.4, while the effects of the prior odds ratio and how to estimate its value are discussed in Section 3.2.6.

The Generic Bayesian approach was adopted to BGA prediction following its previous uses in the forensic science, such as its application in paternity testing as demonstrated by Essen-Möller (1938) and its later application to criminal case work. For the latter, the method was used when DNA evidence was presented in the justice system to compare two mutually exclusive hypotheses. The first instance of the Bayesian technique for BGA prediction was used by Lowe et al. (2001) to infer BGA for crime scene samples to reduce the number of potential suspects in a police investigation. To date, the Generic Bayesian approach has been utilised as an information tool in multiple BGA prediction studies (Cheung et al., 2017, 2018a; Gettings et al., 2018; Jin et al., 2018; Kidd et al., 2014; Phillips et al., 2009; Phillips et al., 2014; Phillips, 2015; Rishishwar et al., 2015; Tvedebrink et al., 2017, 2018; Tvedebrink and Eriksen, 2019).

The 11-M Madrid commuter train bombing is also a key example in the literature for the application of the Generic Bayesian method in BGA prediction (Phillips et al., 2009). In conjunction with the STRUCTURE output shown in Figure 3.4, Phillips et al. (2009) also calculated a log LR scale to create a range of expected log LR values for the individuals (note a logarithmic scale was applied to the LR to create a manageable scale). As shown in Figure 3.6 the distribution of LR values for individuals from the Moroccan population are shown in the left-hand section to be greater than 1, with the distribution of LR for the Spanish population in the right-hand section shown to be less than 1. The seven crime scene samples, shown on the right, are then compared to the estimated sample ranges to determine which, if any, can be assigned ancestry based on LR classification thresholds. Phillips et al. (2009) outline their classification thresholds as: (*i*) Log LR ≥ 100 = Moroccan BGA (which they equate to North African ancestry), (*ii*) Log LR ≤ 0.001 = Spanish BGA (which they equate to European ancestry), and (*iii*) if the Log LR is between 0.001 and 100 the individual remains unassigned as there is minor overlap between the Moroccan and Spanish samples. For the unassigned situation, ancestry is ambiguous. Three crime scene samples (1, 4 and 6) are classified as unassigned as they fall within the classification zone of 0.001 to 100 as previously described.

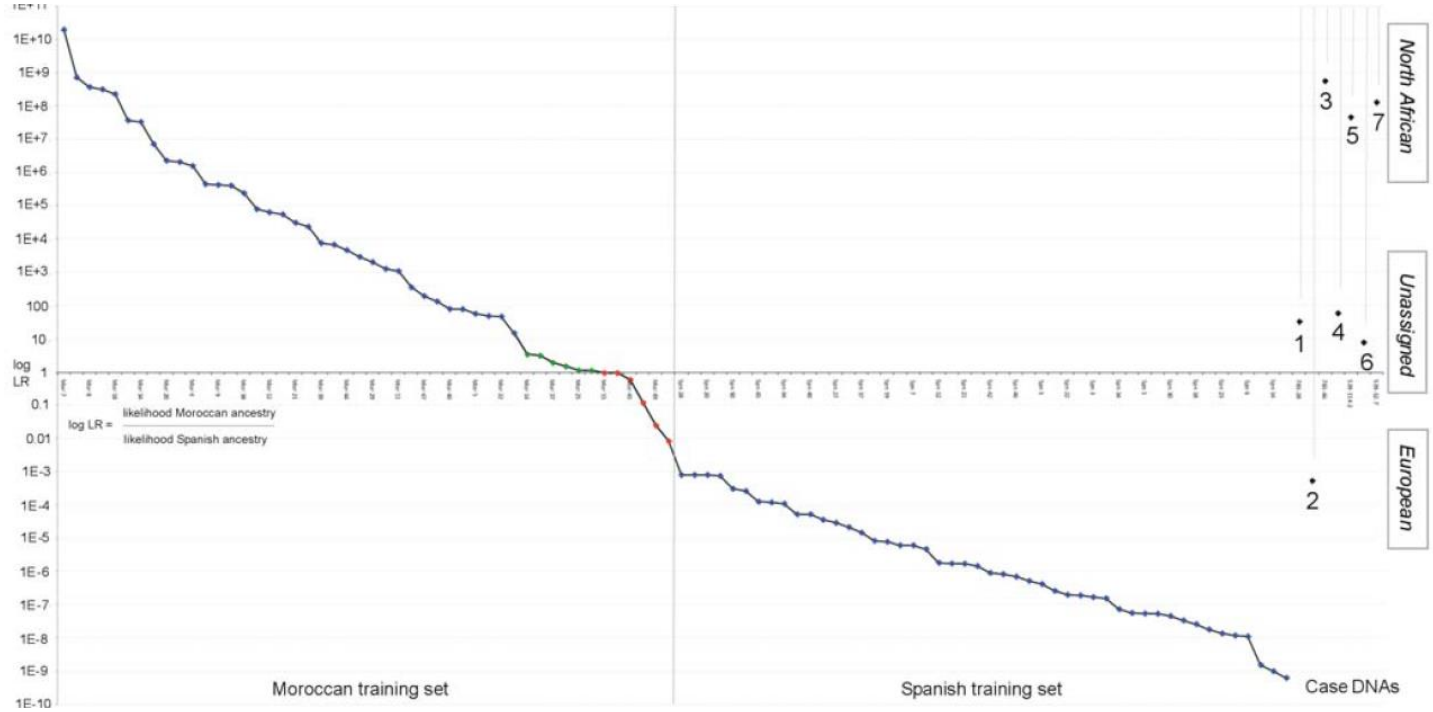


Figure 3.6: Phillips et al. (2009) Log Likelihood Ratio Output
Log likelihood ratio classification system to classify several test samples (Case DNAs) as belonging to either a North African (Moroccan n = 48) or a European (Spanish n = 48) origin (Sourced from Phillips et al. (2009, Figure 1, p.4)).

Limitations that are observed in Phillips et al. (2009)'s analysis that the authors do not comment on are:

1) The sample sizes used to generate the classification thresholds are relatively small (n = 48 each), and these may not accurately reflect the true populations;
2) No margins of error are attached to estimated likelihood ratios, providing little reliability in their respective classifications;
3) Possible overestimation of likelihood ratio values by multiplying genetic markers which were selected through selection criteria, a limitation previously discussed for the STRUCTURE classifier, and addressed by Cheung et al. (2017);
4) The LR inability to handle a relative frequency of zero when an allele is not observed.

GenoGeographer is software that performs BGA prediction using an adjusted version of the Generic Bayesian approach, incorporating a *z*-score analysis to determine if any of the utilised populations are in fact relevant (Mogensen et al., 2020; Tvedebrink et al., 2017, available at https://cran.r-project.org/web/packages/genogeographer/index.html). The software addresses several limitations of the Generic Bayesian classifier, such as the removal of zero probability values (see Concluding Remarks for this section) and the suitability of populations. Another application of the Generic Bayesian approach is seen in Rishishwar et al. (2015), who utilised the method to assign sub-continental African BGA to Afro-Colombians. A technique used by Rishishwar et al. (2015) that has not been utilised widely in other applications is the incorporation of historical data for estimating a value for the prior odds ratio. Rishishwar et al. (2015), based on a simulation study, concluded that incorporation of historical data improved the overall classification accuracy. The significance of this technique is discussed in Section 3.2.6.

Genetic Distance Algorithm (GDA). GDAs are used to measure the cumulative genetic distance of genotypes/alleles over multiple DNA markers between populations of interest. Several statistical measures have been proposed for measuring the divergence between populations, based on varying evolutionary models. Commonly used algorithms are Nei's measure (Nei, 1972), Cavalli-Sforza and Edward's measure (Cavalli-Sforza and Edwards, 1967), and Reynolds, Weir and Cockerham's measure (Reynolds et al., 1983).

Compared to the Bayesian classifiers, GDAs are less commonly utilised for inferring ancestry of an unknown individual. Comparing an in-house GDA algorithm to the other classifiers discussed in this chapter, Cheung et al. (2017) found the GDA to have the lowest classification accuracy for non-admixed individuals. For admixed individuals, the GDA underperformed compared to other classifiers, including STRUCTURE, for basic admixture scenarios, that is with only two contributing populations. However, for complex admixture scenarios, with three or more contributing populations,

the GDA had the highest classification accuracy. As Cheung et al. (2018a, p.109) state "*GDA is a more accurate classifier than STRUCTURE when the degree of admixture increases, and particularly when the ratio is evenly divided between reference populations*". Unlike Bayesian classifiers the GDA makes no assumptions regarding Hardy-Weinberg equilibrium or other assumptions for the genetic structure of the population.

Multinomial Logistic Regression (MLR). MLR is a regression analysis which aims to predict a categorical dependent variable consisting of two or more levels, from a set of independent variables. Prior to BGA prediction, MLR was already being utilised in ancestry studies as a method for phenotype predictions of hair (Walsh et al., 2013) and eye colour (Liu et al., 2009; Walsh et al., 2011; Walsh et al., 2013). In these phenotypic studies, hair and eye colour (the dependent variables) were predicted using numerous SNPs (the independent variables) that were known to have statistically significant association with hair and eye colour. An example of how SNPs can be used to predict hair colour is shown in Figure 3.7.

Figure 3.7: Multinomial Logistic Regression Pathways for Hair Prediction
Walsh et al. (2013, Figure 2, p.105) describes the pathways necessary for which hair colours (shown in boxes) are obtained based on the observed genotypes at the dependent SNPs. By observing which set of alleles are present at the SNPs of interest and the resulting combination, one can predict the phenotype that will occur. For example, an individual

with the following alleles: rs1805005-TT, rs1110400-CC, rs2228479-AA, rs12821256-GG, rs4959270-AA and rs1800407-AA will be predicted to have blonde hair.

McNevin et al. (2013) compares MLR to STRUCTURE to determine whether MLR could be a viable alternative for BGA prediction. The two classifiers' prediction accuracies were compared using several admixed and non-admixed subpopulations from the Human Genome Diversity Project (HGDP) database (Africa n = 98, Europe n = 157, East Asia n = 225, Oceania n = 27, Indigenous Americans n = 64). These authors commented that MLR is a practical substitute for STRUCTURE as the loss of accuracy from STRUCTURE to MLR was minimal. In addition, unlike STRUCTURE, MLR does not require assumptions such as Hardy-Weinberg equilibrium to be valid and is readily performed even in an excel spreadsheet (and other statistical software) as opposed to STRUCTURE's strictly formatted input files and long run-times.

Logistic Model Tree. Landwehr et al. (2005, p.16) describe the LMT algorithm as "… *a standard decision tree structure with logistic regression functions at the leaves…"* providing a combination of these two statistical classifiers. The idea of combining tree induction with logistic regression follows a logical application as the benefits of each approach complement the limitations of the other. Simply put, tree induction can exhibit low bias but high variance, while conversely, logistic regression can have high bias with lower variance (Landwehr et al., 2005). Note that this section only provides a basic summary of the LMT algorithm's methodology as the focus of this thesis is not specifically this method, but rather on a specific application of it (see Section 4.3). The reader is referred to Landwehr et al. (2005) for a full description of the method. Before describing the LMT algorithm, a short summary is provided for tree induction and for logistic regression.

Tree Induction. A decision tree can be described as a classification model that builds events which contain various outcomes using conditional 'rules' occurring over a number of variables (referred to as attributes) (see Figure 3.8 for an example), "*a map of the possible outcomes of a series of related choices.*" (Lucid Chart, n.d.). A decision tree is comprised primarily of "nodes" which generally take two forms: *i*) a question node (sometimes referred to as a chance or terminal node) which describes a range of values for a given attribute, and *ii*) an outcome node which describes the observed classification outcome based on the tree pattern leading up to this node. Nodes are connected to outcomes through branches, which describe the conditional rule applied to the attributes in the node to determine which outcome is observed. The outcomes of a question node can be one of two possibilities: *i*) either the question node results in a brand-new question node (with a new attribute of interest and conditional branches) which extends the length of the tree, or *ii*) the outcome can result in a leaf node, or terminal node, (represented by a square). A leaf node will contain the predicted

class output, which the user is attempting to classify using the available variables (Perner, 2015). Figure 3.8 demonstrates a simplified example of how decision trees are presented and interpreted. The question of interest is determining whether an individual should go to the beach on a given day based on three variables of interest: whether it is raining, the temperature (°C) and whether there are sufficient shark safety measures in place. As seen in Figure 3.8, this tree has three levels based on the number of descending question nodes, with each question node containing a single attribute variable (note, an attribute can be comprised of more than one variable). Each question node in Figure 3.8 consists of binary branches – however, multivariate branches are also possible – resulting in either the next question node or the predicted class output, "Don't go to the beach" or "Go to the beach".

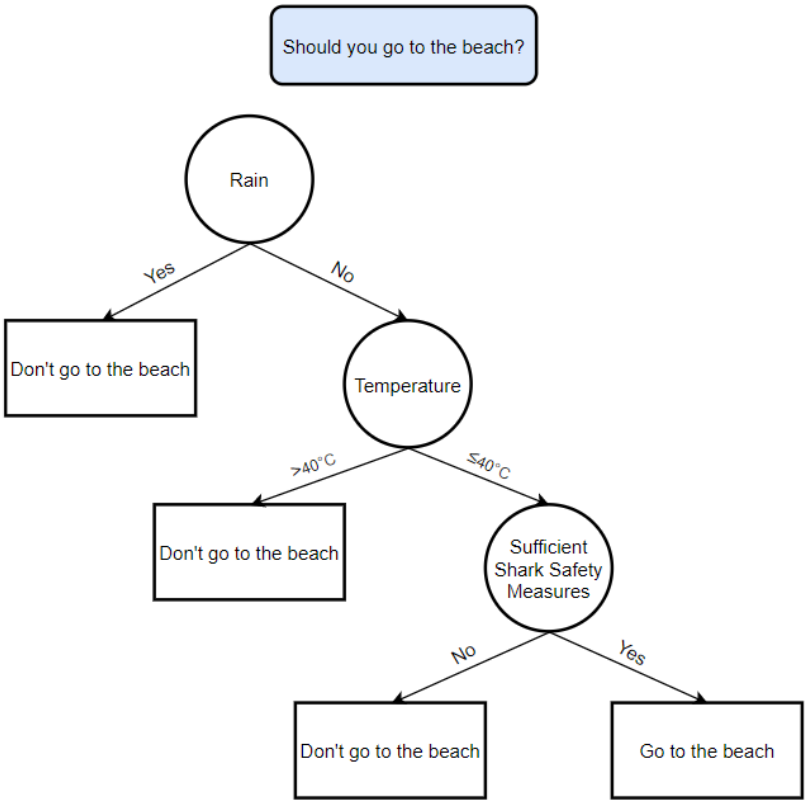


Figure 3.8: Simple Decision Tree
Basic decision tree to support the question, "should you go to the beach?", utilising three decision nodes.

Decision trees can become increasingly complex as more variables are utilised causing the tree itself to become a large network of conditional outcomes.

There are numerous algorithms for creating decision trees, each one producing different trees based on varying methods used for selecting which attributes will be used to build nodes and the criteria to be used for splitting the nodes into branches (Song and Ying, 2015). Notable algorithms found in the literature include: C4.5 (Quinlan, 1994), CART (Breiman et al., 1984), CHAID (Kass, 1980), ID3 (Quinlan, 1986) and M5 (Quinlan, 1992). Each algorithm has advantages and disadvantages for different scenarios and determining which one to use will depend on the available data and the question of interest (Song and Ying, 2015). Each decision tree algorithm has a stopping criterion (similar to Shannon's balance of information versus error) according to which the gain of information is minimised, and tree generation stops. While the difference between stopping criteria is not a topic discussed in this thesis, the logic behind most algorithm's stopping criteria can be broadly described as a threshold after which the data can no longer be partitioned into distinct homogenous sub-groups (Hssina et al., 2014; Mingers, 1989; Singh and Gupta, 2014; Song and Ying 2015). It should be noted that not all input variables will always be used. Some variables may not be deemed informative enough and thus will not be included in the tree.

Logistic Regression. Logistic regression describes the relationship between a dichotomous outcome variable and one or more predictor variables (Peng et al., 2002). As an example, consider a bank detecting credit card fraud application. When a purchase is made using a credit card, the bank receives data on numerous variables such as: where the transaction occurred, the amount of the transaction, date of transaction for this particular individual, and category of purchase. Using the data, the bank can then build a logistic regression model based on the individual's typical purchases, that is, building a characterized profile of what is a standard purchase and what constitutes an out of the ordinary or an "outlier" transaction. When a purchase is made which is outside what the model considers a "normal" purchase, based on a predetermined threshold, the transaction is flagged for further investigation, and is considered potentially fraudulent.

Logistic Model Tree. The first stage of LMT is creating the initial decision tree, which is constructed using a standard classification tree algorithm, such as the C4.5 algorithm. The LMT will build the tree by analysing the available variables and estimating the information gain, that is, the ability to distinguish between the two populations of interest (known as classes), for each individual variable and for sets of variables. The variable/s with the highest information gain is then used to construct the tree. Once the tree has been constructed, logistic regression models are fitted to the leaves of the tree, resulting in the output of the LMT algorithm as a regression formula for each population of interest. An important aspect of the LMT algorithm is the way in which missing values are handled in the training data. Missing values are replaced with the mean of that respective variable. Landwehr

et al. (2005) note that this simplistic approach did not cause noticeable issues during analysis but suggest more sophisticated methods could be an advantage in the future.

By implementing the concept of information theory – how many SNPs are needed to identify the true ancestry – multiple LMTs of different subsets of the available SNP panel can be generated. For example, the LMT is sensitive to rare genotypes, and if a SNP is present where the genotype is fixed in one population but absent in the other, the algorithm will likely construct a tree which solely utilises this one SNP. The limitations of utilising a single LMT are:

i) As remains in a historic military case are likely to be degraded resulting in partial profiles, the SNPs present in the single LMT may not be available;

ii) The sample sizes used to generate the LMT may be small (≤100). Note this has already been flagged in the majority of forensic applications (such as Cheung et al. (2018a) and Phillips et al. (2009)). There is the possibility that the genotype which is believed to be absent based on the sample may in fact be present in the population, the sample size was simply insufficient to detect it.

By obtaining estimates of ancestry from a number of independent models, a better knowledge of the true panel value and thus of the true ancestry will be possible. This parsimonious adaptation of the LMT can ensure that the most informative SNPs are included in the models, while SNPs that only add noise to the analysis are excluded.

Parsimonious Logistic Model Tree. To overcome the issue of utilising a classifier which may consist of only a single SNP, a variation to the basic LMT approach is developed. Where multiple models are coupled with the concept of parsimony. Instead of a single model being obtained from the available SNPs, multiple LMT models are generated iteratively. Using information theory's logic of balancing information and error, the Parsimonious LMT (*p*LMT) approach determines the number of LMT models, *M*, that can be generated from a set of SNPs (without replacement) before a threshold is reached where the gain of information is offset by the gain in error, certainty and uncertainty are balanced. Details of this approach are given in Section 4.3.

Concluding Remarks. When comparing the Generic Bayesian approach to STRUCTURE, GDA, and MLR, Cheung et al. (2017) found that the correct classification rate of the method was only minimally lower than that of STRUCTURE for non-admixed individuals. Since the Generic Bayesian is computationally simpler and it can be readily implemented in a spreadsheet, Cheung et al. (2017) suggested using it as an alternative to STRUCTURE when analysing non-admixed individuals with minimal loss in accuracy. However, for admixed individuals, the Generic Bayesian method was

considered less effective. Unlike STRUCTURE which can output several proportions of the relative contributing populations, the Generic Bayesian approach provides an "all-or-nothing" output, that is, "…*individuals are assigned largely to a single BGA despite the presence of admixture*" (Cheung et al., 2018a, p.109). By assigning individuals to a single population, the Generic Bayesian – and any other "all-or-nothing" classifier – is not useful for determining an individual's admixture proportion. Cheung et al. (2018a) also states that BGA predictions should not be performed using "all-or-nothing" classifiers in general. From the above it becomes apparent that if one's interest is in estimating relative admixture proportions for multiple populations, then classifiers such as STRUCTURE are more suitable, but for scenarios where interest lies purely in assigning an individual to a single BGA, "all-or-nothing" classifiers such as the Generic Bayesian are attractive options due to their computational simplicity.

For Cheung et al. (2017)'s comparison of MLR, STRUCTURE, Generic Bayesian, and the GDA on non-admixed individuals, MLR's overall classification accuracy (98.25%) was only slightly lower than those of STRUCTURE (100%) and the Generic Bayesian (99.5%). As MLR was categorised by Cheung et al. (2018a) as an "all-or-nothing" classifier, MLR was excluded during experiments on admixed individuals. One limitation of MLR raised by Cheung et al. (2018a) is that the method is highly sensitive to single locus effects, that is, having the classification model being driven heavily by a genotype that is fixed in one population and absent in the other population, which may cause other informative SNPs to be overlooked by the model.

There are three additional limitations of the Generic Bayesian approach. As is the case with most Bayesian classifiers (including STRUCTURE), the Generic Bayesian method assumes that all loci are independent of each other. This assumption has been contested in the literature by both Cheung et al. (2017) and McNevin et al. (2013), both of whom have commented on the matter stating DNA markers chosen under selection criteria are unlikely to adhere to such assumptions in reality. From a practical forensics' application perspective, this assumption may be true: markers are often selected using metrics such as linkage disequilibrium (measuring the dependency or correlation between genetic markers) (Phillips, 2015) to ensure a panel consists of independent markers. However, this assumption may not be valid if one considers how these markers may be linked through an individual's heritage. The goal of ancestry prediction is to create a panel where each DNA marker has good discrimination power between two or more populations, therefore, it is possible that these selected markers may be correlated.

The second limitation of the Generic Bayesian, specifically when outputted in the format of the LR, is the difficulty of interpreting the output number. This limitation becomes exceedingly complex when the predicted ancestry outcome is not binary, but rather, comparing multiple populations. In these cases, the recipient is left with a series of pairwise LR statements to weight the unknown. In the literature, two possible solutions to alleviate the difficulty of LR interpretation have been suggested. The first solution is the application of some form of scale, such as the logarithmic scale as in Phillips et al. (2009). The second approach is the use of a qualitative or 'verbal' scale which attempts to substitute numbers with words, which some suggest are easier to follow. Ballantyne et al. (2017) provide an example of a verbal scale (Table 3.2).

Table 3.2: Verbal Scale for the Likelihood Ratio
Verbal scale to assist forensic scientists when providing LR evidence to the jury (Sourced from Ballantyne et al. (2017, Table 1, p.8)).

Verbal Conclusion (Support for or against the referent)	Corresponding Likelihood Ratio
Extremely strong support against	< 0.000001
Very strong support against	0.000001 – 0.001
Strong support against	0.001 – 0.01
Moderate support against	0.01 – 0.1
Slight support against	0.1 – 1
Neutral	1
Slight support for	1 – 10
Moderate support for	10 – 100
Strong support for	100 – 1,000
Very strong support for	1,000 – 1,000,000
Extremely strong support for	> 1,000,000

The issue with utilising a verbal scale is the subjectivity of how an individual interprets that verbal conclusion. While Ballantyne et al. (2017)'s scale deems a LR of 10 – 100 as "Moderate support for", that does not mean to say that another individual would have the same conclusion for this range. Rather, the scale seems to have been arbitrarily ranked in magnitudes of ten, rather than through empirical surveys of individual's beliefs. As stated in Marquis et al. (2016, p.4) "*…numbers allow us to make the distinction that words cannot make…only numbers can cope with this challenge*". A study by Berger et al. (2011, p.47), evaluating various court appeals revolving around the interpretation and reporting of DNA evidence, suggested that "*In those cases where a quantitative likelihood ratio has been calculated…it is the number alone that should be put to the jury*". The

interpretation of LR values remains a limitation in the investigative ancestry context, as ultimately there is still a human element involved in decision making that involves the subjectivity of comprehending large numbers. The interpreter must also be aware when interpreting likelihood ratios that simply calculating that one population is more likely to occur than another does not imply that it is the correct or relevant population (Tvedebrink et al., 2018). An additional limitation to interpreting LR values beyond a verbal scale, is that most BGA studies do not rely on a sole LR, rather, a comparison of several LRs between all possible populations and their pair-wise permutations which are then ranked. An issue with this approach is that subconsciously, most people will tend towards the largest LR value and ignore all remaining outcomes.

The final limitation of the Generic Bayesian method is its inability to handle zero probabilities. If a genotype/allele is absent in a sample, then its frequency estimate will be zero. However, in another sample from the same population the genotype/allele may be seen; it is simply an issue of an event which occurs rarely in a population not being seen in a particular sample. As the Generic Bayesian is calculated using the LR, which is estimated through the multiplication of an individual's genotype frequencies, any zeros in the calculations will result in an uninformative result – either the numerator or the denominator of the LR will be zero. An approach utilised in the literature to account for zero sample probabilities, is to apply a conservative frequency (based on sample size) (Gettings et al., 2018; Graydon et al., 2009; Lowe et al., 2001; National Research Council, 1996; Phillips et al., 2007; Voskoboinik et al., 2018). Different approaches for estimating a conservative minimum frequency were found in the literature (Table 3.3).

Table 3.3: Conservative Minimum Frequency Methods
Examples of conservative minimum frequency estimation methods from the literature where n is the sample size.

Articles where a minimum frequency has been utilised	Conservative minimum frequency formula
Budowle et al. (1991); Gettings et al. (2018); National Research Council (1996); Voskoboinik et al. (2018)	$\frac{5}{2n}$
Graydon et al. (2009); Phillips et al. (2007)	$\frac{1}{2n+1}$
Lowe et al. (2001); Mogensen et al. (2020)	$\frac{1}{n}$

The methods shown in Table 3.3 are ad-hoc, with little empirical or theoretical support. The original minimum frequency proposed for forensic science situations (National Research Council, 1996), stemmed from need to have a minimum number of entries to address sampling error (Budowle et al. 1996).

Based on these limitations of both STRUCTURE and the Generic Bayesian method, the *p*LMT was selected as the analysis machine for inclusion in DNA-MAP, however, as part of this thesis, these two additional classifiers are also applied for comparison purposes (see Chapters 4 and 5 for methods and results, respectively). It is important to note, that the focus of this thesis is the construction of the user-friendly KBDSS, which is constructed in such a way that the selected classifier should be interchangeable with other statistical methods. Therefore, while the choice of which classifier to implement in the KBDSS's prototype is a personal preference (and should not be viewed as the final choice), during this early stage of development it was considered desirable to select and validate a method with a high accuracy and few limitations.

3.2.4 Relevant Populations

Selecting the appropriate populations is imperative for accurate classification. It is first important to define what is a "population" in the context of BGA prediction. Depending on the scope of the study a population can consist of a continental group, if interest lies solely in distinguishing between racial groups such as Asian versus European, or can be as specific as the distinction between two separate human clades within a regional area. Note that as the scope becomes more defined, so too does the difficulty of finding highly variable genetic regions between the populations of interest since there is less chance of biodiversity evolving from geographic separation. The term population can be further defined in terms of homogeneity, determining whether admixture between groups has occurred which may also restrict the ability to accurately assign BGA. For contemporary populations, the issue of admixture is rapidly becoming a greater limitation for BGA studies; with the modernisation of commercial travel, the rate of admixture between distant populations is expected to increase.

Several criteria need to be considered including geographic location, population history, age, biological sex and historical time. Typically for BGA studies, both commercially and in current research, geographic location and population history are the two primary criteria of interest. When creating a reference database for a population, a researcher has two options: (*i*) collection of a new sample, and/or (*ii*) the use of freely available data from online databases. As previously discussed, the databases within commercial ancestry companies are kept as trade-secrets and very rarely can be accessed by researchers. Phillips (2015) states that the following three databases are commonly utilised for research: (*i*) the 1000 Genomes Project (Genomes Project Consortium, 2015), (*ii*) the CEPH Human Genome Diversity Panel (Li et al., 2008), and (*iii*) the Allele Frequency Database (Rajeevan et al., 2012).

Determining which populations, and how many, to include in an analysis relies heavily on the question of interest. For example, if a study is concerned with the distinction and classification between only two populations of interest, additional resources can be directed towards collection of appropriate individuals from these two populations. However, when a study focuses on a global scale, it can be a difficult task to determine which (sub)populations to include, given that it will be impossible to sample all (sub)populations. The researcher should consider the possibility that a classifier may infer an individual belongs to a specific population, however, this may not be the individual's true population of origin, which was not present in the original samples. For example, consider a study consisting of only Asian populations. To then subsequently classify an unknown individual, the sample will likely be inferred to have Asian ancestry given that these are the only reference populations available. However, the true origin of the unknown individual may in fact belong to any number of ancestries, but no samples from the individual's true origin were available in the reference database. Such an outcome may lead to erroneous, classifications, a perspective shared by Kidd et al. (2014) and Themudo et al. (2016). Online databases might consist of numerous, specific populations, however, such populations are not exhaustive (Tvedebrink et al., 2017, 2018; Tvedebrink and Eriksen, 2019).

Researchers/investigators should also be wary of the assumption that their databases include all relevant populations, since publicly available databases such as the 1000 Genomes Project (Genomes Project Consortium, 2015) are used without questioning how the observations in these resources were obtained. Royal et al. (2010) raise several points regarding this issue: (*i*) certain ancestral populations cannot be accurately represented as a definitive sample no longer exists, (*ii*) admixed populations are severely under-represented, and most importantly, (*iii*) proxy populations are often poor representations of the desired population. An example of a doubtful proxy population is in the 1000 Genomes Project's (Genomes Project Consortium, 2015) British individuals from England and Scotland (abbreviated as GBR). Information readily available through the 1000 Genomes Project shows that this GBR sample was collected from individuals in Kent and Cornwall (England), the Orkney Islands, Argyll and Bute (Scotland) (Figure 3.9). It is hard to see how these samples could be a truly representative and unbiased representation of the general British population's genome when the samples are collected from remote areas that may not represent the true nature of the greater British population. Despite this, researchers use this sample as an overall representation of Britain (see for example, Bulik-Sullivan et al. (2015), Khrameeva et al. (2014), and Ramos et al. (2014)), without commenting on the accuracy of the representation.

Figure 3.9: 1000 Genomes Project's GBR Location
Geographic location of the source of the 1000 Genomes Project's GBR sample (n = 91), Cornwall, Kent, Orkney, and Argyll and Bute.

Great care must be exercised when selecting which populations to include in an analysis, and the researcher/investigator should determine which populations are relevant to the question of interest. There is little point in trying to get an accurate estimate, only to realise that the sample used is inadequate or atypical of the true population, and that conclusions inferred and assumed to be applicable to the population of interest, may be invalid and misleading.

3.2.5 Sample Size and Rare Event

Prediction of ancestry relies on the comparison of allelic and/or genotypic frequencies which are estimated using samples from different populations. Sample sizes used typically vary between and within studies. Little, if any, attention is given in the literature to the following important issues which arise when sample sizes are small: (*i*) the effects on the estimation of the allele frequency and (*ii*) the real possibility that some alleles present in a population may not be identified in the sample (Chakraborty, 1992). The importance of "rare" allele detection in ancestry analysis is ensuring that any statistical models generated based on observed allele frequencies are accurate. A rare, or elusive allele is characterised by occurring in low abundance, restricted due to geographic elements, or carrying a low probability of detection (Budowle et al., 1996; Chakraborty, 1992). In the forensic literature, there have been several definitions for what constitutes a rare allele, Budowle et al. (1996)

use a frequency less than or equal to 0.01, while Chakraborty (1992) uses values of 0.05, 0.01, and 0.001 during his analyses. When an algorithm is used to find the most parsimonious set of genetic markers for classifying individuals, it uses all available information. When an allele is rare it might be missed in a small sample and thus, this information is not available to the algorithm which might lead to misclassification.

Loci used for inferring ancestry are typically those found to provide high discrimination power. The best case, and most desirable scenario, is when specific alleles are found in all individuals of one population of interest and none in another. Thus, the alleles common (potentially fixed) in one population are rare (potentially absent) in the other. When the alleles are 'rare' it is likely that these alleles will fail to be detected in any relatively small sample of individuals usually found in the literature (Chakraborty, 1992; Hackshaw, 2008). However, such failure to detect does not mean that the allele is not present in the population: absence of evidence should not be interpreted as evidence of absence. To demonstrate the impact failing to detect a rare event may have, consider an ideal SNP which is fixed with one allelic state in one population while the other allelic state is fixed in the second population. Note that this decision will be based on the available sample data. Despite this expected situation, there will always be the possibility of a small percentage of one of the populations having the unexpected fixed allelic state, either due to a random mutation occurring at some prior generation and by chance being passed down, or through an unknown admixture event. If the prediction-based modelling technique employed to assign BGA was heavily influenced by this SNP (a fair assumption considering its ideal discrimination power), then there is a possibility for misclassification for this small percentage. Due to the SNP's high discrimination power, these individuals who truly belong to one population may be misclassified to the other population simply because the original sample taken was of inadequate size to detect the rare allele and adjust the modelling to account for this rare case.

Articles proposing techniques for dealing with and acknowledging rare events detection have been published in numerous research areas, with some examples shown in Table 3.5.

In these papers (Table 3.5) it is clearly acknowledged that not detecting a particular observation in a sample does not mean that the observation is not present in the population; it simply means that the item was not seen in the sample. Underlying all considerations of the impact of using samples to make decisions about populations is the assumption that the sampling was carried out randomly, that is, every allele could be sampled according to the probability with which it occurs in the population. It is also assumed that the sample is taken from a single homogeneous population of interest. In

assigning ancestry, consideration should also be given to differences in the genotype distributions for different ancestral populations and not just to the allele frequencies.

Table 3.5: Examples of Rare Events in Various Disciplines
Examples from multiple disciplines regarding their equivalent rare event, with respective papers providing methodology to detect these events.

Scientific Discipline	The Rare Event	Article
Ecology	A rare species of flora/fauna	Krebs (1980); Robinson et al. (2018)
Epidemiology	An unexpected disease within a healthy population of people	Kamangar and Islami (2013); Miller et al. (2018)
Geology	A rare mineral	Hystad et al. (2015)
Veterinary Science	A diseased-state animal within a healthy herd	Humphry et al. (2004).
Brewers	A defective bottle	Gojanovic (2007)
Clinical Science	Uncommon drug side-effect	Chow et al. (2007)
Post-Marketing Safety	A severe/irreversible drug reaction	Makuch (2006)
Psychology	A gambling decision with a low probability of success	Rakow et al. (2008)
International Relations	The occurrence of wars, coups, revolutions, economic shocks	King and Zeng (2001)
Air Traffic Management	Aircraft collisions	Nassar et al. (2011)
Transportation	Vehicle accidents	Theofilatos et al. (2016)

Chakraborty's early work from the 1990s forensic literature lists the difficulties associated with observing rare alleles when sampling from a population (Chakraborty, 1992). He outlines the limitations and approaches of sampling that need to be considered to achieve an accurate representation of the population, addressing the effect that a small sample size has on the reliability of detecting a rare allele. The effects of sample size can be seen in studies such as Szabolcsi et al. (2015) where a newly collected sample from a previously sampled population resulted in the detection of 85 previously unobserved alleles; the previous sample was of size 4213 individuals whereas the new sample consisted of 21,473 individuals. Despite the literature advising against the use of small samples due to potentially imprecise estimates (Chakraborty, 1992; Hackshaw, 2008;

President's Council of Advisors on Science and Technology, 2016; Royal et al. 2010), population inference in ancestry analysis is commonly carried out using samples of insufficient size. As stated in the President's Council of Advisor's on Science and Technology report "*When the sample size is small, the estimates may be far from the true value…"* (President's Council of Advisors on Science and Technology, 2016. p.153). It is suggested that this statement should be further expanded to incorporate that a small sample size may also miss events with a low probability of occurrence, namely, that a rare allele may not be seen.

To calculate the minimum sample size required to detect a rare event, three specific methods are compared from the literature, one from within forensic science, and two from other disciplines. These methods were selected based on the criteria of (*i*) adequately answering the question of interest, that is, what is the minimum sample size required to detect a defined rare event, (*ii*) utilise simplistic calculations, and (*iii*) are generalisable to any ancestry prediction scenario? It is acknowledged that other methods for sample size calculation are available in the literature, both in forensics and other disciplines, however they did not meet the previous criteria. For those interested in alternative methods published in the forensic literature, the reader is directed to the following studies (Aitken, 1999; Brenner, 2010, Cereda, 2017; Cereda et al., 2018; Cereda and Gill, 2020).

In the forensic literature, Chakraborty (1992) assumed that the variable of interest (number of observed alleles) has a binomial distribution and used this probability distribution to develop formulae and tables of recommended sample sizes. This distribution is justified as being appropriate when the variable of interest has only two possible outcomes such as the tossing of a coin to give either heads or tails with the variable analysed being the number of heads seen in a fixed number of tosses. In the current application, the variable of interest comes from a "yes" or "no" answer to the question "Does this individual have the rare allele of interest". If so, is there one or two of these rare alleles present?', followed by considering the number of individuals to which the answer is 'yes', together with how many alleles are present for each individual, to obtain a count of the number of alleles. In his methodology, Chakraborty outlines the calculations for estimating the minimum number of individuals needed to observe a specified number of alleles with a given minimum frequency (See Chakraborty (1992, Table 6, p.152)).

Green and Young (1993) outline methodology for estimating the sample size for collection of different species of molluscs, rather than the occurrence of a rare allele. As Green and Young state, "*One can only decide how rare a species one wants to detect, and then allocate sampling effort accordingly*" (Green and Young, 1993, p.356), pointing out that sampling effort, such as sample size,

is directly related to detectable rarity. In their work Green and Young (1993) use the Poisson distribution in place of the binomial, as relevant for the occurrence of rare events. They also consider the negative binomial distribution which specifically addresses the question (in the DNA setting): "how many individuals must be sampled before a specified number of successes (rare alleles) is seen?" In the forensic science setting of detecting a rare allele, this question of interest becomes more specific, namely, 'how many individuals must be sampled before the first occurrence of a single rare allele is seen?' This special case of a negative binomial is known as the geometric distribution in the statistical literature (Kotz, 2006).

As stated in Chakraborty (1992), the Poisson distribution is closely related to the binomial distribution and is a good approximation for modelling when the probability of the allele is rare. Green and Young (1993) show that for rare events, the formula for sample size calculations using the simpler Poisson distribution is approximately the same as those for the negative binomial and is identical with the formula derived from the binomial distribution. They conclude that: "… *the simple Poisson-based formula is usually adequate for estimating the necessary number of samples to detect the presence of a rare species*." (Green and Young, 1993, p.355). Jovanovic and Levy (1997) present the 'The Rule of Three', a very simple method for sample size determination, which they say is part of the 'folklore' seen in clinical research. They develop this simple equation for a binomial distribution, a Poisson distribution, and a Bayesian approach. The use of rare event calculations is shown in Section 4.8.

3.2.6 Prior Probability

In the context of BGA prediction the prior probability is the probability of an individual belonging to a population prior to genotyping. For example, suppose investigators are tasked with identifying remains from a mass-grave where it is known that individuals from two populations are buried. Also known are the number of individuals per population within the grave. The prior probability could be the ratio of the two population numbers. To illustrate the possible effect of the prior probability, consider a scenario where a forensic scientist is attempting to predict ancestry for a set of remains discovered in a given geographic area. The remains are assumed to belong to either Population A or Population B, and the utilised DNA panel is comprised of a single SNP. Based on the observed genotype for the set of remains, the probability of observing this genotype in Population A is 0.95, while for Population B is 0.1. Therefore, it is 9.5 times more likely (0.95/0.1) that the remains originate from Population A than Population B. Now consider, if prior to DNA testing, the forensic scientist has knowledge regarding the possible population sizes of Populations A and B in that given area and can incorporate this knowledge. In this area, there are a total of 1050 remains, 50 of these

are believed to be individuals from Population A, and 1000 from Population B; clearly, Population A is largely outnumbered, and if no DNA testing was available, a set of remains is more likely to belong to Population B. Despite the genotype being more common in Population A (approximately 95% of the 50 = 48 individuals present should have this genotype), there is still a greater chance of the unknown remains belonging to Population B (approximately 10% of the 1000 = 100 individuals present should have this genotype) simply due to the mass disproportion between the two populations. It should be noted that the inclusion of additional SNPs will provide the discrimination power necessary to reduce the effect of the prior. Alternatively, cases with populations with greater distinction between their genomes would have a lessened effect from a prior probability, due to the preliminary discrimination power. However, it is still important to acknowledge the importance the prior probability can have on the resulting ancestry prediction. The incorporation of a prior into BGA prediction will be discussed in Section 4.8.

Budowle et al. (2011, p.2) discuss concerns regarding the prior odds in a forensic context, stating that while "*…the forensic DNA community has made recommendations for using Bayes' Theorem, they have not addressed the variables that should be considered when establishing prior odds…*". The primary concern that Budowle et al. (2011) raise is the methods used to estimate a prior odds value, where the typical approach observed is to use simply $\frac{1}{v}$, where v is the pooled number of potential victims. Budowle et al. (2011) outline how utilising $\frac{1}{v}$ can be a poor estimate, using the work of Pajnic et al. (2010) as an example, where the authors were attempting to identify WWII remains for a given area in Slovenia. A prior odds ratio was chosen based on the estimated number of victims within the mass grave, however, it was suggested that some remains may have been buried in a separate site. Budowle et al. (2011) raise further concerns for this approach as not all remains were successfully identified, and the authors did not consider the possibility that additional victims may be present in the mass grave.

Two other methods for inferring a value for the prior odds were found in the literature for BGA prediction. The most common method is to use a prior odds ratio equal to 1, a so called non-informative prior as it assumes the prior is equal to one for all populations, resulting in the prior having no effect on the posterior probability. When Lowe et al. (2001, p.19) first introduced the Generic Bayesian approach to BGA prediction for providing intelligence to criminal investigations, an equal prior odds was utilised, supported by the following logic "*…base an ethnic classification of the origin of the crime profile on the DNA information alone*". Additionally, Graydon et al. (2009), who was also concerned with the utilisation of BGA prediction for informing criminal casework,

used the Generic Bayesian approach and assumed an equal prior odds ratio. Rishishwar et al. (2015) used the Generic Bayesian approach to determine whether accurate classification could be made between historical sub-populations within Africa. Rather than assume an equal prior odds Rishishwar et al. (2015) uses historical records to estimate the relative contributions of ancestral regions to the modern-day African populations. It was found that the inclusion of a historical prior, versus a non-informative prior, led to a significant reduction on the misclassification error for known cases (Rishishwar et al., 2015). Despite utilising a non-informative prior odds, Lowe et al. (2001, p.19) comments that the selection of the prior should be fluid, updated based on the given scenario, "*The assignment of prior probabilities will depend on the circumstances of the individual case.*", a comment reiterated by Rishishwar et al. (2015). It must be noted that while the inclusion of a prior is beneficial for measuring the size discrepancy between populations, there is the associated limitation of the so-called prior wash-out. Simply put, as the prior-odds ratio tends towards extreme values, around 0 or 1, the posterior probability tends towards these extreme values as well. This prior wash-out can therefore, cause informative results to be considered otherwise uninformative simply due to an extreme prior probability heavily skewing the results. Therefore, it is imperative for the prior to be based on accurate information to ensure any estimations are accurate. Despite the possibility of prior washout occurring even with informed prior probabilities, the impact of using an uninformative prior is if subsequent estimates result in a different answer, when genuine prior information may indicate otherwise.

Another area of forensics where the prior probability is used is paternity testing, when comparing the probability of the alleged man being the father with the probability a random man from the same population is the father. A paternity index is first calculated by comparing the two previously described scenarios, which utilises a 50:50 prior (also referred to as an uninformative or equal prior). The use of a 50:50 prior assumes that every man in the given population has an equal chance of fathering the child, a grossly inappropriate assumption that is not likely to hold true. Alternatively, if a non-equal prior is selected, the resulting probability of paternity will then change. For example, for a LR of it being 10 times more likely that the alleged father is the true father rather than a random man from the population, the resulting probability of paternity assuming an equal prior is 0.83 (83%). If a prior probability of 0.95 was used however, the resulting probability of paternity is then changed to 0.9 (90%). Therefore, the choice of prior selection is important due to the changes it can cause to the subsequent paternity index; the limitation, however, is that selecting an appropriate prior would be extremely difficult.

It is acknowledged that there may be other prior information that could be used as prior inputs, for example anthropological evaluation indicating an individual belongs to a given racial group based on bone structure, or personal belongings found on the remains. However, for the UWC-A framework these priors are excluded due to the remains found being highly fragmented (it is rare to find remains completely intact) and the high possibility of disturbance from looters or wild animals leading to disturbed burials.

3.2.7 Degraded/Partial Profile

Obtaining a partial, or incomplete, DNA profile is common for forensic samples due to degradation or stochastic errors in the chemical process (Cheung et al., 2017). To determine which classifier is suitable for forensic application, considerations are required for how the classifier handles both complete and incomplete DNA profiles.

Cheung et al. (2017) discusses the limitation of how incomplete profiles affect certain classifiers and simulated "degraded" samples by randomly removing SNPs from the original profiles. These degraded samples were then classified using STRUCTURE, Generic Bayesian, GDA, and an MLR approach. Of the original 142 SNPs used in their panel, test samples had 10% (127 SNPs remaining), 50% (71 SNPs remaining), 70% (42 SNPs remaining), and 90% (14 SNPs remaining) of SNPs missing from the finale profile. Cheung et al. (2017) were able to rank the classifiers based on their ability to correctly assign BGA across various levels of degradation, with STRUCTURE being consistently the most accurate, followed by the Generic Bayesian approach, GDA, and then MLR. It is important to know how missing data is handled by each classifier, as if done incorrectly it could lead to biased results and possible misclassifications. For the Generic Bayesian and GDA, missing data is ignored, namely, missing SNPs are not utilised in subsequent analysis. The correct approach for the MLR, would be to remove the SNPs initially, before the classifier constructs the models used for later predictions.

3.2.8 Margin of Error

Since estimates of genotype frequencies are obtained from a sample taken from a population, a measure of precision of the results obtained should be attached to give a measure of confidence, that is an expression of the margin of error that can be expected in the result. This measure of precision can be obtained by applying a confidence interval to resulting inferences obtained with sample data. The application of a confidence interval is important to achieve accurate reporting, as an outputted result may vary significantly had a different sample had been used, and a confidence interval will account for most of this variation. Currently, there is no standardised method for estimating a measure

of error in BGA prediction. Methodology for calculating confidence intervals varies and is reliant on the statistical modelling used. An example of a previously utilised method for calculating confidence intervals used in forensics is the Delta method (see Chakraborty et al. (1993) and Curran et al. (2002)). The Delta method is used to estimate the variance of a function and is useful in forensic science and BGA prediction for classifiers that utilise functions such as the likelihood ratio, a ratio of two proportions. Bootstrapping is another method for applying a margin of error and was employed in a recent ancient DNA study (Wright et al., 2018), however, it is noted that this is a computationally intensive process and leads to substantially increased processing time.

Methods used in this research for calculating a confidence interval are discussed in Section 4.9.

3.3 Conclusion

In this chapter several key factors were identified that should be considered to achieve accurate BGA prediction. These factors were:

1. Population-level and Family-level Admixture: how certain individuals may be genetically related to two or more populations and the limitations this scenario can introduce;
2. Parsimony: selecting the minimum number of SNPs required to achieve accurate discrimination using information theory;
3. Relevant populations: ensuring that the samples collected for analysis are an adequate representation of the true populations of interest;
4. Sample size and rare events: the possibility that a genotype believed to be absent in a population is present as a rare event and whether the collected sample has detected this event;
5. Prior probability: are the populations of interest equally represented;
6. Degraded/Partial profile: how does missing data affect the classification;
7. Margin of error: the inclusion of a measure of precision to provide a level of confidence in obtained results.

Selecting a suitable approach for BGA prediction should be based on both the question of interest, and what is being measured. The next chapter will outline the methodology for each factor that was implemented into this thesis' KBDSS, with the associated justifications.

Chapter 4 – Materials and Methods

4.1 Introduction

This chapter outlines how the various factors regarding BGA prediction, as discussed in Chapter 3, were implemented in this thesis. Each analysis was performed using the case study described in Section 4.2. Finally, these methods will be combined and integrated to build the KBDSS known as DNA-MAP, which is discussed in greater detail in Chapter 6. This chapter will be structured as shown in Table 4.1.

Table 4.1: Implemented Methods for Addressing BGA Prediction Factors
A list of the seven BGA prediction factors that are tested in this chapter, with their respective method of testing.

Factors	Method
Relevant Population	Case Study
Admixture	Simulation Tool
Classifier	Logistic Model Tree, Generic Bayes and STRUCTURE
Rare event	Green and Young
Prior	User input, Bayes formula
Partial/degraded DNA	Systematic Removal of Markers

4.2 Materials

4.2.1 Case Study, Relevant Populations, DNA Panel

In this thesis, a case study will be used for developing and demonstrating the KBDSS developed. The case study used is the ongoing recovery of missing WWII Australian soldiers in the South-East Asia Pacific being carried out by UWC-A. Australia has been involved in numerous conflicts in different parts of the world for over a century. Table 4.2 summarises the most notable conflicts Australia has been involved in since becoming an independent Commonwealth from the British in 1901 (National Museum of Australia, n.d.). While the number of Australian soldiers who remain unrecovered from each area of conflict is not shown in Table 4.2, it is estimated that thousands are still missing (Unrecovered War Casualties – Army, n.d.).

Table 4.2: Australia's History of War Participations

List of significant conflicts Australian troops have been involved in since 1900 (Australian War Memorial, n.d.).

War Fought (Geographic regions)	*Timeline*	*Estimated Number of Australian Soldiers Involved*	*Estimated Number of Australian Soldiers Deceased*
Boer War (South Africa)	1899 – 1902	16, 175	251
Boxer Rebellion (China)	1900 – 1901	300 – 500	6
First World War (South-East Asia Pacific, Middle East, Europe)	1914 – 1918	416, 809	60, 000
Second World War (South-East Asia Pacific, Middle East, Europe)	1939 – 1945	1, 000, 000	27, 000
Occupation of Japan	1946 – 1951	16, 000	0
Korean War	1950 – 1953	17, 000	340
Malayan Emergency	1950 – 1960	7, 000	39
Indonesian Confrontation	1963 – 1966	3, 500	23
Vietnam War	1962 – 1975	60, 000	521
The First Gulf War (Iraq)	1990 – 1991	1800	0
Afghanistan	2001 – present	400	41
The Second Gulf War (Iraq)	2003 – 2009	2000	0

It is estimated that over 2000 Australian soldiers are currently unaccounted for in the Southeast Asia-Pacific region from WWII (Unrecovered War Casualties – Army, n.d.), the region that will be the focus of this thesis. Papua New Guinea (PNG) is just one country where large-scale battles took place, with the Kokoda Track being a notable geographic area where numerous engagements were fought. The two primary armies of interest in this thesis which fought in the Southeast Asia-Pacific during WWII were the Australian and Japanese. Additional nations involved were the North Americans, Chinese, British, and New Zealand military forces and the local populations of Papuans, and New Guineans (Australian Government – Department of Veteran Affairs, 2009). Figure 4.1 outlines a detailed map of the Kokoda Track, where significant casualties were suffered by all participants in areas such as Isurava, Sanananda, and Buna (Jackson, 2019).

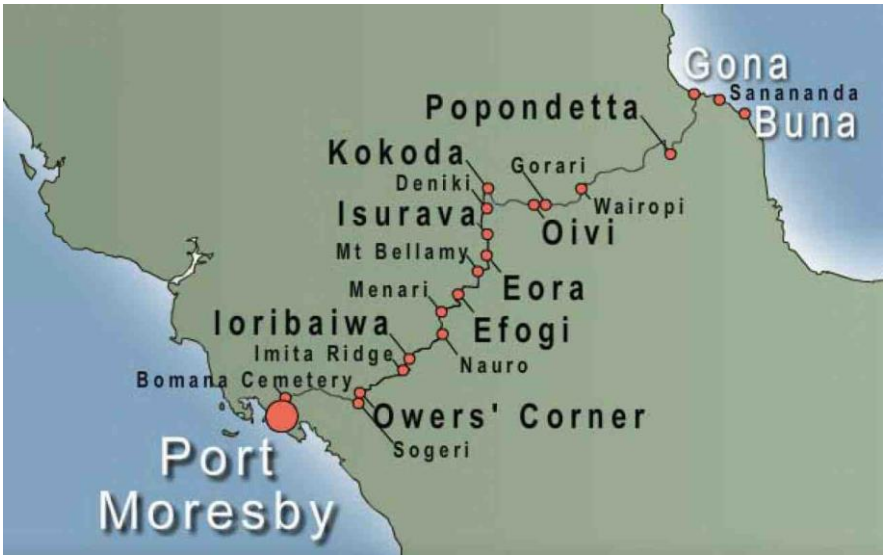


Figure 4.1: Kokoda Track
Detailed overview of the Kokoda Track, obtained from https://anzacportal.dva.gov.au/history/conflicts/kokoda-track/kokoda-track/about-kokoda-track-1942-and-today

To assist with the recovery of thousands of unaccounted for Australian soldiers from historical conflicts throughout the world, UWC-A, was formed. Their task is to investigate areas where it is believed there may be the remains of Australian soldiers (Unrecovered War Casualties – Army, 2012). If remains are recovered, identification is attempted with three possible outcomes (Figure 4.2):

1. Complete Identification: remains are assigned to a specific individual, e.g. "*M. Madge MM, 2/1 Field Regiment R.A.A, 7th June 1944 Age 43*"
2. Partial Identification (ancestry): remains are assigned to one of the populations of interest from the given geographical location, e.g. "*An Australian soldier of the 1939 – 45 war*"
3. No Identification: remains are assigned to a time period based on the given geographical location, e.g. "*A soldier of the 1939 – 45 war, known only to God*".

Figure 4.2: Military Burials
Three outcomes of UWC-A recovery operations: 1). Complete identification; 2). Partial identification (ancestry); 3). No identification. Images provided by Felicity Poulsen – Personal Communication, 2017.

While identification is the desired goal, in real casework success is hindered by various factors including:

(i) *Lack of family reference samples*: To determine if a set of remains belong to an individual in modern missing persons cases, a known sample is required for comparison; this sample could belong to the individual or to a close family member (such as a sibling or parent). A constraint of UWC-A is that such samples are not available, so distant maternal or paternal relatives are sought.

(ii) *Environmental*: Often remains are subject to harsh environmental conditions such as prolonged exposure to ultraviolet light, heat, moisture, microbes and scavengers (Figure 4.3). These circumstances can cause DNA degradation and reduce the amount of information produced by genetic testing.

Figure 4.3: UWC-A Investigators
UWC-A investigators excavating a set of skeletal remains from a WWII soldier, in PNG. Image taken from https://www.army.gov.au/our-work/unrecovered-war-casualties/world-war-two-papua-and-new-guinea.

Accurate methods for ancestry prediction is the first step required by UWC-A to facilitate decisions about the final resting place of each soldier. For DNA-MAP to be implemented into UWC-A casework, the system needs to be tailored to their case work needs. Thus, in the case of PNG it requires: (*i*) a method of ancestry prediction for distinguishing between two populations of interest with a high degree of accuracy (*ii*) a clear explanation of assumptions and limitations, and (*iii*) a report of the results designed for the intended end-user. The main user of DNA-MAP is UWC-A forensic biologists. UWC-A investigators and the Identification Decision Board will use the conclusions and opinions of the forensic biologists and can be considered as 'consumers' of the information. Forensic biologists will input data on a case-by-case basis in order to infer potential BGA for a set of unknown remains. When creating and calibrating DNA-MAP, the end user and report consumers will be kept as the focus, particularly when developing the areas of reporting language and user-interface options.

The importance of the work UWC-A performs is in providing closure to the families of the service personnel who lost their lives, and in ensuring that soldiers who fought for their country receive a respectful burial. "*One of the most important reasons to identify unknown persons is because non-identification may result in numerous issues at emotional and legal level for the surviving family members and friends.*" (Beauthier et al., 2009, p.54). As stated in the honouring of World War 1 soldier Private Thullier Lake Cardew, "*Identifying Private Cardew and honouring him with a headstone that bears his name is one small way we honour every man and woman who serves in defence of our nation*" (Tehan, 2017, p.1)

An important aspect of this research is the distinction between the WWII era populations and their respective present-day counterparts. Prior to WWII, Australia was exposed to numerous waves of immigration by individuals from Asian populations, notably, Chinese and Japanese (Parliament of Australia, n.d.). The significance of noting all Asian immigration, rather than Japanese alone, is that the proposed classification approach in this thesis works on a binary basis; for an unknown set of remains the probabilities of belonging to the Australian population and to the Japanese population will be assessed and compared. An individual with Asian ancestry (regardless of which specific Asian country) will still be assigned to one of these two populations and is expected to be more likely associated with the Japanese population based on the genetic similarity of these geographical regions.

Post-WWII, the contemporary populations of Australia and Japan have been affected by a degree of multicultural influence, especially in Australia. The influence can be seen in Census data provided by the Australian Bureau of Statistics (ABS). During the Census of 1933, the last one prior to WWII, approximately 8000 (0.12%) of the total 6,629,839 censored individuals in Australia nominated themselves as Chinese, and 2000 (0.03%) nominated as Japanese ancestry (Australian Bureau of Statistics, 1933). In the 2016 Census, with a total population of 24,130,000 individuals, these equivalent figures were approximately 1.18 million (≈5%) and 41,000 (≈0.2), respectively (Australian Bureau of Statistics, 2017). The increases in both Asian ancestries within the current Australian population indicate that present-day Australia may not accurately represent its WWII era counterpart. Samples which will be used to represent Australian WWII era soldiers must be collected carefully based on appropriate criteria. The specification of European/British as the common ancestry for Australian soldiers arises from the Defence Act of 1909 which stated that individuals ‘not substantially of European origin or descent’ were exempted from enlisting (Australian Government, 1909). Additionally, historical data from the Australian Bureau of Statistics (1933) indicates that British nationality was the highest ancestry self-declared by Australian individuals in the 1933 Census, approximately 6.5 million out of the total 6.630 million (98%). Note that there were exceptions, Australian soldiers with non-British ancestry, and although the choice of these two populations for this thesis is a limitation, it is done so to demonstrate a proof of concept.

To ensure the data utilised in this thesis is representative of Australian WWII era soldiers, samples need to be taken from:

1) Individuals who could have been alive during the WWII period, as these individuals are the same generation as the Australian WWII era soldiers;
2) Direct descendants of individuals from the Australian WWII era generation, since the offspring will inherit genetic profiles similar to those of the soldiers with minor variation.

Using these two criteria, the timeframe of birth dates for including Australian individuals who self-declared European/British ancestry is "1918 to 1939". Collection of samples representing Australian WWII era soldiers and generation of the relevant DNA profiles were not performed as part of this thesis; this was conducted by Ghaiyed (2020).

The data used in this case study was provided by two sources:

1) Individuals who met the criteria to be classified as a WWII era Australian – provided by Ghaiyed (2020);
2) Contemporary Japanese individuals were collected from the publicly available 1000 Genomes Project online database (Genomes Project Consortium, 2015).

From these sources, the following data was used in this thesis:

1) Complete profiles of WWII era Australians (n = 108) – Provided by Ghaiyed;
2) Partial profiles of WWII era Australians (n = 80) – Provided by Ghaiyed;
3) Contemporary Japanese individuals (n = 104) – Collected from the 1000 Genomes Project.

As WWII era Japanese data was not available, contemporary data was utilised as a proxy. The ancestry panel used in this thesis, Ghaiyed Population Specific Panel (GPSP) (Ghaiyed, 2020) is comprised of 45 autosomal SNPs that were selected for their ability to differentiate between Australian and Japanese individuals. Note that the experiments described in this thesis only utilise 40 of the original 45 SNPs, with five SNPs being removed due to genotypes not being obtained for any individuals. The process of how these SNPs were selected is not discussed in this thesis, the reader is referred to Ghaiyed (2020) for full details. Summaries of the set of SNPs with their corresponding allele and genotype frequencies for each population can be found in Appendix 1 and Appendix 2, respectively. Note that due to the small sample sizes, test data will be simulated based on the frequencies obtained from the original training data to avoid overfitting the model. While it would be ideal to have completely independent test data, this is not available and therefore, simulation is the next option to mitigate the chance of overfitting. It is also noted that the number of variables can

also impact the possibility of overfitting, wherein that, the inclusion of each variable requires an increased number of observations to reduce the chance of overfitting the data based on a small sample size being used in conjunction with many variables.

4.3 Methods – Parsimonious Logistic Model Tree (*p*LMT)

4.3.1 Experimental Overview

The following section outlines the experimental process used for developing, executing and testing the *p*LMT classifier in this thesis. Following sections will expand on each of the following steps.

1) Data Input. The relevant population datasets are uploaded.

2) LMT Generation. Using the relevant datasets established in (1), the LMTs are generated in a parsimonious way to provide an overall estimate of the Australian membership probability (a geometric mean, which averages estimates from each iterated model), that is, the probability that the individual belongs to the Australian population. From this point onwards, this estimate will be collectively referred to as the Geometric Mean of Australian Membership Probabilities (GMAMP).

3) Simulate Known Data. To test the effectiveness of a classifier, known individuals from the two major populations of interest (Australia and Japan) and individuals with known levels of admixture of these two populations are required. However, as individuals with known levels of admixture are unavailable, and the only known non-admixed individuals are those from the original population datasets used in Step 1 (Data Input), samples of varying degrees of admixture will be simulated to provide testing data. The inclusion of admixture scenarios is to observe the accuracy of the classifier for complex cases and to determine the point at which the classifier can no longer accurately discriminate between the two populations of interest. Relevant admixture scenarios are defined using appropriate ancestors, and the admixture simulation tool, SimAdmixtR (Kennedy (2019), accessible at https://dkenn.shinyapps .io/ww2-admixture/). For each ancestral scenario the tool is used to simulate 10,000 individuals. SimAdmixtR is initially validated by comparing theoretical genotype estimates with the observed simulated genotype estimates.

4) Analyse Simulated Data and Establish Classification Thresholds. The simulated individuals from each scenario used in (3) are submitted to the models generated in (2) to estimate the Australian membership probability of being an Australian for all individuals within each scenario. The distributions of the resulting probabilities are examined using boxplots to determine the points where one can no longer confidently assign simulated individuals to the major populations (Australian and Japanese). From these graphs, thresholds are defined to be used in establishing classification

guidelines for unknown samples. These thresholds establish the points at which ancestry can confidently be assigned, and at what stage the outcome is ambiguous. The threshold for each of the two major populations are selected to represent approximately a 90% success rate for either Australian or Japanese ancestry assignment.

5) Validate Classification Thresholds. A second set of simulated individuals are created for the scenarios in (3). These second group of simulated individuals are submitted to the models generated in (2) to estimate the Australian membership probability for all individuals. These individuals are then classified using the thresholds previously established in (4). The resulting estimated ancestry is compared with actual ancestry which is known from the information used to generate the simulated individuals. Two errors are considered, 'direct error' where an individual is assigned to an incorrect ancestry, and 'indirect error' where an individual with ambiguous ancestry is assigned to a specific population group.

4.3.2 *p*LMT Algorithm: Data Input

Datasets are uploaded from the two relevant populations. The dataset consists of the raw genotype values for all individuals in the sample for the complete set of SNPs in the DNA panel. The panel used is Ghaiyed's (2020) Population Specific Panel, comprised of 40 biallelic SNPs selected for discriminating between Australia and Japan.

The WWII era Australians (n = 108) and contemporary Japanese individuals (n = 104) described in Section 4.2 are used as the population datasets.

4.3.3 *p*LMT Algorithm: LMT Generation

The Parsimonious Logistic Model Tree's methodology is as follows:

1. Relevant population data is uploaded as the working dataset which contains the total number of panel SNPs, S.
2. An LMT is generated using the dataset from (1). The resulting selected SNPs, $\{s_m\}$, in the model, together with their coefficients are recorded as "*Model m*". The LMT is fitted using 10-fold cross-validation, repeated ten times, with the accuracy across each run averaged and recorded as the model's average accuracy. Note that due to the model's accuracy being used as the stopping criterion for model generation, it is imperative to ensure the process is robust. Minor variations between a model's accuracy between runs on the same data may result in a model being accepted in one run but not in another. To avoid this inconsistency, multiple runs of cross-validation are performed, and the averaged accuracy is used as the criterion. This technique was also implemented by Landwehr et al. (2005) when experimenting with the

original LMT algorithm. If the average classification accuracy of Model *i* is below a specified threshold (default 99%), the *p*LMT proceeds to step 4. If the accuracy is greater than or equal to the specified threshold (default 99%), the algorithm proceeds to step 3.

3. The SNPs, $\{s_m\}$, are removed from the working dataset and, providing there are remaining SNPs in the working dataset, the algorithm repeats step 2 generating a new model using the SNPs remaining in the working dataset after those from the previous model are removed. If no SNPs remain in the working dataset the algorithm proceeds to step 4.
4. Once accuracy falls below the specified threshold (default 99%), or if all available SNPs are utilised, the *p*LMT algorithm stops, and *M* models are obtained (where *M* is ≥1). Note, the model whose accuracy falls below 99% is excluded from the *M* models.

The output of the *p*LMT algorithm is a set of *M* models, where each model contains a subset of SNPs and their corresponding coefficients. The output from each LMT regression model gives an estimate of the logarithm of the odds ratio between memberships of the individual in the two populations. This estimate is then used to obtain the predicted membership probability for the required population of interest, in this situation, Australian. Equation 4.1, based on Landwehr et al. (2005, p.18), provides the probability membership for the Australian population (calculated for each of the models), where $F_k(x)$ is the outputted regression function for the *k*th population (*k* being either Australian or Japanese):

$$p_{Australian} = \frac{e^{F_{Australian}(x)}}{e^{F_{Australian}(x)} + e^{F_{Japanese}(x)}} \tag{4.1}$$

The output of the *p*LMT at this stage is a series of predicted probabilities of population membership, $p_{Australian,m}$, for m = 1, 2, …, *M*, for the given individual. These independent estimates are then combined into a single value, to give a final estimate of the probability of ancestry which will be provided to the end user. The combined population probability is calculated using the geometric mean which was selected over the standard arithmetic mean as outliers have a reduced effect on the geometric mean (Manikandan, 2011) (Equation 4.2).

$$Geometric\ Mean = \left(\prod_{m=1}^{M} p_m\right)^{\frac{1}{M}} \tag{4.2}$$

4.3.4 *p*LMT Algorithm: Simulate Known Data

The "SimAdmixtR" tool created by Dr. Daniel Kennedy (Kennedy, 2019) utilises SNP allele frequencies from two populations to create the genetic profiles of simulated individuals based on a

nominated pedigree. Figure 4.4 shows an example of an admixed pedigree, an Australian individual with a Japanese great-grandparent. Simulated data will be used to provide a large sample size for subsequent experiments, as the original datasets are quite small, consisting of approximately 100 individuals.

The Admixture Simulation tool was created as a package for the statistical software R (R Core Team, 2019) called "SimAdmixtR" (Kennedy, 2019) and is currently available online as a user-friendly Shiny application (https://dkenn.shinyapps.io/ww2-admixture/).

It is acknowledged that the Australian sample may have admixture already present, however, as the sample size available is small (n = 108), it may not represent the true measure of admixture in the population. To accommodate for this possibility, individuals of known degrees of admixture are simulated. The pedigrees of interest for this thesis are based on four generations, going back to great-grandparents. The selection of four generations was made based on recommendations from previous UWC-A research (Ghaiyed, 2016; Poulsen, 2015), and the expectation that admixture beyond four generations will be washed-out.

To simulate individuals using SimAdmixtR three files are required: the allele frequencies data file, the simulation details file, and an example of the required STRUCTURE input. A description of each file and a summary of the Admixture Tool's process is described in Appendix 3.

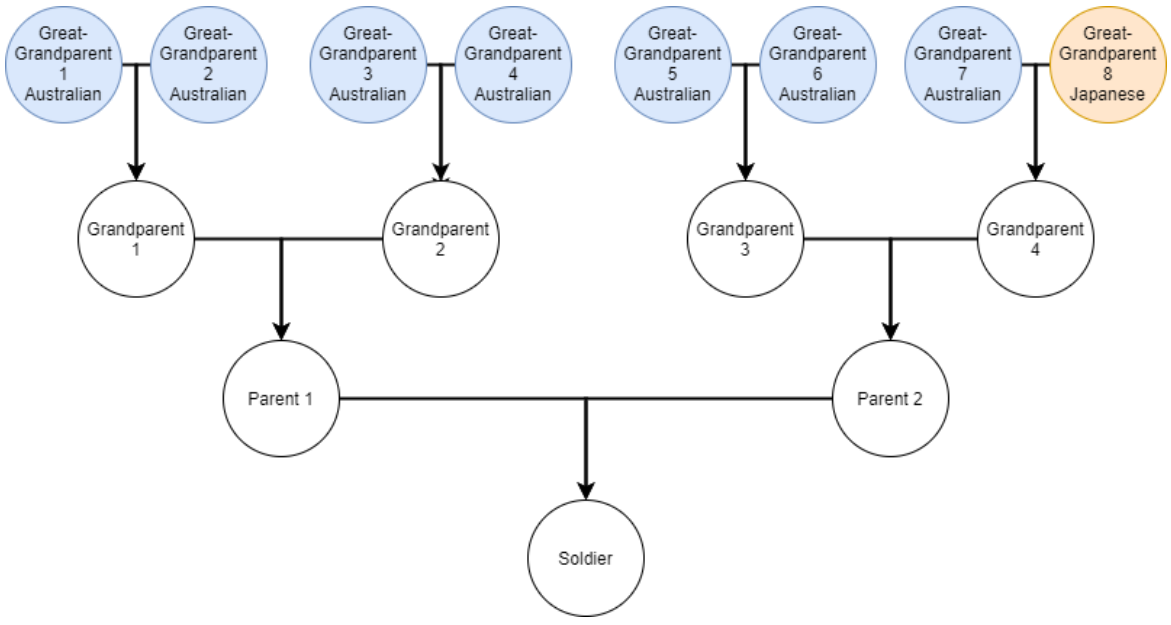


Figure 4.4 Simulated Pedigree
Example pedigree of an Australian soldier with a single Japanese great-grandparent (orange), providing approximately 1/8th of the soldier's ancestry.

Ten scenarios of admixture are simulated in this thesis, with n = 10,000 individuals simulated in each scenario. An *ad hoc* value of 10,000 was selected as it is estimated to be a large enough number to ensure robust inferences but also to not be prohibitively computationally slow. The following data are used to estimate input allele frequencies: (*i*) WWII era Australians (n = 108), and (*ii*) Contemporary Japanese individuals with (n = 104). The scenarios are detailed in Table 4.3.

Two independent groups of simulated individuals are created and used at different stages of the development: Simulation Group 1, consisting of seven scenarios (1, 2, 4, 6, 8, 9, 10), and Simulation Group 2, which consists of all ten scenarios. In each scenario for each Group, 10,000 individuals are simulated giving a total number of 70,000 individuals for Group 1 and a completely different set of 100,000 individuals for Group 2.

Table 4.3: Admixture Scenario
Simulated admixture scenarios with their respective scenario ID for subsequent analyses. Each scenario is simulated with n = 10,000 observations. The correct interpretation of Australian or Japanese for a scenario's key was selected based on whether the scenario consisted of primarily (≥75%) Australian or Japanese ancestors, scenarios which do not meet this criterion are to be interpreted as ambiguous. The scenario ID indicates the respective ratio of Australian to Japanese ancestors' proportion, that is, what percentage of ancestors at the great-grandparent level belong to each population. Note that the subscript of *a* and *b* on Scenarios 3 and 4 is simply to distinguish between these two scenarios which share an approximately equal average pedigree proportions.

Scenario #	Scenario ID	Admixture Scenario	Representing Pedigree Proportions			Simulation Group
			Australian Proportion	Japanese Proportion	Correct Interpretation	
1	100/0	An individual with all Australian ancestors	100%	0%	Australian	1 and 2
2	87.2/12.5	An individual with one Japanese great-grandparent	87.5%	12.5%	Australian	1 and 2
3	75/25a	An individual with a Japanese great-grandparent on both the maternal and paternal lineage	75%	25%	Australian	2
4	75/25b	An individual with a Japanese grandparent	75%	25%	Australian	1 and 2
5	62.5/37.5	An individual with one Japanese grandparent and one Japanese great-grandparent	62.5%	37.5%	Ambiguous	2
6	50/50	An individual with an Australian parent and a Japanese parent	50%	50%	Ambiguous	1 and 2
7	45/55	An individual with a Japanese parent and Japanese great-grandparent	45%	55%	Ambiguous	2
8	25/75	An individual with one Australian grandparent	25%	75%	Japanese	1 and 2
9	12.5/87.5	An individual with one Australian great-grandparent	12.5%	87.5%	Japanese	1 and 2
10	0/100	An individual with all Japanese ancestors	0%	100%	Japanese	1 and 2

In a binary context such as applies in the case study, based on conservative expectations, scenarios 1, 2, 3 and 4 will produce individuals who should be assigned Australian ancestry, scenarios 5, 6 and 7 will produce individuals who cannot be distinguished into either population, and scenarios 8, 9 and 10 will produce individuals who should be assigned Japanese ancestry. The choice of assigning an ambiguous outcome to scenarios between 25% and 75% admixture from a second population was *ad hoc*. The reasoning behind the selection is as follows: an individual with no more than one grandparent (that is 25%) deviating from the common ancestry shared by the remaining lineage could be described as consisting primarily of a single ancestry. However, for cases where more than one

grandparent deviates from the commonality (that is, greater than 25%) the individual could instead be described as a mixture of the two resulting ancestries.

Validation of the software, SimAdmixtR, was carried out by comparing the observed genotype frequencies obtained from each admixture scenario with the expected genotype frequencies as estimated algebraically using Mendelian inheritance calculations. This validation is to test the assumption that the SimAdmixtR software is correctly simulating Mendelian inheritance. Note that the simulations assume that the observed genotype frequencies accurately represent the true population values.

4.3.5 *p*LMT: Analyse Simulated Data and Establish Thresholds

The individuals in Simulation Group 1 (seven scenarios: 1, 2, 4, 6, 8, 9 and 10, n = 10,000 per scenario) are processed through the M models generated from the original data using the *p*LMT algorithm described in Section 4.3.3. The distribution of the GMAMP for all seven scenarios are calculated and viewed using boxplots. Thresholds of classification are established by viewing the boxplots to determine where approximately 90% of the Australian and approximately 90% of the Japanese data lie. The area in between is regarded as Ambiguous and is where neither ancestry can confidently be defined. Note that this phase indirectly identifies the level of admixture where predicting ancestry is not possible.

4.3.6 *p*LMT Algorithm: Validate Thresholds

Simulation Group 2, as described in Section 4.3.4, are submitted to the same M models created in Section 4.3.3, and the GMAMP is estimated for individuals in all ten scenarios. The individuals are classified by comparing their calculated GMAMP with the thresholds established in Section 4.3.5. A misclassification rate is then estimated, using two classes of an erroneous classification:

i) Direct Error – Occurs when an individual who should be assigned as Australian (scenarios 1 – 4, Table 4.3) or Japanese (scenarios 8 – 10, Table 4.3), is assigned to the incorrect population;

ii) Indirect Error – Occurs when an individual from scenarios 5 – 7 (Table 4.3) is assigned to one of the populations instead of being left unassigned.

The direct error occurs when an individual whose pedigree consists mainly of ancestors belonging to one population (≥75%) is misclassified to the wrong population. The indirect error, however, occurs when the ancestry could belong to either population, that is, it is in one of the unassignable scenarios (5 – 7), but the individual is assigned to a population. From a genetic perspective, an Australian soldier with 50/50 admixture and a Japanese soldier with 50/50 admixture are indistinguishable. The

difference in the context of this case study, is that they are soldiers who fought for different armies. Hence it is a historical reason which makes them different not a genetic one, therefore, assignment based on their genetic profile cannot accurately be determined, however, may still provide useful intelligence to the user regarding possible family-level admixture. The relationship between genetic admixture and assignment of ancestry is discussed in Section 7.1.2.

4.4 Methods – Generic Bayesian

4.4.1 Experimental Overview

The following section outlines the proposed experimental process for performing and testing the Generic Bayesian classifier as used in this thesis.

1) Data Input. See Section 4.3

2) Simulate Known Data. See Section 4.3.

3) Analyse Simulated Data and Establish Thresholds. The natural log of the likelihood ratio is estimated for all simulated individuals and thresholds are established using the same methodology outlined in Section 4.3, but with the variable used to establish the threshold being the log likelihood ratio instead of the Geometric Mean of the Australian/Japanese membership probability.

4) Validate Thresholds. See Section 4.3. Again, the variable used is the log likelihood ratio instead of the Geometric Mean Australian/Japanese membership.

4.4.2 Generic Bayesian: Data Input

See Section 4.3.

4.4.3 Generic Bayesian: Simulate Known Data

The same two groups of simulated individuals as are discussed in Section 4.3.4 are used for this analysis. For each scenario in each group, 10,000 individuals are simulated. Simulation Group 1 consists of seven scenarios (1, 2, 4, 6, 8, 9, 10) giving a total number of 70,000 simulated individuals. Simulation Group 2 consists of all ten scenarios giving a total number of 100,000 individuals. The scenarios are defined in Table 4.3.

4.4.4 Generic Bayesian: Analyse Simulated Data and Establish Thresholds

To estimate the LR the probability of observing the genotype for each SNP in each population k, s_k, seen in the simulated individual's profile is recorded. The probability of observing the simulated individual's complete genotype at all SNPs, S, in the kth population is:

$$\Pr\left(Genotype_S|Population\ k\right) = \prod_{s=1}^{S} p_{sk} \quad (4.3)$$

Where p_{sk} is the relative sample frequency of observing the individual's genotype on the *s*th SNP in the *k*th population. An individual will have two probabilities, Pr $(Genotype_S|Population\ k)$, one for the frequency of observing that genotype in each of the populations of interest. If $p_{sk} = 0$, then the zero probability is replaced using the conservative replacement formula, Equation 4.4, as proposed in Phillips et al. (2007), where *n* is the number of individuals. Note the formula has been adjusted to be *n* instead of 2*n* as originally proposed in Phillips et al. (2007) as interest here lies in genotypes, not alleles.

$$Conservative\ Replacement = \frac{1}{n+1} \quad (4.4)$$

The likelihood ratio is then calculated using Equation 4.15;

$$\mathrm{LR}_i = \frac{\Pr\left(Genotype_S|Population\ A\right)}{\Pr\left(Genotype_S|Population\ B\right)} \quad (4.5)$$

The likelihood ratio is estimated for all individuals in Simulation Group 1 (seven scenarios, *n* = 10,000 per scenario).

Thresholds are established by viewing the boxplots of the resulting LRs to determine where approximately 90% of the Australian and approximately 90% of the Japanese individuals lie. Note that the relative frequencies used in Equation 4.3 are estimates based on the samples used, therefore, these values are subject to sampling error, which is explored in Section 4.9.

4.4.5 Generic Bayesian: Validate Thresholds

The second Group of simulated individuals as described in Section 4.3.6 is again used for validation of the LR thresholds.

4.5 Methods – STRUCTURE

4.5.1 Experimental Overview

The following section outlines the proposed experimental process used in this thesis for performing and testing the STRUCTURE classifier.

1) Data Input. See Section 4.3.1 point 1.

2) Simulate Known Data. See Section 4.3.4.

3) Analyse Simulated Data and Establish Thresholds. The membership proportion, *Q*, for the Australian population is estimated by STRUCTURE for all simulated individuals. Thresholds are established using the same methodology outlined in Section 4.3.5 with the *Q* value from STRUCTURE used in place of the GMAMP obtained from the *p*LMT.

4) Validate Thresholds. See Section 4.3.6 with the *Q* value from STRUCTURE used in place of the GMAMP obtained from the *p*LMT.

4.5.2 STRUCTURE: Data Input

See Section 4.3.1 point 1.

4.5.3 STRUCTURE: Simulate Known Data

The same two groups of simulated individuals as discussed in Section 4.3.4 are used for this analysis. For each scenario in each group, 10,000 individuals are simulated. Simulation Group 1 consists of seven scenarios (1, 2, 4, 6, 8, 9, 10) giving a total number of 70,000 simulated individuals. Simulation Group 2 consists of all ten scenarios giving a total number of 100,000 individuals. The scenarios are defined in Table 4.3.

4.5.4 STRUCTURE: Analyse Simulated Data and Establish Thresholds

The following parameters were used to analyse the simulated data using the STRUCTURE program. A total of 10,000 Markov Chain Monte Carlo (MCMC) replicates are performed following an initial burn-in period of 10,000 replicates using the Admixture model. *K*, the number of populations present in the training data, is set at two as the number of populations present is known. The membership proportion (*Q* value) for the Australian population is estimated and recorded for each individual in the test data.

Thresholds are established by viewing the boxplots of the distribution of the resulting *Q* values to determine where approximately 90% of the Australian and approximately 90% of the Japanese individuals lie.

4.5.5 STRUCTURE: Validate Thresholds

The second Group of simulated individuals as described in Section 4.3.6 is again used for validation of the Australian membership proportion thresholds.

4.6 Classifier Comparison on Degraded Samples

To observe how the *p*LMT, Generic Bayesian and STRUCTURE classifiers handle degraded remains the classifiers are each applied to a WWII era Australian sample with missing data (n = 80) in which individuals have between 10 and 39 SNPs out of the original 40 SNPs available.

For the *p*LMT classifier, if any SNPs are not available for the unknown individual, the corresponding SNPs are removed from the original population dataset before the analysis to develop the initial models. This removal of SNPs ensures that any subsequent statistical modelling is derived only from the subset of SNPs that are present in the genotype of the unknown individual. The genotype from the unknown individual becomes the driving force of the model, allowing the classification framework to adapt on a case-by-case basis to handle missing data, a technique previously unused in BGA prediction classifiers. STRUCTURE ignores missing data when calculating *Q* values, to ensure the resulting model is based purely on what information is available (Pritchard et al., 2009).

For the Generic Bayesian classifier, if a SNP is not available for the unknown individual, the relative probabilities are replaced with a value of one. This replacement ensures that the missing SNP does not affect the resulting likelihood ratio as it cannot be known what the true genotype was. Note that this is an additional approach used in conjunction with the minimum allele frequency, where values of zeros for genotypes not observed in the training sample are replaced with a conservative frequency based on the sample size.

Each of the classifier's estimated outputs (GMAMP, likelihood ratio and *Q* value of Australian ancestry) for the n = 75 degraded WWII Australian samples are plotted against the number of SNPs available for that individual. A linear regression is performed, using Microsoft Excel, to explore the effect the number of SNPs may have on the resulting classifier's output. A classifier's weakness for accurately classifying degraded samples will result from two factors, (*i*) if a linear decline in the estimated probability of Australian ancestry occurs as the number of SNPs decrease and (*ii*) if the known degraded Australian samples are assigned as ambiguous or misclassified as Japanese. Classifications for factor (*ii*) occur using the relevant thresholds established for each classifier in Sections 5.2.3, 5.3.1 and 5.4.1.

The degraded samples are then categorised using the thresholds established in each classifier's respective section, to observe whether one classifier outperformed the other. Performance level is determined by: (*i*) the number of individuals correctly classified as Australian as opposed to Ambiguous, and (*ii*) the number of individuals incorrectly classified as Japanese. The purpose of this

experiment will be to observe any trends resulting from the loss of SNPs, that is, how will each classifier perform for different numbers of missing SNPs?

4.7 SNP Removal Experiment

Due to degradation and other possible stochastic errors, it is likely that casework samples will not obtain a complete SNP panel profile. In keeping with the concept of information theory discussed in Section 3.2.2, it is important to determine the minimum number of SNPs required to still achieve the same accuracy that would be obtained if the complete panel were available. To determine the minimum number of SNPs needed, on average, to generate a credible *p*LMT, SNPs are randomly removed using RStudio (R Core Team, 2019). For Simulation Group 2's scenarios of 100% Australian and 100% Japanese (n = 10,000 each), SNPs are removed in groups of 5, 10, 15, 20, 25, 30, and 35, with each subset being selected randomly. Each grouping is repeated independently 100 times. For each of the 700 replications, a *p*LMT is computed using the remaining SNPs, giving 700 different *p*LMTs. For each *p*LMT, the GMAMP is computed for the 20,000 individuals who are then categorised based on the previously defined thresholds. Individuals in the ambiguous classification are then omitted as they provide no interest in specific ancestry determination. For the individuals classified as either Australian or Japanese two possible outcomes are recorded, correct or a direct error. Note that there is no possibility of indirect error as only non-admixed individuals are used. For each grouping, the minimum, mean and maximum numbers of correct and direct error for each of the Australian and Japanese populations are estimated. The results will be reviewed to identify the minimum number of SNPs that are required to still retain a valid result.

4.8 The Effect of the Prior

At this stage of analysis, the output of the *p*LMT is a sample estimate of an unknown individual's GMAMP. This GMAMP is obtained with the assumption that the two populations are approximately equally represented in the combined sample used for analysis and in the true populations being considered. However, this may not be the case and there may be additional, pertinent information that needs to be incorporated. In particular, there may be genuine knowledge of the expected probability of unknown remains being Australian before any DNA is measured. That is, there may be information that allows a sound estimate of the prior odds. For example. UWC-A may have historical records which show that in the battlefield of interest there are 10 unaccounted for Australian soldiers and 1000 Japanese. Thus, without considering any further information, the probability that a sample found in this area is from an Australian soldier is 10/1010 (0.0099), giving a prior odds ratio of 10/1010 divided by 1000/1010 (0.0099/0.9901) which is 0.01. However, suppose the DNA profile seen in the recovered sample is one which all 10 of the Australians have, but which is not seen in any

of the Japanese. Clearly this would change the probability of the remains being from an Australian. The hope is that by taking into consideration other information as well as the DNA and including it in the final analysis, it is possible to get a better estimate especially if that estimate can carry some measured form of confidence, that is, a margin of error.

To combine the prior odds with the GMAMP estimate from the *p*LMT, Bayes' theorem is proposed, however, first it will be necessary to determine the probability of obtaining such a geometric mean in each of the two populations, Australian and Japanese. Once there are estimates for these values, a likelihood ratio can be established and used in conjunction with the prior odds to obtain a posterior estimate of Australian membership that does not assume equal representation in the analysis. Note that the exact same value of the estimated GMAMP could be obtained by: (*i*) individuals having the exact same genotype as the original sample, or (*ii*) individuals whose genotype causes the utilised regression coefficients to estimate the same GMAMP by chance.

To estimate these equivalent variables, an Empirical Cumulative Distribution Function (ECDF) of estimated GMAMP is obtained by using the individuals from the simulated scenarios, 100% Australian (*n* = 10,000) and 100% Japanese (*n* = 10,000), from Simulation Group 2. An area in the ECDF of ±0.025 around the unknown sample's GMAMP (*u*GMAMP) will be used to obtain the probability of seeing this particular geometric mean in each of the populations. A likelihood ratio will be built from these probabilities:

$$LR = \frac{\Pr\left(uGMAMP|Australian\right)}{\Pr\left(uGMAMP|Japanese\right)}$$

Using Bayes' theorem, the posterior probability of an Australian (Equation 4.6) given this general mean can be written as:

$$Posterior\ Probability\ (Australian|UGM) = \frac{1}{1 + \frac{1}{LR} * \frac{1}{Prior\ Odds}} \qquad (4.6)$$

If one of the probabilities in the LR is zero, that is, no individual in the sample is observed to have an approximately similar GMAMP, the LR cannot be calculated. This zero may occur because there genuinely is no possibility of a profile in the population generating such a UGM, or it may simply be that the original dataset did not contain the SNP structure required to get the GMAMP of interest as the profile required represents a rare event in the population which could not be picked up in the sample size used. In the Generic Bayesian classifier, this issue is dealt with by replacing a zero frequency with a conservative value derived using ad-hoc methods, namely one. Rather than draw on

this or some similar ad-hoc method, it is acknowledged that rare events typically follow a Poisson distribution, and the approach taken here is to follow the methodology of Green and Young (1993) which draws on this statistical distribution to obtain Equation 4.7. In Equation 4.7, which was developed to estimate the sample size needed for a given situation, *m* is the probability of the rare event occurring, and β is the probability that the rare event will not be seen in the sample when it really is present in the population, meaning that $(1-\beta)$ is the 'confidence' that the event will be seen in the sample if it is really is in the population.

$$n = -\frac{\ln(\beta)}{m} \tag{4.7}$$

Equation 4.7 can be re-arranged to solve for *m* to determine the rarest event that could be detected with reliability of $(1-\beta)$ for a given sample size. The result is given in Equation 4.8, where *n* is the sample size used in the original dataset.

$$m = \frac{-\ln(\beta)}{n} \tag{4.8}$$

For the case of 95% 'confidence', Equation 4.8 can be simplified to:

$$m = \frac{-\ln(0.05)}{n}$$

Equation 4.8 can therefore, be used as a conservative frequency estimation method to replace a value of zero in the previous LR. If a UGM was not detected in the available sample, one possible reason is that the profile required to obtain said UGM is a rare event in the population, therefore, Equation 4.8 can be used to estimate a conservative frequency for observing the UGM that is also dependent on the available sample size. By using Equation 4.8, the calculations become such that for analyses with a large sample size available, one can have a high degree of confidence that the UGM is extremely unlikely to occur in the population, which is then reflected in the resulting LR.

To observe the effect of the prior odds ratio on the resulting posterior probability of Australian ancestry for the *p*LMT classifier, a sensitivity analysis was performed where the variables tested were: (*i*) the prior odds ratio, (*ii*) the sample size used in the original data, and (*iii*) the GMAMP. Values for the prior odds ratio were chosen using two criteria: (*i*) values that are expected for UWC-A, and (*ii*) generic values which may be relevant in most cases; these values are shown in Table 4.4. In addition to these prior odd values, the following sample sizes were selected: 100, 200, 300, 400 and 500, alongside the following GMAMP values: 0.01, 0.9, 0.95 and 0.99.

It is noted that Budowle et al. (2011) state that prior odds obtained through war manifests may not be entirely reliable. Therefore, a sensitivity analysis testing the effect of the prior odds is important to demonstrate the care that should be taken when estimating a value for the prior odds. It will also allow decision makers to decide whether or not to proceed in an area where the expected prior is such that no amount of further DNA analysis could result in a change to the probability associated with unknown remains. Such a decision could be invaluable in allocating resources to areas where a result is a real possibility.

Table 4.4: Prior Odds Ratio Values
Values for the prior odds ratio variable used in the sensitivity analysis, with the respective English statement describing the value's corresponding scenario.

Prior odds Ratio Value	Respective English Statement
0.5	The ratio of individuals from population 1 to population 2 is 1:2
0.3	The ratio of individuals from population 1 to population 2 is approximately 1:4*
0.1	The ratio of individuals from population 1 to population 2 is 1:10
0.05[a]	The ratio of individuals from population 1 to population 2 is 1:20
0.01[b]	The ratio of individuals from population 1 to population 2 is 1:100

*Rounded up from $1:3.\overline{33}$. [a] This prior odds ratio value is expected in PNG areas such as Buna. [b] This prior odds ratio value is expected in PNG areas such as Sanananda.

4.9 Applicability to Alternative Populations

As part of the case study utilised in this thesis, it is assumed that there are only two outcome populations of interest, Australian or Japanese. However, it is acknowledged that individuals from additional populations were present during the time of conflict, including Americans, Chinese, British and Papuans. To determine how individuals from populations outside the primary interest (Australian versus Japanese) are classified using the nominated panel, freely available samples from alternative populations will be analysed using the *p*LMT classifier. Individuals from the following populations were obtained from a combination of the 1000 Genomes Project and HGDP database:

- British (GBR; $n = 91$);
- Chinese (CHB; $n = 103$);
- Papuan (OCE; $n = 26$);
- American from European descent (CEU; $n = 99$).

4.10 Estimating the Margin of Error and a Measure of Confidence

The two outputs that can be used to infer ancestry when utilising the *p*LMT method are the GMAMP (Equation 4.2), if interest does not lie in the prior odds, and the posterior probability (Equation 4.6),

if the prior odds is of importance. Each of these two outputs has its own respective method for the calculation of a margin of error in the form of a confidence interval. The estimation of a confidence interval for each output is described in the following section.

Calculation of a Confidence Interval for the Geometric Mean

The geometric mean (Equation 4.2, repeated below), together with its variance, is required for a series of *M random variables*, each an output from a different model: p_1, p_2, …, p_m, Each of these has its own variance as measured by the Mean Square Error from the relevant model.

$$Geometric\ Mean = \left(\prod_{m=1}^{M} p_m\right)^{\frac{1}{M}}$$

To facilitate the calculation of the combined variance, the logarithm of the GM can be considered, requiring the addition of the logarithms of the k estimates (Equation 4.10). Note that natural logarithms (to the base e) are used.

$$\mathrm{Ln}(GM) = \frac{\ln(p_1) + \ln(p_2) + \cdots + \ln(p_m)}{M} \qquad (4.10)$$

The following two basic properties of variations for functions of variables can then be used (Adams and Clarkson, 1934):

1. The variance of a sum of independent variables is the sum of their variances.
2. The variance of a variable, p, divided by a constant, say a, is $V\left(\frac{p}{a}\right) = \frac{V(p)}{a^2}$

Thus, the variance of the logarithm of the GM is:

$$\mathrm{V}[\mathrm{Ln}(GM)] = \frac{V[\ln(p_1)] + V[\ln(p_2)] + \cdots + V[\ln(p_m)]}{M^2} \qquad (4.11)$$

Before this variance can be computed, the variance of the logarithms of the original random variables must be found. To do so, the Delta method for estimating the variance of a function is proposed. The Delta method is a statistical technique, which uses a Taylor expansion to approximate the expected values for functions of random variables when it is not possible to directly evaluate the expectation itself (Oehlert, 1992). The Delta method can be described in simple terms as follows. Suppose there is a function of a proportion ($f(p)$) for which a variance is required. Providing $f(p)$ is differentiable (that is, has an estimable derivative with respect to (p)), then its variance (V) can be approximated as shown in Equation 4.12.

$$V\big(f(\hat{p})\big) = \left(\frac{\partial\big(f(\hat{p})\big)}{\partial\hat{p}}\right)^2 \times V(\hat{p}) \tag{4.12}$$

Using the Delta method (Chakraborty et al., 1993; Curran et al., 2002), the variance of the logarithm of a random variable can be calculated as follows:

$$V(\ln X) \approx \frac{V(X)}{[E(X)]^2} \tag{4.13}$$

where E(X) is the expected value of the variable.

Note that this approximation is reported to be satisfactory provided that (Curran et al., 2002):

$$\frac{E(X)}{\sigma} > 2.5 \tag{4.14}$$

Where E(X) is the expected value and σ is the standard deviation of the random variable.

Once the variance of ln(GM) is calculated it can be used to obtain confidence intervals for the logarithm of the geometric mean using the standard *z*-score method. These confidence intervals can then be back transformed (using exponentiation) to give the confidence interval for the estimated GM.

In the current problem, the random variables are independent estimates of the probability of BGA, outputted by the *p*LMT for each of the models.

Posterior Probability Model. As the posterior probability is obtained through a function that contains a likelihood ratio of two probabilities, each of which has its own variance, the Delta method is utilised again. Note that it is assumed that the prior odds ratio is a constant and no variance is required to be estimated.

To estimate the variance of the likelihood ratio function ($\hat{x}/\hat{y}$), the Delta method is used. The final variance of the function ($\hat{x}/\hat{y}$), is given as Equation 4.15, while the full derivation is provided in Appendix 9, where n_1 and n_2 are the sample sizes used to estimate $\hat{x}$ and $\hat{y}$ respectively:

$$V\left(\frac{\hat{x}}{\hat{y}}\right) = \frac{\hat{x}}{\hat{y}^2} \times \left[\left(\frac{1-\hat{x}}{n_1}\right) + \left(\frac{\hat{x}(1-\hat{y})}{\hat{y}n_2}\right)\right] \tag{4.15}$$

Note that it is standard to also consider the covariance of the function when utilising the Delta method. However, as the likelihood ratio utilised in BGA prediction is estimated from samples taken from two independent populations, Australian and Japanese, there is no reason to expect a non-zero covariance, except by random chance, and thus is assumed to be zero and is omitted. As the variance

within the posterior probability calculation is assumed to be limited to solely the LR, the approach taken in this thesis to estimate a confidence interval of the posterior probability is to calculate a lower and upper limit for the likelihood ratio. These limits are then inputted into the calculation of the posterior probability, estimating a lower and upper confidence limit for the posterior respectively. As per standard statistical theory for the estimation of two-tailed $100(1-\alpha)\%$ confidence interval (CI), using z scores and therefore assuming normality, the CI for the likelihood ratio of probability of Australian and Japanese BGA, Equation 4.16 is used, where *z* is the corresponding *z* score.

$$\pm CI = \frac{\hat{x}}{\hat{y}} \pm z\sqrt{V\left(\frac{\hat{x}}{\hat{y}}\right)} \tag{4.16}$$

Equation 4.16 can then be incorporated into Equation 4.17 (a reiteration of Equation 4.6, now with the likelihood ratio adjusted to utilise the lower and upper limits obtained in Equation 4.16) to estimate a confidence interval for the posterior probability.

$$Two-tailed\ CI\ for\ Posterior\ Probability = \frac{1}{1+\frac{1}{Prior\ Odds}\times\frac{1}{\pm CI}} \tag{4.17}$$

Chapter 5 – Results and Discussion

5.1 Introduction

This chapter will outline the results of the experiments discussed in Chapter 4, with accompanying discussions where required. The structure will coincide with the order provided in Chapter 4, with the *p*LMT classifier, the Generic Bayesian classifier, and some additional sensitivity tests.

5.2 Parsimonious Logistic Model Tree (pLMT)

5.2.1 Logistic Model Tree Generation

For the training dataset of WWII era Australians (n = 108) and contemporary Japanese individuals (n = 104), a *p*LMT was generated which resulted in M = 5 independent LMTs which met the requirement of an averaged 10-fold cross-validation accuracy level across ten runs of 99%. Note that the same model is obtained for every repetition of the cross-validation, it is simply the accuracy which varies and is subsequently averaged. The SNPs for those models are shown in Table 5.1. Of the 40 utilised SNPs from the GPSP, 34 of them were used in the *p*LMT. Note that the SNPs shown in Table 5.1b demonstrate only the expected model for these 40 SNPs, at this stage, no unknown samples have been analysed. A detailed list of the SNPs and their coefficients as estimated in each model is provided in Appendix 4.

Table 5.1: SNPs Included in the *p*LMT Models
A summary list of the 40 SNPs used in this thesis, and whether the SNP was incorporated into the pLMT models. a) Lists which SNPs were utilised together for a given pLMT, b) summarises the full panel outlining which SNPs were and were not used - note the SNPs are in rank order for original discrimination power.

a)

LMT 1	LMT 2	LMT 3	LMT 4	LMT 5	Excluded
rs1426654	**rs12913832**	**rs28777**	**rs9809818**	**rs1876482**	**rs6754311**
	rs2196051	**rs820371**	**rs4683510**	**rs10455681**	**rs4787040**
	rs3811801	**rs4749305**	**rs7997709**	**rs9319336**	**rs2357442**
		rs6494411	**rs1448485**	**rs192655**	**rs1393350**
		rs9286879	**rs730570**	**rs11725412**	**rs12203592**
		rs10496971	**rs722869**	**rs4918664**	**rs4959270**
		rs683	**rs1250233**	**rs4781011**	
			rs1366220	**rs1471939**	
			rs2758988	**rs1950993**	
			rs984654	**rs4984913**	
			rs4463276		
			rs4833103		
			rs3907047		

b)

SNP	Used?	SNP	Used?
rs1426654	YES	rs10496971	YES
rs9809818	YES	rs2758988	YES
rs28777	YES	rs9319336	YES
rs12913832	YES	rs192655	YES
rs4683510	YES	rs984654	YES
rs820371	YES	rs11725412	YES
rs4749305	YES	rs4918664	YES
rs6494411	YES	rs4463276	YES
rs7997709	YES	rs683	YES
rs1448485	YES	rs4787040	NO
rs730570	YES	rs4781011	YES
rs1876482	YES	rs1471939	YES
rs722869	YES	rs1950993	YES
rs1250233	YES	rs4984913	YES
rs9286879	YES	rs4833103	YES
rs2196051	YES	rs3907047	YES
rs6754311	NO	rs2357442	NO
rs10455681	YES	rs1393350	NO
rs1366220	YES	rs12203592	NO
rs3811801	YES	rs4959270	NO

Of the six SNPs excluded from the *p*LMT, four had the lowest discrimination power, where discrimination power is equated to the absolute difference in allele frequencies between the Australian (n = 108) and Japanese (n = 104) samples. The exclusion of these four SNPs demonstrates the utility of a parsimonious model, where these low discriminatory SNPs are not used as they are likely to only introduce noise and reduce the efficiency of the subsequent classification. For SNP "rs6753411", its exclusion is likely linked to the SNP having insufficient discrimination power when used in tandem with other SNPs in the panel.

5.2.2 Validation of SimAdmixtR

A comparison was then made between the observed (calculated from the simulated data) and expected (determined algebraically using Mendelian inheritance formulae) genotype frequencies to determine SimAdmixtR's accuracy, results shown in Appendix 5. Using the 10,000 individuals created for scenarios 1 – 10, the estimated genotype frequencies for each SNP were calculated for each scenario. These estimated frequencies were then compared to the expected genotype frequencies as determined algebraically using the Australian and Japanese databases and Mendelian inheritance formulae and the absolute difference between the estimated and expected frequencies were recorded for each

scenario. The minimum, mean, maximum and standard deviation for the absolute difference across all ten scenarios are then calculated for the three genotypes at each SNP.

It should be noted that some departure from the theoretical expected frequencies are to be expected due to the randomness of the above process, however, the observed genotype frequencies did not deviate significantly (≥0.1) from their expected values.

5.2.3 Analysis of Simulated Data and Establishment of Thresholds

The individuals from the seven scenarios (1, 2, 4, 6, 8, 9 and 10) in Simulation Group 1 were analysed using the *p*LMT classifier. The resulting distributions of GMAMPs are shown in the following boxplots (Figure 5.2). The key indicating each boxplot is based on the admixture scenario as originally described in Table 4.3, where the key describes the ratio of Australian/Japanese ancestor proportions. For example, 75/25 describes individuals simulated based on a pedigree with 75% Australian great-grandparents and 25% Japanese great-grandparents.

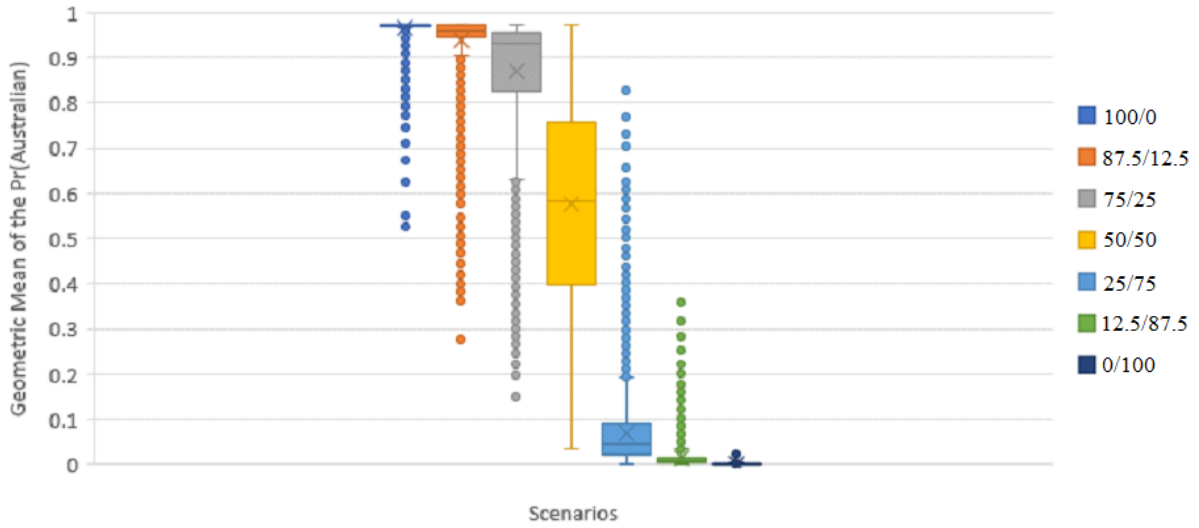


Figure 5.2: Distribution of Simulation Group 1's GMAMP
Distribution of Simulation Group 1's GMAMP *for seven scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportions, from the top down, the scenarios used are 1, 2, 4, 6, 8, 9 and 10.*

Based on the distributions of the Australian scenarios and the Japanese scenarios, the following classification thresholds were selected:

$$\begin{aligned} \text{Australian:} &\quad \text{GMAMP} \geq 0.8 \\ \text{Ambiguous:} &\quad 0.1 < \text{GMAMP} < 0.8 \\ \text{Japanese:} &\quad \text{GMAMP} \leq 0.1 \end{aligned}$$

These thresholds were selected as approximately ≥90% of the Australian distributions (Scenarios 1, 2 and 4, Table 4.3: Correct Interpretation = Australian) were observed to be greater than 0.8, while ≥90% of the Japanese distributions (Scenarios 8, 9 and 10, Table 4.3: Correct Interpretation = Japanese) were below 0.1. Due to the overlap of Australian and Japanese distributions between 0.1 and 0.8, this region was classified as Ambiguous. Individuals with probabilities within the ambiguous region cannot confidently be assigned to one of the populations and should remain unclassified to avoid error. Note that a single instance was observed in the 75% Japanese scenario (25/75, Scenario 8), of a Japanese individual being misclassified as Australian based on these thresholds. These thresholds could be changed based on the desired accuracy, but any such change will also affect the misclassification rate. For example, a more conservative threshold, such as 0.9, will result in less misclassifications, but more samples being classified as ambiguous. While samples that fall within the ambiguous range may be indicative of family-level admixture, it is important to note that there is the alternative possibility the obtained sample does not originate from either of the populations of interest. It is important to acknowledge then when outlining interpretation guidelines based on thresholds such as these, there is always the possibility of false positives and negatives. In the context of the UWC-A scenario, a false negative would be an individual who truly originates from Australia, yet they are misclassified as Japanese, and a false positive is the reversal of this scenario. To investigate the possibility of either of these false outcomes occurring, validation of thresholds is required (Section 5.2.4). However, said validation will only relate to the thresholds as they currently stand, if a threshold is adjusted in either direction, the probabilities of a false positive or a false negative are altered as well. The occurrence of false positives and negatives is an unavoidable aspect of any prediction-based modelling, and the KBDSS must ensure that the user is aware of their possible existence and of how probable each is, understands their possible impact, and is advised of measure that can be taken to mitigate their effects.

5.2.4 Validation of Thresholds

Simulation Group 2 (consisting of all ten scenarios) was then processed through the *M* models obtained from the original data using the *p*LMT classifier, and individuals were categorised using the thresholds that were established using Simulation Group 1. The distributions of the resulting GMAMP for the ten scenarios are shown in Figure 5.3.

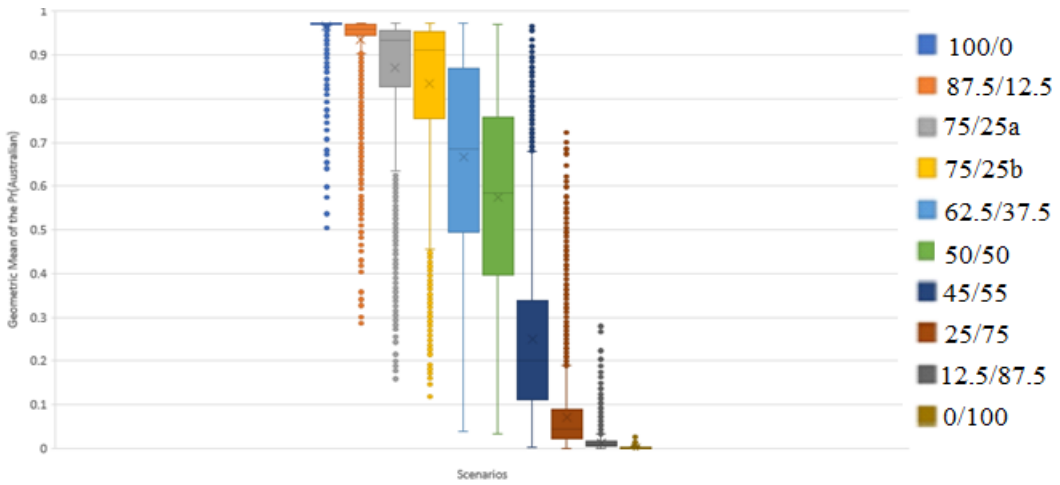


Figure 5.3: Distribution of Simulation Group 2's GMAMP
Distribution of simulation group 2's GMAMP *for ten scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportions, from the top down, the scenarios used are 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.*

Table 5.2 outlines the count of inferred ancestry for the 10,000 individuals within each of Simulation Group 2's admixture scenarios using the classification thresholds.

Table 5.2: Counts of Classification Outcomes for Simulation Group 2 for *p*LMT
Australian: GMAMP ≥ 0.8, Ambiguous: 0.1 < GMAMP < 0.8, and Japanese: GMAMP ≤ 0.1. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportion. Shown are the counts of individuals assigned to each inferred ancestry within a scenario (n = 10,000), alongside the respective percent.

Scenario Number	Represented Average Pedigree Proportions	Inferred Ancestry		
		Australian	Ambiguous	Japanese
1	100/0	9948 (99.48%)	52 (0.52%)	**0 (0%)**
2	87.5/12.5	9496 (94.96%)	504 (5.04%)	**0 (0%)**
3	75/25a	8064 (80.64%)	1936 (19.36%)	**0 (0%)**
4	75/25b	7086 (70.86%)	2914 (29.14%)	**0 (0%)**
5	62.5/37.5	3606 (36.06%)	6370 (63.70%)	24 (0.24%)
6	50/50	1988 (19.88%)	7967 (79.67%)	45 (0.45%)
7	45/55	116 (1.16%)	7744 (77.44%)	2140 (21.40%)
8	25/75	**0 (0%)**	2109 (21.09%)	7891 (78.91%)
9	12.5/87.5	**0 (0%)**	61 (0.61%)	9939 (99.39%)
10	0/100	**0 (0%)**	2 (0.02%)	9998 (99.98%)

Errors observed in this table were assigned using the previously defined *direct* and *indirect* errors terms. The appropriate error rate calculated for each of the ten scenarios, together with 95% Wilson confidence interval where relevant are given in Table 5.3.

Table 5.3: *p*LMT Error Rates

Error rates for the pLMT classifier on each of the ten admixture scenarios tested, a two-tailed 95% Wilson confidence interval was included where relevant (Sergeant, 2018). N/A = Not available, namely, that this error cannot occur in this given scenario. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportion.

Scenario Number	Represented Average Pedigree Proportions	Direct Error (95% CI)	Indirect Error (95% CI)
1	100/0	<0.0004	N/A
2	87.5/12.5	<0.0004	N/A
3	75/25a	<0.0004	N/A
4	75/25b	<0.0004	N/A
5	62.5/37.5	N/A	0.3630 (0.3536 – 0.3725)
6	50/50	N/A	0.2033 (0.1955 – 0.2113)
7	45/55	N/A	0.2256 (0.2175 – 0.2339)
8	25/75	<0.0004	N/A
9	12.5/87.5	<0.0004	N/A
10	0/100	<0.0004	N/A

The difference between a direct versus an indirect error is the resulting outcome. A direct error is akin to a false negative or positive (depending which population is used as the reference population), that is, an Australian soldier misclassified as Japanese would be a true misclassification, similarly a Japanese soldier misclassified as Australian. An indirect error would result in a soldier receiving an ambiguous classification when the correct response should be an Australian or Japanese classification. The impact of the indirect error is that the soldier's family receive no resolution and the soldier's identity remains in limbo indefinitely. Note that it is expected that, given the exclusions in place at the time of WWII, the majority of remains that will be discovered by the UWC-A will consist of Australian and Japanese individuals with little, if no, admixture between the two populations. However, it is important to explore these complex, admixed pedigrees to understand the capabilities and limitations of the proposed model.

For Simulation Group 2, there was not a single observed direct error, with the indirect error occurring in the unknown scenarios ranging between approximately 19.5% - 37.2% of the GM. Note that while no instances of direct error were observed in this test simulation group, as previously discussed a single instance of direct error did occur for Simulation Group 1 in the 75% Japanese scenario. Therefore, it is appropriate to acknowledge a direct error rate of <0.0004 (95% CI), to account for the possibility of another direct error in a different sample of simulated individuals. Based on the results from Table 5.3, this experiment has also indirectly identified that Australian and Japanese individuals can still be reliably classified for pedigrees consisting of as low as 75% of their majority ancestry. The region between these reliable classifications consists of pedigrees that are less likely to occur naturally, and in these instances BGA cannot confidently be assigned.

If the *p*LMT's direct error rate is considered as the primary metric for comparison, the classifier can be compared to other methods observed in the literature, such as Cheung et al. (2017). Based on the

results from Table 5.3, in comparison with Cheung et al. (2017, Table 1, p.905), the *p*LMT classifier has an error rate similar to that of STRUCTURE (<0.01), which had the highest accuracy of the several methods compared.

5.3 Generic Bayesian

5.3.1 Analyse Simulated Data

The likelihood ratio was calculated for all individuals from Simulation Group 1 using the original dataset of WWII era Australians (*n* = 108) and contemporary Japanese individuals (*n* = 104) to provide relative genotype frequencies. The natural logarithmic scale was applied to these likelihood ratios to assist with scaling and interpretation. The distributions of the resulting likelihood ratios for each scenario are shown in Figure 5.4, which is the equivalent of Figure 5.2 for the *p*LMT.

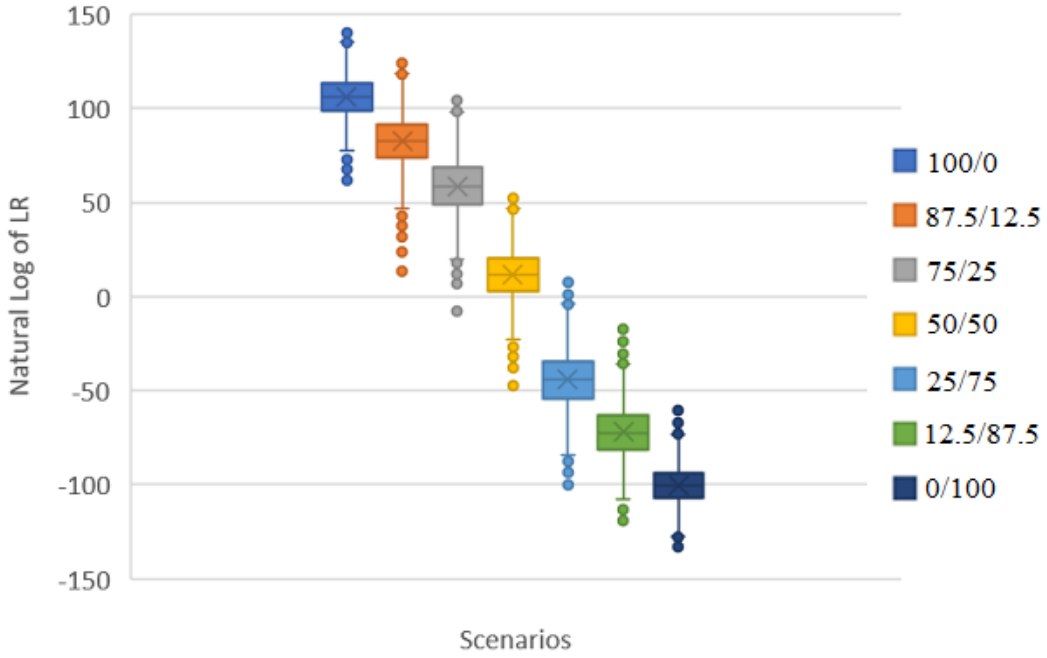


Figure 5.4: Distribution of Simulation Group 1's Natural Log of the Likelihood Ratio (LR)
Distribution of Simulation Group 1's natural log of the likelihood ratio for seven scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportion, from the top down, the scenarios used are 1, 2, 4, 6, 8, 9 and 10.

Based on the distributions of the Australian scenarios and the Japanese scenarios, the following thresholds were selected:

Australian:	Natural Log of Likelihood Ratio ≥ 50
Ambiguous:	-25 < Natural Log of Likelihood Ratio < 50
Japanese:	Natural Log of Likelihood Ratio ≤ -25

These thresholds were selected using the methodology as discussed in Section 5.2.3, with the criterion that approximately 90% of the population, that is, Australian = Scenarios 1, 2 and 4, and Japanese = Scenarios 8, 9 and 10, was covered by the threshold.

5.3.2 Test Thresholds

Simulation Group 2 was then analysed using the Generic Bayesian classifier, and individuals were categorised using the thresholds established on Simulation Group 1. The resulting distributions of the natural log of the likelihood ratio for the ten scenarios are shown in the following boxplots (Figure 5.5).

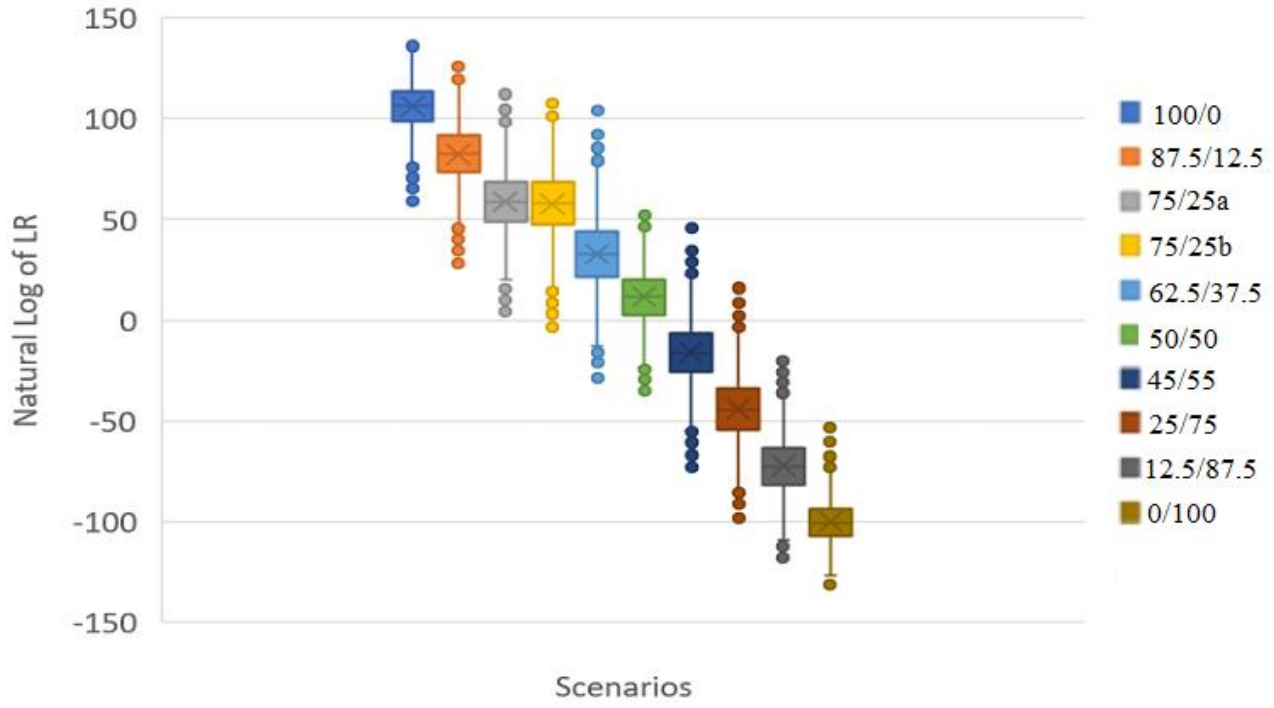


Figure 5.5: Distribution of Simulation Group 2's Natural Log of the LR
Distribution of Simulation Group 2's natural log of the likelihood ratio for ten scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportions, from the top down, the scenarios used are 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

Table 5.4 presents the count of inferred ancestry for the 10,000 individuals within each of Simulation Group 2's admixture scenarios using the classification thresholds.

Table 5.4: Counts of Classification Outcomes for Simulation Group 2 for Generic Bayesian
Australian = Natural Log of Likelihood Ratio ≥ 50, Ambiguous = 50 > Natural Log of Likelihood Ratio > -25, Japanese = Natural Log of Likelihood Ratio ≤ -25. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportions. Shown are the counts of individuals assigned to each inferred ancestry within a scenario (n = 10,000), alongside the respective percent.

Scenario Number	Represented Average Pedigree Proportions	Inferred Ancestry		
		Australian	Ambiguous	Japanese
1	100/0	10000 (100%)	0 (0%)	**0 (0%)**
2	87.5/12.5	9878 (98.78%)	122 (1.22%)	**0 (0%)**
3	75/25a	7315 (73.15%)	2685 (26.85%)	**0 (0%)**
4	75/25b	6942 (69.42%)	3058 (30.58%)	**0 (0%)**
5	62.5/37.5	1490 (14.90%)	8507 (85.07%)	3 (0.03%)
6	50/50	14 (0.14%)	9960 (99.60%)	26 (0.26%)
7	45/55	0 (0%)	7325 (73.25%)	2675 (26.75%)
8	25/75	**0 (0%)**	1012 (10.12%)	8988 (89.88%)
9	12.5/87.5	**0 (0%)**	6 (0.06%)	9994 (99.94%)
10	0/100	**0 (0%)**	1 (0.01%)	9999 (99.99%)

Errors observed in this table were assigned using the previously defined *direct* and *indirect* error terms, and the appropriate error rate calculated for each of the ten scenarios, with a 95% Wilson confidence interval being applied where relevant (Table 5.5).

Table 5.5: Generic Bayesian Error Rates
Error rates for the Generic Bayesian classifier on each of the ten admixture scenarios tested together with a 95% Wilson confidence interval where relevant (Sergeant, 2018). N/A = Not available, namely, that this error cannot occur in this given scenario. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportion.

Scenario Number	Represented Average Pedigree Proportions	Direct Error (95% CI)	Indirect Error (95% CI)
1	100/0	<0.0004	N/A
2	87.5/12.5	<0.0004	N/A
3	75/25a	<0.0004	N/A
4	75/25b	<0.0004	N/A
5	62.5/37.5	N/A	0.1493 (0.1424 – 0.1564)
6	50/50	N/A	0.0040 (0.0029 – 0.0054)
7	45/55	N/A	0.2675 (0.2589 – 0.2763)
8	25/75	<0.0004	N/A
9	12.5/87.5	<0.0004	N/A
10	0/100	<0.0004	N/A

For the Generic Bayesian classifier, there were no observed direct errors, on either Simulation Groups 1 or 2, and the resulting indirect error rate ranged from 0.3% - 27.6%, compared to the *p*LMT's indirect error rate range of 20% - 38%.

5.4 STRUCTURE

5.4.1 Analyse Simulated Data

The Australian membership proportion (*Q* value) was estimated by STRUCTURE for all individuals from Simulation Group 1 using the original dataset of WWII era Australians (*n* = 108) and modern-day Japanese individuals (*n* = 104) to provide relative genotype frequencies. The distributions of the resulting likelihood ratios for each scenario are shown in Figure 5.6 (equivalent to Figure 5.2 for *p*LMT).

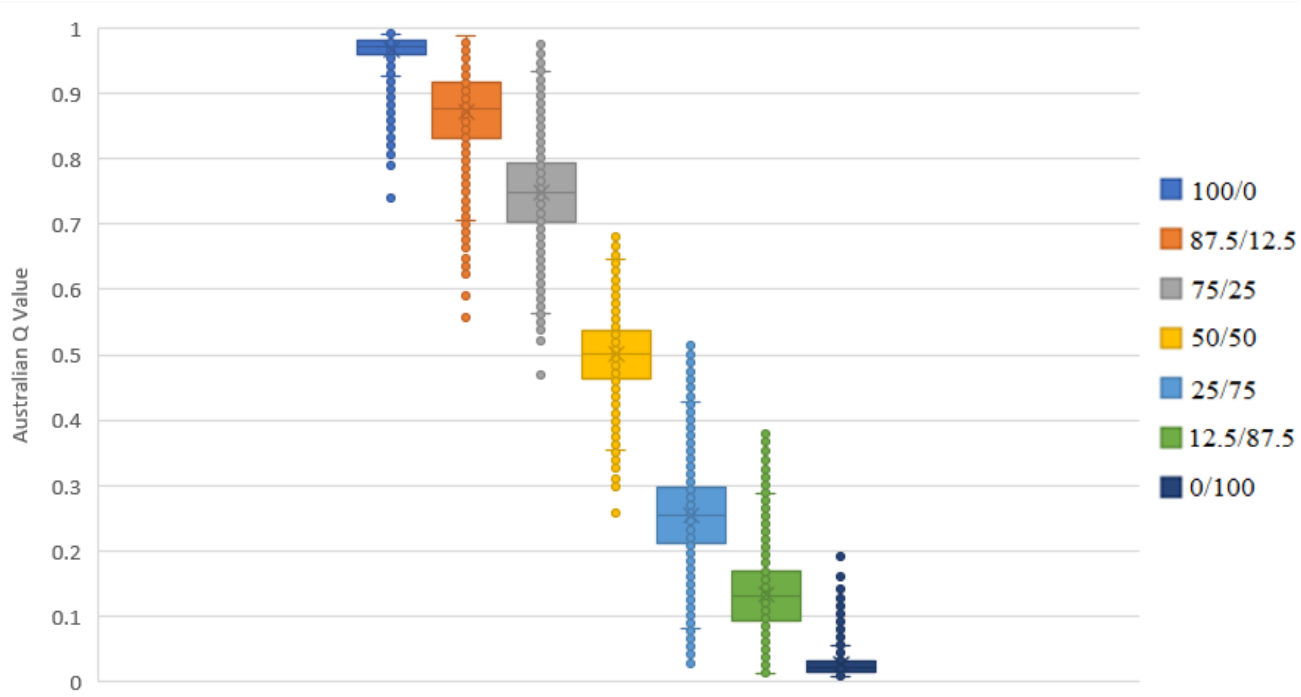


Figure 5.6: Distribution of Simulation Group 1's Australian Membership Proportion (*Q* Value)
Distribution of Simulation Group 1's Australian Q value for seven scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportions, from the top down, the scenarios used are 1, 2, 4, 6, 8, 9 and 10.

Thresholds were selected using the methodology as discussed in Section 5.2.3, with the criterion that approximately 90% of the population, that is, Australian = Scenarios 1, 2 and 4, and Japanese = Scenarios 8, 9 and 10, was covered by the threshold. Based on the distributions of the Australian scenarios and the Japanese scenarios, the following thresholds were selected:

Australian:	Australian *Q* Value ≥ 0.7
Ambiguous:	$0.3 <$ Australian *Q* Value < 0.7
Japanese:	Australian *Q* Value ≤ 0.3

5.4.2 Test Thresholds

Simulation Group 2 was then analysed using STRUCTURE, and individuals were categorised using the thresholds established with Simulation Group 1. The resulting distributions of the Australian *Q* Value for the ten scenarios are shown in the following boxplots (Figure 5.7).

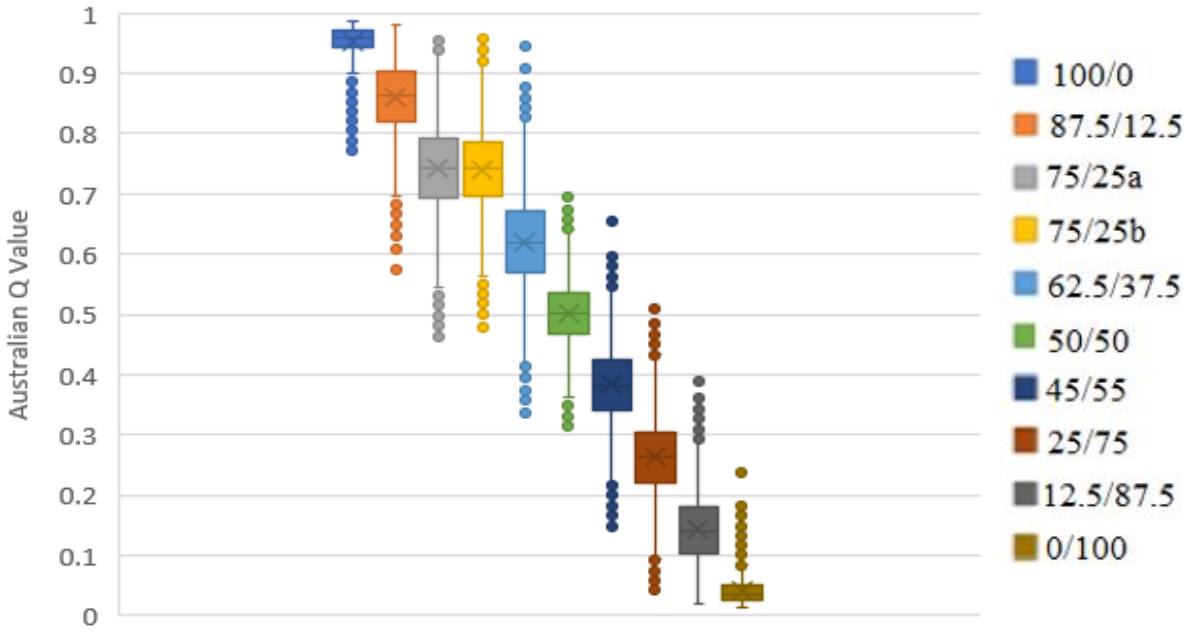


Figure 5.7: Distribution of Simulation Group 2's Australian Membership Proportion (*Q* Value)
Distribution of Simulation Group 2's Australian Q value for ten scenarios (with full descriptions provided in Table 4.3). The key represents the ratio of Australian/Japanese ancestor proportions, from the top down, the scenarios used are 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

Table 5.4 outlines the count of inferred ancestry for the 10,000 individuals within each of Simulation Group 2's admixture scenarios using the classification thresholds.

Table 5.6: Counts of Classification Outcomes for Simulation Group 2 for STRUCTURE
Australian = Australian Q Value ≥ 0.7, Ambiguous = 0.3 < Australian Q Value < 0.7, Japanese = Australian Q Value ≤ -0.3. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportions. Shown are the counts of individuals assigned to each inferred ancestry within a scenario (n = 10,000), alongside the respective percent.

Scenario Number	Represented Average Pedigree Proportions	Inferred Ancestry		
		Australian	Ambiguous	Japanese
1	100/0	10000 (100%)	0 (0%)	**0 (0%)**
2	87.5/12.5	9936 (99.36%)	64 (0.64%)	**0 (0%)**
3	75/25a	7222 (72.22%)	2778 (27.78%)	**0 (0%)**
4	75/25b	7384 (73.84%)	2616 (26.16%)	**0 (0%)**
5	62.5/37.5	1482 (14.82%)	8518 (85.18%)	0 (0%)
6	50/50	0 (0%)	10000 (100%)	0 (0%)
7	45/55	0 (0%)	9149 (91.49%)	851 (8.51%)
8	25/75	**0 (0%)**	2727 (27.27%)	7273 (72.73%)
9	12.5/87.5	**0 (0%)**	43 (0.43%)	9957 (99.57%)
10	0/100	**0 (0%)**	0 (0%)	10000 (100%)

Errors observed in this table were assigned using the previously defined *direct* and *indirect* errors terms, and the appropriate error rate calculated for each of the ten scenarios, with a 95% Wilson confidence interval being applied where relevant (Table 5.7).

Table 5.7: STRUCTURE Error Rates

Error rates for the STRUCTURE classifier on each of the ten admixture scenarios tested where a 95% Wilson confidence interval was calculated (Sergeant, 2018). N/A = Not available, namely, that this error cannot occur in this given scenario. The key (represented average pedigree proportions) represents the ratio of Australian/Japanese ancestor proportion.

Scenario Number	Represented Average Pedigree Proportions	Direct Error (95% CI)	Indirect Error (95% CI)
1	100/0	<0.0004	N/A
2	87.5/12.5	<0.0004	N/A
3	75/25a	<0.0004	N/A
4	75/25b	<0.0004	N/A
5	62.5/37.5	N/A	0.1482 (0.1413 – 0.1553)
6	50/50	N/A	<0.0004
7	45/55	N/A	0.0851 (0.0798 – 0.0907)
8	25/75	<0.0004	N/A
9	12.5/87.5	<0.0004	N/A
10	0/100	<0.0004	N/A

For STRUCTURE, there were no observed direct errors, on either Simulation Group 1 or 2, and the resulting indirect error rate was lower than the *p*LMT, ranging from 8.51% - 14.82%.

Comparing the error rates across the three classifiers demonstrates that for a complete panel profile (Table 5.8), all three methods are effective based on no observed direct errors in the used sample, however, STRUCTURE had highest performance based on having the lowest maximum indirect error rate.

Table 5.8: Summarised Classifier Error Rates

The direct and indirect error rate range for the three utilised classifiers: pLMT, Generic Bayesian and STRUCTURE. Values were obtained from the recorded minimum and maximum for the two error rates across all ten scenarios from Tables 5.3, 5.5 and 5.7 respectively. Note that for the Direct Error, there were no observations in this sample, therefore, a minimum conservative estimate is used in its place.

Classifiers	Direct Error	Indirect Error (95% CI)
Parsimonious Logistic Model Tree	< 0.0004	19.5% - 37.2%
Generic Bayesian	< 0.0004	0.3% - 27.6%
STRUCTURE	< 0.0004	8.5% - 14.8%

5.5 Classifier Comparison on Degraded Samples

To observe how the three classifiers compared when inferring BGA for degraded remains, a second WWII Australian sample (n = 80) consisting of individuals with missing data, mimicking degradation, was utilised (collected in collaboration with Ghaiyed (2020)). Based on the results of the SNP removal experiment (see Section 5.6), a minimum threshold of ten SNPs was selected. This selection resulted in five samples being excluded for having less than 10 SNPs available, providing a final sample of 75 degraded Australians.

For the *p*LMT, the data for all individuals were analysed separately as the intended use of the *p*LMT's procedure, having their own respective *p*LMT generated based on which SNPs were available. The GMAMP was then estimated for each of the 75 individuals. Appendix 6 lists the estimated GMAMP and available SNPs for each of the 75 individuals. For the Generic Bayesian, probabilities for missing SNPs were replaced with a value of one, resulting in these SNPs having no effect on the resulting likelihood ratio, thus effectively being removed. The natural log of the likelihood ratio was estimated for each of the 75 individuals. Full details are provided in Appendix 7. STRUCTURE's standard procedure to ignore missing data was utilised for these runs, and full details for the Australian *Q* values obtained from STRUCTURE are provided in Appendix 8.

Since interest lies in the effect on the ancestry estimation caused by the number of SNPs available, the following figures depict each classifier's estimated output of interest for each of the n = 75 WWII degraded Australian samples, against the number of SNPs missing for that individual (ranging from 5 to 28 out the total possible 40). Note that the SNPs missing (or present) will not necessarily be the same between individuals. A simple linear regression is applied to each plot to observe any effect the number of SNPs may have on the classifier's estimated output.

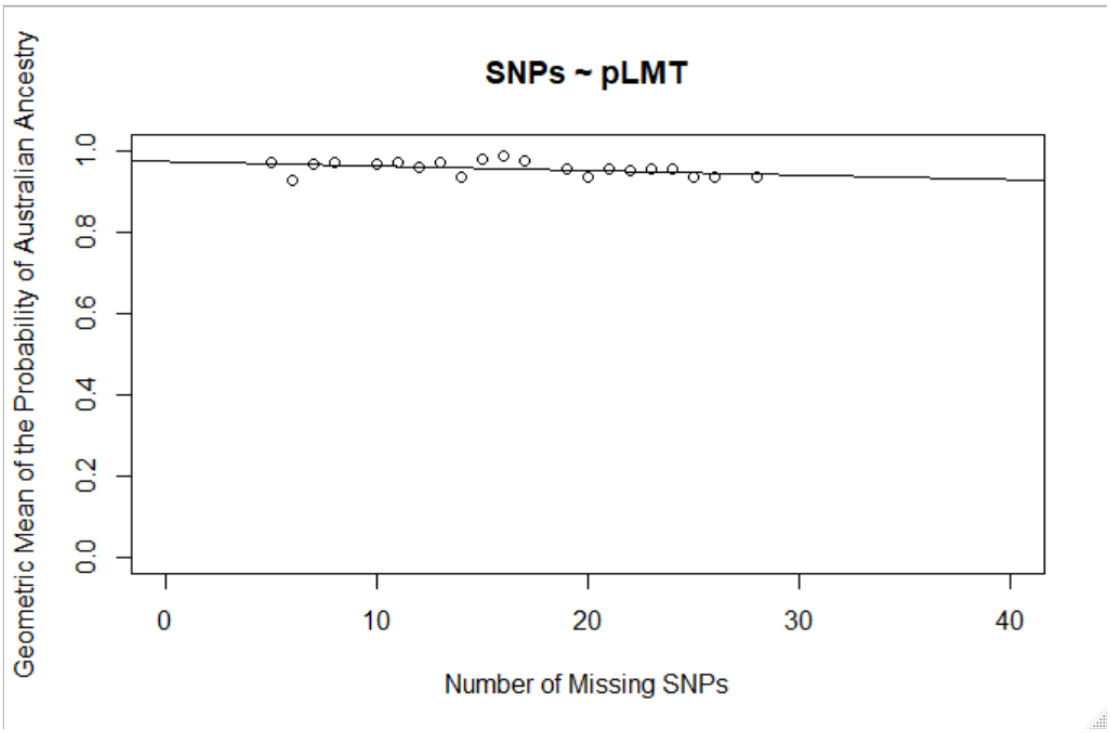
In addition, each of the n = 75 WWII degraded Australian samples are categorised using the thresholds established in Sections 5.2.3, 5.3.1 and 5.4.1. Table 5.9 details the averaged output of each classifier (GMAMP, natural log of the likelihood ratio of Australian ancestry and the Q value of Australian membership proportion) for the number of missing SNPs (out of the total possible 40).

Table 5.9: Averaged Classifier Outputs for Degraded Samples

The averaged outputs for the pLMT (GMAMP), Generic Bayesian (natural log likelihood ratio of Australian ancestry) and STRUCTURE (Q value of Australian membership proportion) at various levels of missing SNPs ranging from 5 to 28 (out of 40). Number of observations is also included for each level.

No. Missing SNPs	No. Observations	*p*LMT (GMAMP)	Generic Bayesian (Log Likelihood Ratio)	STRUCTURE (Q Value)
5	1	0.971	98.52	0.999
6	3	0.931	87.07	0.990
7	3	0.970	84.53	0.997
8	2	0.972	92.09	0.999
10	1	0.968	87.78	0.999
11	5	0.972	74.77	0.995
12	3	0.959	66.67	0.987
13	1	0.972	67.28	0.991
14	4	0.935	70.03	0.997
15	2	0.982	79.81	0.999
16	3	0.988	69.88	0.997
17	8	0.976	65.38	0.997
19	5	0.955	52.38	0.996
20	7	0.938	44.69	0.979
21	2	0.956	40.92	0.995
22	5	0.953	45.44	0.997
23	5	0.955	38.05	0.993
24	8	0.956	36.22	0.991
25	2	0.936	39.69	0.998
26	3	0.937	41.02	0.997
28	1	0.936	27.52	0.997

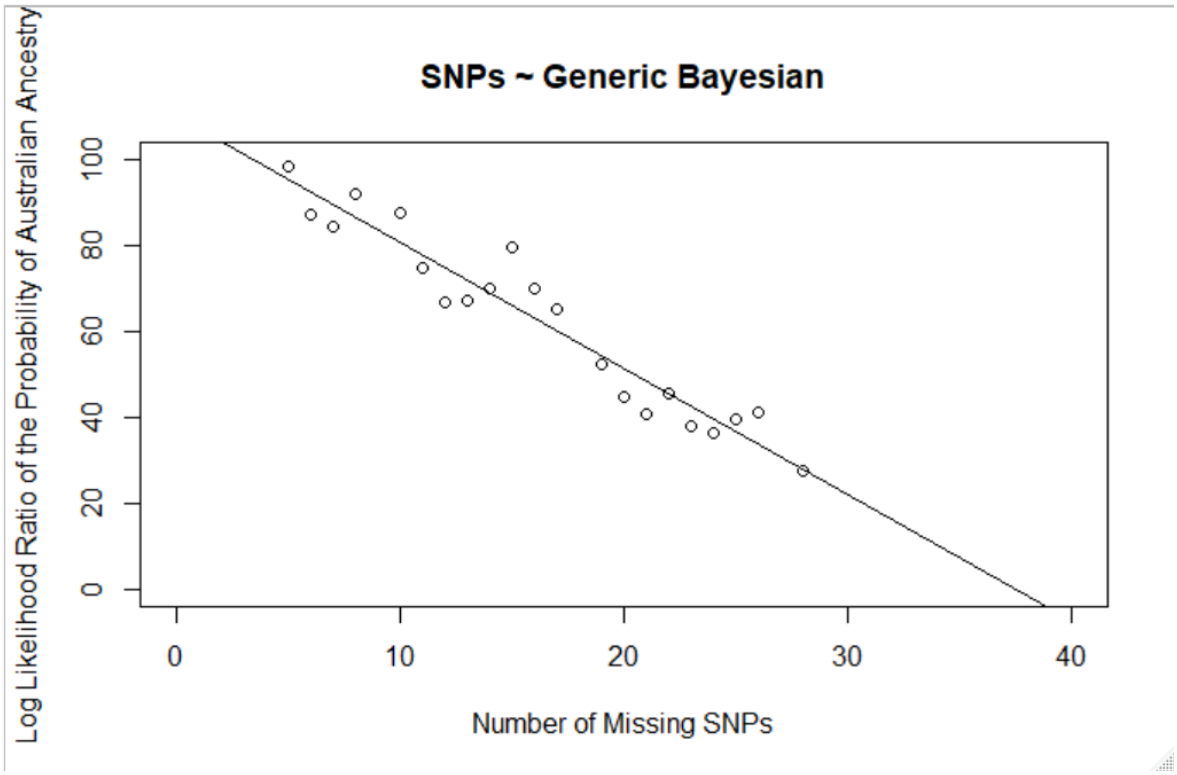
Note that the outputs of Figures 5.8, 5.9 and 5.10 depict the different measures used by the *p*LMT (GMAMP), Generic Bayesian (log likelihood ratio) and STRUCTURE (Q value) classifiers.



	Coefficients	Standard Error	P-value
Intercept	0.9755714	0.0089376	<2e-16
Missing SNPs	-0.0010877	0.0005061	0.0447

R-Squared = 0.1956

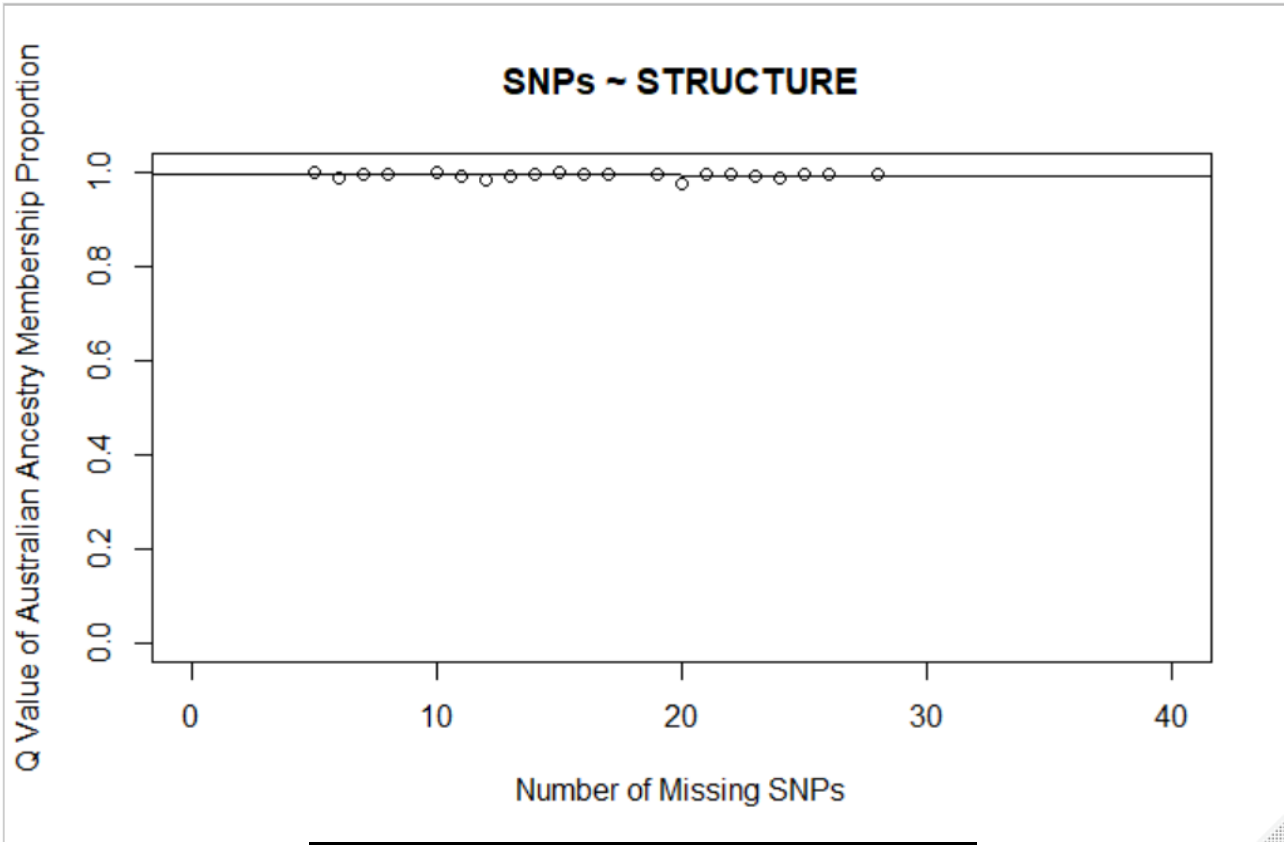
Figure 5.8: GMAMP Distribution in the Degraded WWII Australian Sample
Averaged GMAMP for individuals in the degraded WWII era Australian sample (n = 75), based on the number of missing SNPs out of the total possible 40 available. Accompanied by the regression coefficients, standard error, significance value and R-squared value.



	Coefficients	Standard Error	P-value
Intercept	110.2607	3.4315	< 2e-16
Number of SNPs	-2.9408	0.1943	4.7e-12

R-Squared = 0.9234

Figure 5.9: Log Likelihood Ratio Distribution in the Degraded WWII Australia Sample
Averaged log likelihood ratios for individuals in the degraded WWII era Australian sample (n = 75), based on the number of missing SNPs from the possible 40 available. Accompanied by the regression coefficients, standard error, significance value and R-squared value.



	Coefficients	Standard Error	P-value
Intercept	0.9954675	0.0029053	<2e-16
Number of SNPs	-0.0000474	0.0001645	0.776

R-Squared = 0.0044

Figure 5.10: Australian *Q* Value Distribution in the Degraded WWII Australia Sample
Averaged Australian Q values for individuals in the degraded WWII era Australian sample (n = 75), based on the number of missing SNPs from the possible 40 available. Accompanied by the regression coefficients, standard error, significance value and R-squared value.

By plotting each classifier's averaged output against the number of missing SNPs, a significant reduction was observed for the *p*LMT classifier (*p*-value = 0.0447) and the Generic Bayesian classifier (*p*-value ≤ 0.01). For these classifiers, the loss of each SNP resulted in a loss of 0.001 for the GMAMP and 2.941 for the natural log likelihood of Australian ancestry. Note that converting the natural log value of 2.941 back to a normal likelihood ratio results in a reduction of approximately 19 for each missing SNP. The R-squared values for these two classifiers indicate that approximately 20% and 92% of the variation for the *p*LMT and Generic Bayesian classifiers, respectively, is accounted for by the number of missing SNPs. The Generic Bayesian's high R-squared value indicates that the classifier's performance is limited as the degree of degradation increases.

For STRUCTURE, it was observed that the number of missing SNPs did not have a significant effect on the resulting Q-values of Australian membership proportion (*p*-value = 0.776). In addition, the R-squared indicates that virtually no variation is occurring from the number of missing SNPs. This occurrence is likely due in part to STRUCTURE's algorithm and how it handles missing data.

Following regression, the previously established classification thresholds for the three classifiers in Sections 5.2.3, 5.3.1 and 5.4.1, were applied to the degraded WWII Australian individuals (See Appendices 6, 7 and 8). Both the *p*LMT and STRUCTURE classifiers resulted in 100% of individuals (n = 75) being correctly classified as Australian. The Generic Bayesian classifier however, resulted in the following outcomes: 40 (53%) were classified as Australian (Log LR ≥ 50), 35 (47%) were classified as Ambiguous (-25 < Log LR < 50), and 0 (0%) were classified as Japanese (Log LR ≤ -25). As it is expected for most historical cases will be degraded, the utilised classifier must be readily adaptable to DNA profiles with missing data. While the Generic Bayesian did not incorrectly classify any of the individuals, almost half of the sample (35/75) were classified as ambiguous.

5.6 SNP Removal Experiment

For the two scenarios, 100/0 (Scenario 1) Australian (n = 10,000) and 0/100 (Scenario 10) Japanese (n = 10,000), SNPs were randomly removed from the individuals in Simulation Group 2. SNP removal was carried out in sets of 5, 10, 15, 20, 25, 30, and 35 and 100 independent repetitions were completed for each set. Ambiguous classifications were removed from further analysis as there was no further interest in monitoring indirect error situations. The remaining individuals were classified into either correct or a direct error. The distributions of these classifications across the 100 repetitions are summarised as minimum, mean and maximum counts in Table 5.10. For example, when 5 SNPs were removed from the 100% Australian group, across the 100 replications, the mean number of the 10,000 individuals who were correctly classified was 9915, with the minimum being 9728 and the maximum 9991. For this example, no direct errors were observed for this group but there were some ambiguous classifications (not shown in Table 5.10). Similarly, when 5 SNPs were removed from the 100% Japanese group. The minimum, mean and maximum individuals correctly classified were 9966, 10000 and 10000, respectively. Again, there were no direct errors, the remaining individuals resulted in an ambiguous classification (not shown). Note that for this section, only the *p*LMT classifier was used.

Table 5.10: Artificial SNP Removal Summary

SNPs were artificially removed from a 100% Australian and 100% Japanese scenario (n = 10,000 each) in sets of five SNPs with 100 iterations per set. The minimum (min), mean and maximum (max) counts (based on 100 repetitions) for the subsequent correct and direct error categories are shown rounded up to the nearest integer, N/P = Not Possible. Each entry in the table is a count out of 10,000

		Correct			Direct Error		
SNP Removed	Scenarios	Min	Mean	Max	Min	Mean	Max
5 SNPs	Australian	9728	9915	9991	0	0	0
	Japanese	9966	10000	10000	0	0	0
10 SNPs	Australian	9713	9906	9989	0	0	0
	Japanese	9929	9999	10000	0	0	0
15 SNPs	Australian	9607	9898	9992	0	1	9
	Japanese	9930	9996	10000	0	0	0
20 SNPs	Australian	9540	9893	9998	0	1	1
	Japanese	9745	9981	10000	0	0	0
25 SNPs	Australian	9588	9869	9994	0	1	4
	Japanese	0[a]	9812	9994	5	8	114
30 SNPs	Australian	8836	9791	9994	0	2	23
	Japanese	N/P[b]	9464[b]	9989	5	13	114
35 SNPs	Australian	N/P	N/P	N/P	N/P	N/P	N/P
	Japanese	N/P	N/P	N/P	N/P	N/P	N/P

[a] One repetition resulted in a *p*LMT where no Japanese individuals were correctly assigned, that is, the individuals were either classified as Ambiguous (not shown) or a direct error (114/10,000)

[b] Three repetitions resulted in *p*LMTs that were unable to generate a single LMT, therefore, the average will be marginally skewed due to the presence of 'false' zeros.

Table 5.10 can be interpreted as follows, using the example of 5 SNPs removed. Initially, five SNPs were randomly removed, and these same five SNPs were removed from the 100% Australian (n = 10,000) and the 100% Japanese (n = 10,000) samples. A *p*LMT was generated based on the remaining available SNPs and these 20,000 individuals were classified based on the *p*LMT and previously established GMAMP thresholds. The count was then recorded for the number of individuals who were correctly assigned to their respective population and the number of direct errors. This entire process was repeated 100 times, with a new set of five SNPs selected at random for each repetition. The minimum, average (mean) and maximum were then estimated for the count of correct and direct error classifications over the 100 iterations.

It was observed that when 35 SNPs were removed (that is 5 SNPs were left for analysis), in nearly all repetitions there were too few SNPs available for the *p*LMT classifier to successfully construct a single LMT to predict ancestry. Therefore, it is indicative that the *p*LMT requires at least ten SNPs out of the 40 utilised SNPs to create models. Note that this statement is directly related to this specific panel (GPSP) of 40 SNPs and may not be applicable to other panels. As the generation of models depends on the discrimination power of the available SNPs, a different panel which contains SNPs

of a reduced power may require more than a minimum of ten. While recommendation for this panel is a minimum of ten SNPs available, the current analysis suggests that it is possible for the *p*LMT to be ineffective with only fifteen SNPs available (25 removed from the 40) (Table 5.10[a]).

For severely degraded remains, 30 SNPs missing, the *p*LMT was still able to create models with a high classification rate. On average, with only ten SNPs available (30 removed) the *p*LMT was able to correctly assign 97.9% of Australian individuals (9791/10,000), and 94.6% of Japanese individuals (9464/10,000). Also, with only ten SNPs (30 SNPs missing), the maximum numbers of direct errors in the 100% Australian and 100% Japanese sample were 23 and 114, respectively.

The classification thresholds used during this experiment remained static, based on the same thresholds established on individuals with complete panel profiles. However, there may be merit in establishing thresholds based on the number of SNPs that are available for a degraded set of remains, which may improve the *p*LMT's classification ability. Despite using static thresholds, the *p*LMT was still able to classify individuals from these two samples with a high success rate. In comparison to other classifiers in the literature, the *p*LMT's classification accuracy for degraded remains is relatively high; comparing the classifier to STRUCTURE and the Generic Bayesian results from Cheung et al. (2017) indicate the three classifiers have similar classification accuracies for highly degraded remains. However, the results observed in this thesis indicate that the Generic Bayesian classifier has limited prediction power as the number of SNPs available is reduced.

5.7 Factors Affecting the Posterior Probability

Three factors, prior odds, original sample size and the resulting BGA probability, are identified as affecting the posterior probability of Australian ancestry (PPAA) with varying amounts of sensitivity. A sensitivity analysis was performed using these three factors:

1. The GMAMP;
2. The sample size of the original data used to estimate the minimum genotype frequency using Green and Young's formula (Equation 4.8);
3. The prior odds ratio.

In the sensitivity analysis, only GMAMP relevant to the Australian and Japanese populations are shown. These values are obtained using the distributions observed in Figure 5.3, that is, the GMAMP observed in the individuals with all Australian ancestors' scenario, approximately > 0.9, and the individuals with all Japanese ancestor's scenario, approximately < 0.03.

The original data's sample size affects the PPAA during the calculation of the likelihood ratio, which is later used to estimate the PPAA. The ECDF curve is used to estimate the conditional probabilities of observing a GMAMP in each population. If a value of 0 is obtained, it is replaced using Green and Young's formula to account for the possibility of a rare event. This effect is demonstrated in Table 5.11.

Table 5.11: Conditional Probabilities of the GMAMP in the Two Populations and the Resulting LR
The conditional probabilities of observing the GMAMP in the Australian and Japanese ECDF curve. If a sample of zero was obtained, a conservative frequency was estimated (shown in brackets) using the Green and Young formula where the original data's sample size (n) = 100 (a) and 300 (b), demonstrating the effect of sample size on the resulting LR, and thus, the posterior probability (not shown).

a). Original Data's Sample Size = 100

	GMAMP			
	0.99	**0.95**	**0.9**	**0.01**
Australian Proportion	0.7987	0.9462	0.043	0 (0.03)
Japanese Proportion	0 (0.03)	0 (0.03)	0 (0.03)	0.9998
Resulting LR	26.62	31.54	1.43	0.03

b). Original Data's Sample Size = 300

	GMAMP			
	0.99	**0.95**	**0.9**	**0.01**
Australian Proportion	0.7987	0.9462	0.043	0 (0.01)
Japanese Proportion	0 (0.01)	0 (0.01)	0 (0.01)	0.9998
Resulting LR	79.87	94.62	4.3	0.01

As the sample size for the original data is increased, the conservative estimated frequency is reduced (using the examples n = 100 and 300) which alters the resulting likelihood ratio used to estimate the PPAA.

Table 5.12 extends on the results of Table 5.11, demonstrating how the original data's sample size and the prior odds ratio affect the resulting PPAA. Note that the likelihood ratio used to estimate each PPAA is not shown in Table 5.12.

Table 5.12: Posterior Probabilities of Australian Ancestry
Posterior probabilities of Australian Ancestry resulting from the sensitivity analysis where the GMAMP from the *p*LMT, sample size of original data and prior odds are combined. Note the LR used during the estimation is not shown.

a). Prior Odds = 0.5, the Japanese soldiers outnumber the Australian soldiers 2 to 1.

	Sample Size for Original Data (*n*)				
GMAMP	100	200	300	400	500
0.99	0.930	0.964	0.976	0.982	0.985
0.95	0.940	0.969	0.979	0.984	0.987
0.9	0.418	0.589	0.683	0.742	0.782
0.01	0.015	0.007	0.005	0.004	0.003

b). Prior Odds = 0.3, the Japanese soldiers outnumber the Australian soldiers 4 to 1.

	Sample Size for Original Data				
GMAMP	100	200	300	400	500
0.99	0.889	0.941	0.960	0.970	0.976
0.95	0.905	0.950	0.966	0.974	0.979
0.9	0.301	0.463	0.564	0.633	0.683
0.01	0.009	0.004	0.003	0.002	0.002

c). Prior Odds = 0.1, the Japanese soldiers outnumber the Australian soldiers 10 to 1.

	Sample Size for Original Data				
GMAMP	100	200	300	400	500
0.99	0.727	0.842	0.889	0.914	0.930
0.95	0.760	0.863	0.905	0.927	0.940
0.9	0.126	0.223	0.301	0.365	0.418
0.01	0.003	0.001	0.001	0.001	0.001

d). Prior Odds = 0.05 (Similar to the expected value for Buna in Papua New Guinea), the unrecovered Japanese soldiers outnumber the Australian soldiers 20 to 1.

	Sample Size for Original Data				
GMAMP	100	200	300	400	500
0.99	0.571	0.727	0.800	0.842	0.870
0.95	0.612	0.760	0.826	0.863	0.888
0.9	0.067	0.126	0.177	0.223	0.264
0.01	0.001	0.001	0.001	0.001	0.001

e). Prior Odds = 0.01 (Similar to the expected value for Sanananda in Papua New Guinea), the unrecovered Japanese soldiers outnumber the Australian soldiers 100 to 1.

	Sample Size for Original Data				
GMAMP	100	200	300	400	500
0.99	0.210	0.348	0.444	0.516	0.571
0.95	0.240	0.387	0.487	0.558	0.612
0.9	0.014	0.028	0.041	0.054	0.067
0.01	0.001	0.001	0.001	0.001	0.001

The results of the sensitivity analysis indicate that both the prior odds ratio and the original data's sample size have a substantial effect on the estimated PPAA. An individual with a GMAMP of 0.95, indicative of Australian BGA, combined with a prior odds ratio of 0.5 (Table 5.12a) would result in a PPAA of 0.94 (original data's sample size = 100). If that same individual was discovered in an area

where the appropriate prior odds ratio was 0.05 (Table 5.12d), the resulting PPAA would be reduced to 0.612, an ambiguous classification. In addition, for a prior odds ratio of 0.05 (Table 5.12d), a sample size increase from 100 to 500 can increase the resulting PPAA from a GMAMP of 0.95 from an ambiguous classification, PPAA = 0.612, to an Australian classification, PPAA = 0.888.

Ultimately, in areas with an extremely low prior odds ratio, such as 0.01 it is unlikely that any Australian remains will be classified unless the original sample size is increased substantially.

A limitation of this approach for incorporating the prior, is that it cannot be used when the obtained GMAMP is outside the expected distribution of GMAMPs for both populations, that is, the ECDF of conditional probabilities. If the GMAMP is not observed in either population's ECDF curve, the resulting LR will be 1, as the Green and Young formula will be used to update the zero value conditional probabilities obtained for both populations. However, in this scenario, this method of incorporating the prior may indirectly alert the user that the unknown test sample may not truly belong to either of the populations of interest, as the GMAMP is not observed in either of the populations. Additionally, this method of incorporating the prior is limited if the unknown test sample's GMAMP is on the extreme ends of their true population's ECDF conditional probability distribution. This limitation is demonstrated in Table 5.12, where the conditional probability of observing a GMAMP of 0.9 in the Australian population is 0.043. As the proportion of individuals within the ECDF curve of the Australian sample with an approximate GMAMP of 0.9 is small, the resulting conservative value estimated for the Japanese sample causes the likelihood ratio to also be small.

With the prior odds and original data's sample size having a noticeable effect on the PPAA, it is important to recognise how a confidence interval may also affect the decision-making process. The use of a confidence interval, and the choice of how it is reported, one-tailed or two-tailed, may cause the outputted PPAA to result in a different classification. This effect is another means of how sample size can alter the result, and the subsequent decision-making.

5.8 Classifying Alternative Populations

Individuals from alternative populations which may be of interest for the utilised case study were analysed using the *p*LMT classifier. Note that for the British, Chinese and American samples, all 40 of the GPSP SNPs were available for testing. For the Papuan sample, only 30 GPSP SNPs were available, resulting in a *p*LMT consisting of four LMT groups which utilised 26 of those SNPs. The distributions of the GMAMPs for each sample are shown in Figure 5.11.

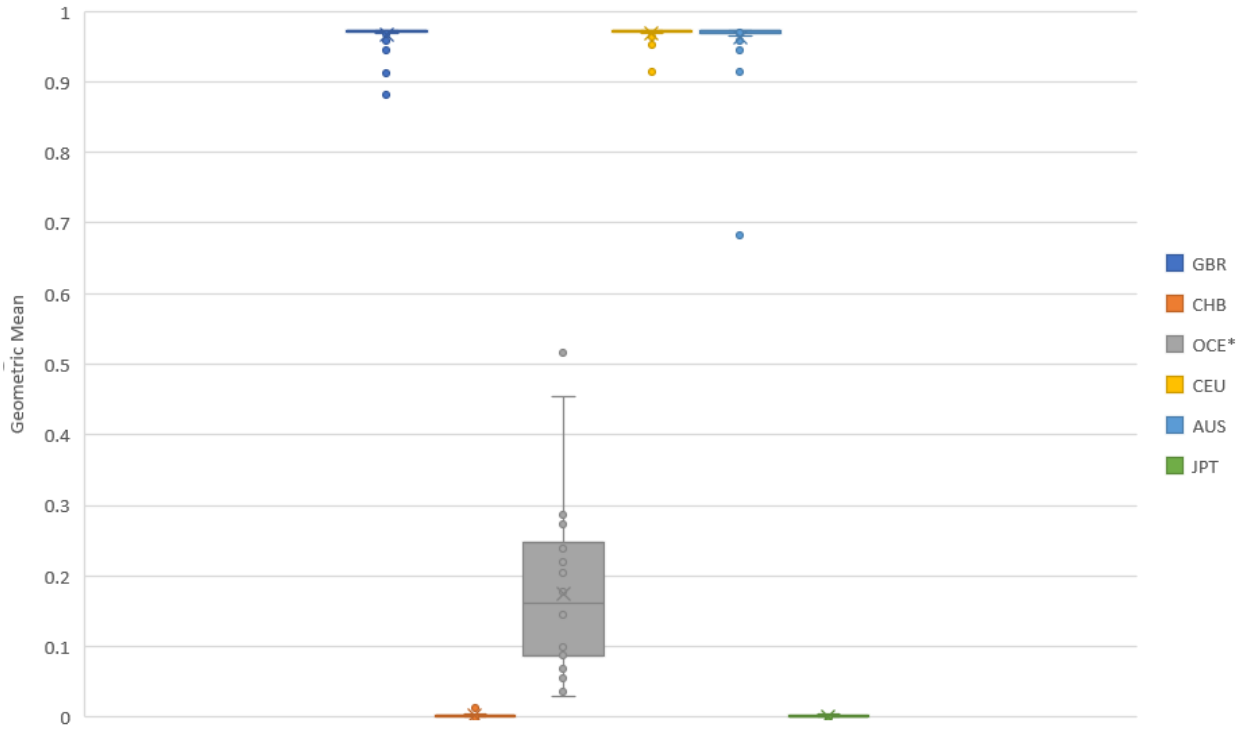


Figure 5.11: Distribution of GMAMPs for Alternative Populations
Observed GMAMP for alternative populations, British (GBR) = 91, Chinese (CHB) = 103, Papuan (OCE) = 26, American (CEU) = 99, Australian (AUS) = 108 and Japanese (JPT) = 104.

Since the *p*LMT classifier works on a binary model, additional populations are forced onto the same probability scale based on which of the two primary clusters (Australian or Japanese) they are most closely related towards. As expected, Caucasian-based populations such as British (GBR) and American (CEU) have GMAMPs which tend Australian, while Asian-based populations like Chinese (CHB) are more closely related to the Japanese population. Of interest is the Papuan (OCE) population, which was observed to have a distribution separate to the two primary clusters, indicating that it may be possible to distinguish Papuan individuals from other soldiers. However, as the Papuan sample only had 30 out of the GPSP's 40 SNPs available, it is possible that should the sample be reanalysed but with all 40 SNPs available, this distribution may differ.

5.9 Concluding Statement

In this chapter comparisons were made between three classifiers: (*i*) the Generic Bayesian, which is still commonly used in forensic practice, (*ii*) STRUCTURE, considered the 'golden standard' of BGA prediction, and (*iii*) the *p*LMT, a classifier proposed in this thesis for DNA-MAP. While all three classifiers performed equally for complete panel profiles, with regards to no observed direct errors, only STRUCTURE and the *p*LMT classifiers remained consistently effective for degraded DNA profiles, and with high accuracy, whereas the effectiveness of the Generic Bayesian increasingly decreased in a linear fashion with the number of SNPs missing. The ability to handle missing data is a key benefit for a BGA prediction classifier.

While STRUCTURE slightly outperformed the *p*LMT classifier in terms of having a reduced indirect error rate, it is important to compare other factors between the two remaining classifiers. To analyse Simulation Group 1 (n = 70,000) and Group 2 (n = 100,000), STRUCTURE took approximately four and six hours, respectively on a laptop. The equivalent analyses using the *p*LMT classifier took approximately <20 minutes. Therefore, the *p*LMT classifier offers significant advantages in terms of computing speed and with minimal reduction in classification accuracy.

Chapter 6 – DNA-Military Ancestry Predictor

6.1 Stages of DNA-MAP

There are three distinct stages to DNA-MAP's process: (*i*) the input stage, (*ii*) the statistical modelling stage, and (*iii*) the reported outputs stage. The following sections outlines each stage, detailing both the user-friendly front-end and the algorithms used in the background. In its current stage, DNA-MAP is a prototype, and suggestions of future updates are outlined which will be implemented prior to being made publicly available. This chapter has been written while still utilising the UWC-A case study to provide context to the reader.

6.1.1 The Input Stage

When launching DNA-MAP's Shiny application, available at (**https://dna-map.shinyapps.io/shinydeployed/**), the user is directed to the "Inputs" tab, which contains the following user-inputs. (Appendix 10 of this thesis provides a OneDrive and Dropbox link containing example files for Inputs *b*, *c* and *i*).

(a) "**Do You Wish to Input a Prior Odds?"** The first task the user must complete is selecting whether they wish to utilise a prior odds model. Two action buttons are presented to the user: (*i*) "No prior model is needed" and (*ii*) "I want to use a Prior Model". Note that by default, the "Prior Odds" model is enabled. Enabling or disabling a prior model will indicate to DNA-MAP whether to include or hide the tab "Prior". If the user selects not to utilise a prior model, the user will only be presented with an ancestry prediction statement containing the GMAMP. However, if the user does opt to utilise a prior model, two additional inputs (Inputs *i* and *j*) are then required, which will result in DNA-MAP to carrying out an analysis utilising an ECDF curve to estimate the posterior probability. Using the prior model, will result in the display of both the GMAMP and the posterior probability as separate statements in the output.

(b) "Training Data Containing Samples Collected from the Populations of Interest". A Comma Separated Values (.csv) file containing the raw genotype data for the two populations of interest (Table 6.1) is required. The file's required format is as follows: each row represents a single individual sampled from the populations of interest and each column should contain a single SNP with the given raw genotype shown for each individual. A column headed "POP" is required, which contains the name of an individual's declared population, which DNA-MAP uses as the classification variable (dependent variable) when performing the *p*LMT analysis. A search function in DNA-MAP's algorithm checks column names and identifies the column titled "POP" (case sensitive) as the classification variable. Thus: (*i*) each row is assumed to be one individual, (*ii*) a column named

"POP" is required and is taken to be the column which contains the dependent variable (in a logistic regression), (*iii*) all other columns are assumed to be explanatory variables (independent variables) and are used as such. The intended, default use of DNA-MAP, is for the outcome variable ("POP") to be binary and the rest of the variables to be SNP genotypes. It is required that the input file has no additional columns, the only columns are the outcome variable and the explanatory SNP variables. Therefore, the ordering of the columns does not affect the resulting *p*LMT, with the "POP" column be taken as the dependent variable and all other variables assumed to be outcome, SNP data.

Table 6.1: Example of the Population Data File
An example of the format required by the "Population Data" user-input (.csv) file, columns represent the genetic markers used with the last column containing the population variable used for classification.

rs1426654	rs9809818	rs28777	POP
AA	AA	AA	AUS
AA	AA	AC	AUS
AA	AA	AC	AUS
AA	AC	AA	AUS
GG	CC	CC	JPT
GG	CC	CC	JPT
GG	CC	CC	JPT
GG	CC	AC	JPT

(c) "Unknown Sample for Ancestry Prediction." A comma separated file (.csv) file containing the raw genotype data for an unknown individual (Table 6.2) is required. Note that in its current phase, DNA-MAP analyses one unknown sample per run. Note that at this stage of DNA-MAP's prototype development, only a single unknown sample can be analysed at a time due to the need for the SNPs present to drive the modelling. Therefore this file is assumed to contain data of a single individual. Section 6.2 outlines how prior to DNA-MAP becoming publicly available, the system will be altered to allow the user to analyse multiple unknown samples at a given time, with each sample having its own unique modelling performed for the SNPs available. This file requires an additional column to identify the unknown sample. The location of this column in the file is irrelevant, DNA-MAP is made aware of the column of interest through the user's compliance with Input (*h*). As the population to which this sample fits the best, is the outcome of this analysis, no equivalent to "POP" column of the previous file is assumed for the data file with the sample from the unknown sample. The remaining explanatory SNP columns in this file must match those that are present in the population data file. If a SNP is missing from the unknown sample under question, the cell for this SNP is left blank. This allows DNA-MAP to remove SNPs that are missing in the unknown individual from the original population data prior to *p*LMT analysis. Empty cells for SNPs in this file are removed and the values

of the remaining columns are recorded. The SNP columns that had missing values in this file, are then removed from the file containing the population data (Input (*b*)).

Table 6.2: Example Unknown Individuals File
An example of the format required by the "Unknown Data" user-input (.csv) file.

Unknown Sample ID	rs1426654	rs9809818	rs28777	rs …	rs4959270
UNK1	AA	AA	AA	…	GG

(d) "Max Number of pLMT Models Before Algorithm Stops". The maximum number of models to be generated by the *p*LMT algorithm before DNA-MAP forces model generation to stop and utilise the available models, must be provided. This user-input is included for cases where many SNPs are available (several hundred) and the process of generating models may exceed the user's computational limits. The default value is 50 and the available range is 1 – 1000.

(e) "Number of 10-fold Cross-Validation Iterations per Model". The number of instances that ten-fold cross-validation is performed and averaged for a single given *p*LMT model must be provided. Note that large values will substantially increase the algorithm's run-time. The default value is 10, and it is recommended that the user treats this value as a minimum. This input has been implemented as a measure of precision to ensure a consistent result is achieved when repeatedly fitting the same model to the same data. Therefore, while lowering this value does reduce DNA-MAP's overall run-time, the importance of a precise analysis should be prioritised wherever possible. The user may opt for a higher number of iterations in scenarios with: (1) a panel containing many markers (>100), (2) a panel which contains only SNPs with allele frequencies between 0.6 – 0.8, and (3) when using DNA-MAP for a real-world application to ensure high precision. The available range on this input is 1 to 1000.

(f) "Classification Threshold for a pLMT Model to be Accepted". The cut-off threshold that instructs DNA-MAP when to cease generating models must be provided. A model's average classification accuracy calculated over the multiple iterations of ten-fold cross-validation, is tested against this threshold to determine whether to accept the model or not. Default value is 0.99 and the available range is 0 to 1.

(g) "Desired Level of Confidence". The user's desired level of confidence in the result which is used during the estimation of confidence intervals for the output and when creating conservative values for the detection of a rare event (Equation 4.8). There are three options available, a 90%, 95% or a 99% level of confidence, which are presented in statements as a two-tailed Wilson confidence interval (See Chapter 4 for methodology). The choice of confidence should be based on one of the

following reasons: (1) the user's beliefs, (2) guidelines or standards that require a minimum level of confidence, or (3) on the available sample size, that is, knowing a reduced confidence is the only option due to a small sample size.

(h) "Name of Designated Identification Column in Your Unknown File". This input acts as a quality assurance measure, allowing the user to nominate which column in their unknown sample file (Input *c*) contains the sample's identification tag. DNA-MAP will find the column whose text string matches the input and record the identification code given to the unknown sample. The user simply needs to provide the column name they have used for identifying their sample (see Table 6.2, column 1).

Figure 6.1 provides a visual overview of the "Inputs" tab that the user is first directed to upon DNA-MAP's launch.

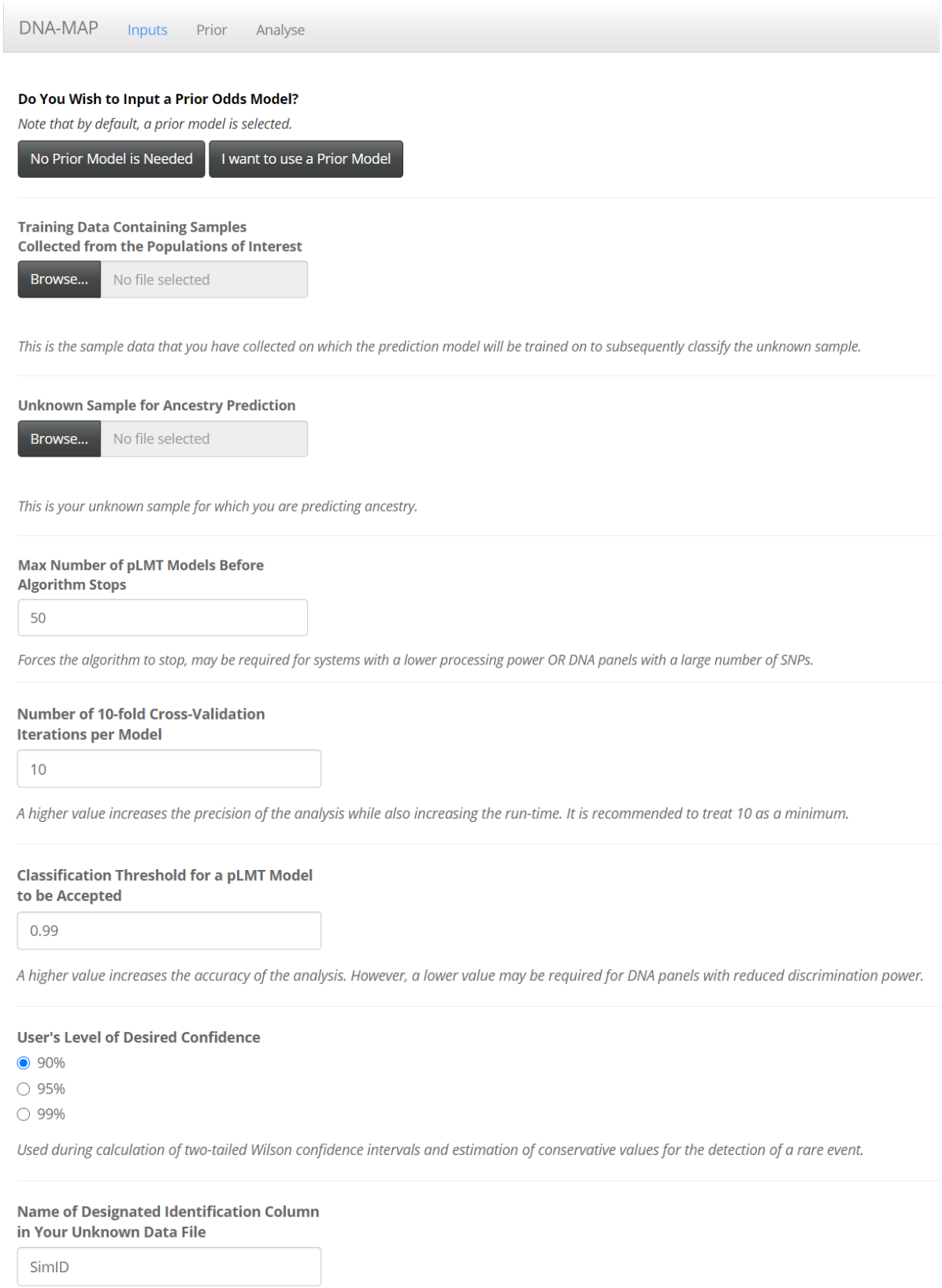


Figure 6.1: Prototype User-Interface – "Inputs" Tab.
A prototype of DNA-MAP's Shiny application user-interface for the initial Inputs stage of operations. Users will automatically begin on the Inputs tab upon launch of DNA-MAP where all user-input files and variables are uploaded.

If the user has enabled the "Prior Odds" Model, by selecting the "I want to use a Prior Odds" action button, the tab "Prior" appears at the top of the user's screen. This tab contains the two remaining inputs and indicates to DNA-MAP to calculate the posterior probability. Note that if the user has disabled the "Prior Odds" model, the "Prior" tab is removed from DNA-MAP's user interface, and the following two inputs are treated as "null" inputs and are no longer required by DNA-MAP.

(i) "Sample Data of Known Individuals not Included in the Training Data". A comma separated values file (.csv) containing the raw genotype values for a sample of individuals from both populations of interest, that have not been used in the model development (i.e. Input (*b*) above). Note that these individuals can be obtained through simulation or by taking a subset from the original training data prior to analysis, if the available sample size is large. For simulating individuals, the relative frequencies of alleles of the SNP panel in the available data (or obtained from the literature) can be used in conjunction with appropriate software such as SimAdmixtR (Kennedy, 2019). This file should be identical in format to the population data file (Input (*b*), Table 6.1). Missing SNPs observed in the individuals file to be classified (i.e. Input (*c*)) are also removed from this data file using the procedure described above. The data contained in this file are used to create the Empirical Cumulative Distribution Function (ECDF) curve required to incorporate the prior odds ratio.

(j) "Prior Knowledge of the Estimated Ratio of Two Populations for a Given Geographic Area". The prior odds value that the user has estimated for their situation. This value is assumed to be the ratio of the probabilities that an individual could belong to a given population versus another in a geographic area. To demonstrate, consider an area with 200 Australian soldiers, (n_a = 200) and 400 Japanese soldiers (n_b = 400). Therefore, the ratio of the probability that the individual could be an Australian soldier $\left(\Pr(Australian) = \frac{n_a}{n_a+n_b} = \frac{200}{200+400} = \frac{1}{3}\right)$ versus the probability that the individual could be a Japanese soldier $\left(\Pr(Japanese) = \frac{n_b}{n_a+n_b} = \frac{400}{200+400} = \frac{2}{3}\right)$, is equal to $\frac{\Pr(Australian)}{\Pr(Japanese)} = \frac{\frac{1}{3}}{\frac{2}{3}} = 0.5$. This value of 0.5 can also be interpreted as there are half as many Australian soldiers to Japanese soldiers for the area of interest.

Note that no default value is available for this input as it is assumed that for the user to enable the "Prior Odds" model in the first place, they must have access to some information regarding a possible estimate of the prior odds. Otherwise, if the "Prior Odds" model is disabled, Inputs *i* and *j* are not required. Figure 6.2 provides a visual overview of the optional "Prior" tab that the user has access to after enabling DNA-MAP to utilise the "Prior Odds" model.

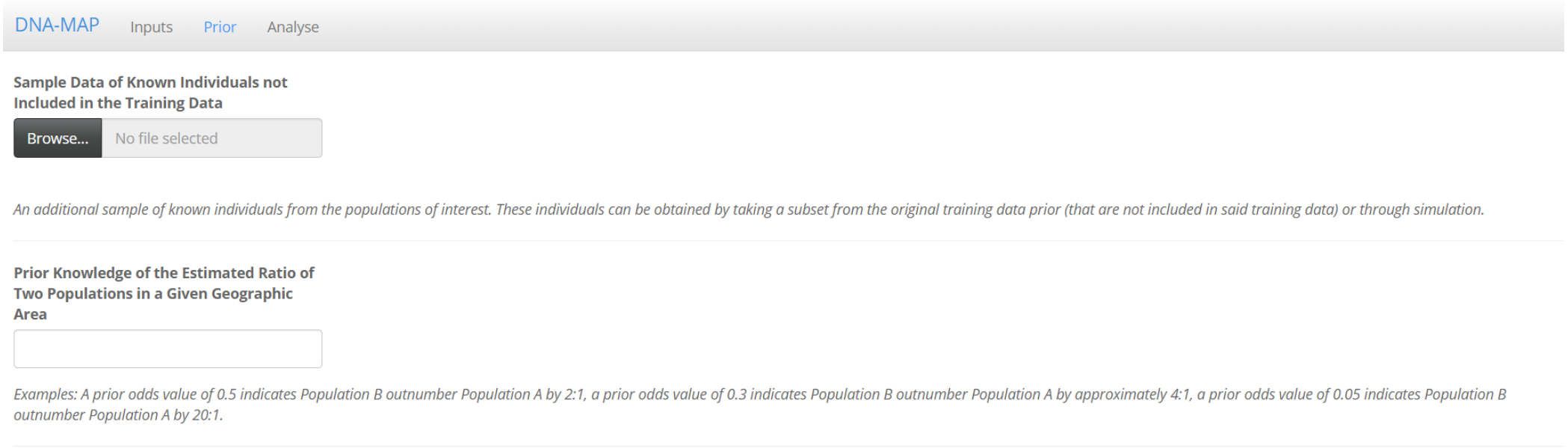


Figure 6.2: Prototype User-Interface – Prior Tab.
A prototype of DNA-MAP's Shiny application user-interface for the Prior tab. Users will have access to this tab once enabling a prior odds model using the action buttons available on the Inputs tab.

Once the user has successfully uploaded all required files and nominated all input variables, the user should then proceed to the "Analyse" tab to complete the process.

6.1.2 Statistical Modelling Stage

The following sections occur behind the front-end in DNA-MAP's algorithm, invisible to the user.

Match-Up Phase. The first phase of the analytical stage is the "match-up" process. SNPs that are missing for the test individual in the "*Unknown Sample for Ancestry Prediction*" file are removed from the "*Training Data Containing Samples Collected from the Populations of Interest*" and "*Sample Data of Known Individuals not Included in the Training Data*" files; the latter occurring only if the Prior Odds model is enabled. This process is performed as described in the "*Unknown Sample for Ancestry Prediction*" input. This allows the unknown individual to drive the subsequent *p*LMT models.

***p*LMT Phase.** With the user-inputs uploaded and the data files checked for errors, the *p*LMT algorithm is then applied to the "*Training Data Containing Samples Collected from the Populations of Interest*" file using the methodology as described in Chapter 4.3. A summary of the process follows, outlining where user-inputs are utilised.

The Process

1) DNA-MAP will select the file uploaded to "*Training Data Containing Samples Collected from the Populations of Interest*" input as the working dataset;
2) An LMT model is trained on the working dataset under *x* iterations of 10-fold cross-validation, where *x* is the numeric value for the "*Number of 10-fold Cross-Validation Iterations per Model*" user-input;
3) The obtained LMT's classification accuracy averaged over the *x* iterations is compared to the accuracy threshold which equals the numeric value for the "*Classification Threshold for a pLMT Model to be Accepted*" user-input;
4) If the LMT model's averaged classification accuracy is greater than or equal to the accuracy threshold then the model is accepted and stored internally. Any SNPs utilised in this LMT are then removed from the "*Training Data Containing Samples Collected from the Populations of Interest*" file. If the initial LMT model's averaged classification accuracy is below the accuracy threshold, no further analysis is performed;

5) If the model's accuracy was accepted, then Steps 2 – 4 are repeated continuously until one of the following three conditions is reached:
    a. There are no more SNPs available in the "*Training Data Containing Samples Collected from the Populations of Interest"* file;
    b. The number of generated LMT models has reached the maximum threshold which equals the numeric value for the "*Max Number of pLMT Models Before the Algorithm Stops*" user-input;
    c. The LMT model's averaged classification accuracy is less than the accuracy threshold.
6) Once one of these conditions is met, DNA-MAP ceases model creation and records the models created;
7) The regression coefficients are extracted from the models and are then applied one model at a time, to the respective individual in the "*Unknown Sample for Ancestry Prediction*" file to obtain the probabilities of membership from each model;
8) The geometric mean of the resulting probabilities of membership is then calculated, and a confidence interval is estimated using the methodology as shown in Section 4.10.

Posterior Probability Phase. This optional phase allows the incorporation of prior knowledge (such as historical records) of the expected relative population size with the estimated geometric mean to obtain a posterior probability. After the *p*LMT stage, if the user has enabled the "Prior Odds" model, the following steps are carried out:

1) DNA-MAP uses the user's "*Sample Data of Known Individuals not Included in the Training Data*" file as a working dataset of known individuals from both populations, where the SNPs missing from the unknown individual have been removed;
2) The individuals in the "*Sample Data of Known Individuals not Included in the Training Data*" file are analysed through the same *p*LMT model used for the unknown individual, creating a distribution of geometric means of ancestry for each population;
3) An empirical cumulative distribution function (ECDF) curve is constructed for each population's geometric mean of ancestry distribution;
4) The geometric mean of ancestry that was calculated for the individual in the "*Unknown Sample for Ancestry Prediction*" file is recorded (assume it is recorded as "*g*");
5) A likelihood ratio (LR) of *g*'s occurrence in each population is calculated by taking a cross-section of each population's respective ECDF (as shown in Section 4.8);
6) The posterior probability of Australian ancestry is then calculated using Equation 6.1.

7) A confidence interval is calculated for the posterior probability using the methodology outlined in Section 4.10 by first calculating a confidence interval to the LR and incorporating those limits into Equation 6.1.
    a. Note that smaller sample sizes (n < 150), may result in the lower confidence interval for the LR being negative. To ensure the resulting confidence interval for the posterior probability is not affected by this negative, an if statement is present to replace the negative LR with a minimum conservative estimate using the Green and Young method.

6.1.3 Reported Outputs Stage

The reported outputs stage summarises the information that is pertinent to the user from the analytical stage, which is provided to the user on the "Analysis" tab. Note that the length of the analysis stage will depend primarily on two factors: the user's selected Number of Iterations per Model and the total number of SNPs in the panel. For a value of 10 iterations of cross-validation per model on the GPSP (40 SNPs), DNA-MAP had a run time of approximately two minutes. Note that the GPSP was also tested with a value of 100 iterations which extended the software's runtime to 20 minutes. These runs were completed using a Windows operating system; results may vary for other operating systems.

No Prior Information Interpretation. If the user has disabled the "Prior Odds" model, the geometric mean is used to classify the individual sample under question. A simple statement is provided to the user describing the estimated geometric mean of the probability of Australian ancestry with the appropriate confidence intervals. See the following outputs (Figure 6.3).

Unknown Sample ID: As a quality assurance measure, the ID of the unknown sample is provided.

Statement of Results: An ancestry prediction statement (with no posterior probability) with its accompanying confidence interval.

Number of Models: The number of successful LMT models that were accepted and utilised for analysis.

SNPs not Used in Models: Which SNPs, if any, were unsuccessful in creating an LMT model above the required threshold. A ratio of the number of SNPs used versus the total available will also be provided.

Figure 6.3: Prototype User-Interface – Example Output with No Prior.
A prototype of DNA-MAP's Shiny application user-interface for the reported Output. Users are provided with these outputs by disabling the Prior Odds model.

Prior Incorporation Phase – Posterior Probability Interpretation. If the user has enabled the "Prior Odds" model, then in addition to the outputs provided the **No Prior Information Interpretation** section (Figure 6.3). The posterior probability, and its applied confidence intervals, are also provided in a simple, English statement for the user (Figure 6.4).

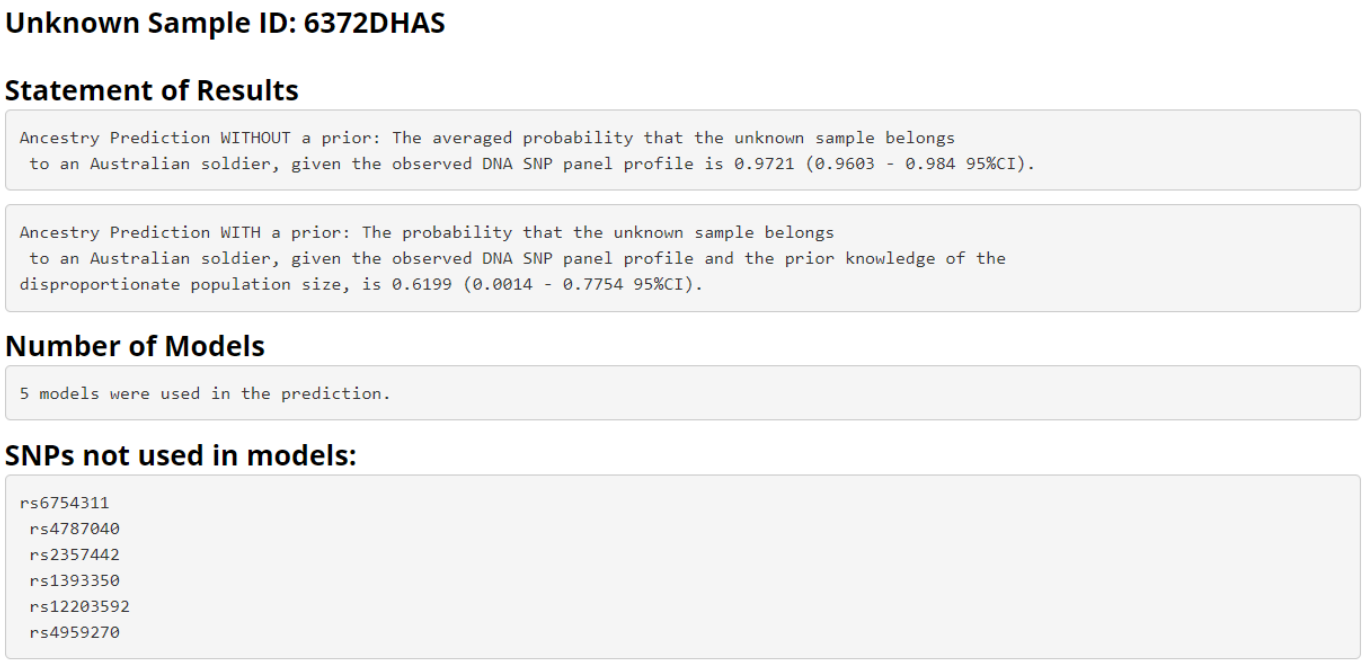


Figure 6.4: Prototype User-Interface – Example Output with Prior.
A prototype of DNA-MAP's Shiny application user-interface for the reported Output. Users are provided with these outputs by enabling the Prior Odds model.

Feedback Loop/Warnings: As the final interaction with the user, DNA-MAP may provide suggestions to the user on where alterations could be made to the inputted data and what effect these would have on the obtained BGA prediction. These suggestions, with their respective trigger and response, are:

Trigger: If the "*Training Data Containing Samples Collected from the Populations of Interest*" file contains a sample size less than 300 individuals (Note that this is based on the results found in this thesis, see Section 5.7).

Response: WARNING: The uploaded training data contains less than 300 individuals total. This may result in subsequent models not representing the true differences between the populations of interest. Additionally, resulting confidence intervals and the posterior probability will be affected. Consider increasing the sample size AND/OR lowering the desired level of confidence.

Trigger: If the number of available SNPs for the unknown sample in the "*Unknown Sample for Ancestry Prediction*" file is less than ten (Note that this is based on the results found in this thesis, see Section 5.6).

Response: WARNING: The number of available SNPs for the unknown sample is less than 10. These remaining SNPs may not provide the necessary discrimination for accurate ancestry prediction or

may result in informative models. Consider using a DNA panel with additional SNPs or lowering the threshold for model acceptance so more SNPs are utilised.

Trigger: The outputted number of successful models used in the analysis is less than three.

Response: WARNING: The number of models used to generate the ancestry prediction statements is less than three. These models may not provide adequate discrimination, or this may be the result of a small number of highly discriminating SNPs. Consider lowering the threshold for model acceptance.

6.1.4 Supplementary Download File

After the analysis is complete, the user has access to a download action button which will provide a supplementary comma separated values (.csv) file. This output contains the original unknown sample data uploaded in the "*Unknown Sample for Ancestry Prediction*" file, with the addition of multiple estimates. Those that are pertinent to the user are:

- "Pr_AUS" = GMAMP;
- "Lower_GM_CI" = Lower Confidence Interval for the GMAMP;
- "Upper_GM_CI" = Upper Confidence Interval for the GMAMP;
- "Post_PrAUS" = Posterior Probability of Australian Ancestry;
- "lower.Post_PrAUS" = Lower Confidence Interval for the Posterior Probability of Australian Ancestry;
- "upper.Post_PrAUS" = Upper Confidence Interval for the Posterior Probability of Australian Ancestry.

Note that these outputs will be renamed prior to DNA-MAP becoming publicly available to be relevant to the user's situation.

6.1.5 Optional User Manual

An important aspect of DNA-MAP is that it provides a user-friendly option to statistical prediction of BGA without the user requiring the necessary background in statistics or programming. However, there is still a minimum level of understanding that the user should have to ensure the software is being used correctly. DNA-MAP's GUI will provide brief statements for user inputs, phase transitions through DNA-MAP's process, outputs and any assumptions made during the modelling. These statements will allow the user to have a basic understanding to ensure proper application and interpretation. Beyond these statements, an optional user manual will be available for download upon DNA-MAP's official public access release that will go into greater detail of the software's operations. While the user manual will be available for download from DNA-MAP's front-end, it is

acknowledged that the greater level of detail will not be understood by all users, and rather, may serve to overwhelm certain users. Instead, it is provided as a secondary level of information for those who wish to learn more about DNA-MAP's processes.

6.1.6 Applicability to Other Populations

In its current format, DNA-MAP can perform the analysis on any two given populations. However, the majority of the outputted reports and estimates are written using the term "Australian soldier" to reflect the case study used to develop this prototype. If the user wished to utilise DNA-MAP for a case study outside of the UWC-A framework on which this prototype has been built, the following steps can be taken:

1. In the "*Training Data Containing Samples Collected from the Populations of Interest*" file, have the two populations still represented as "AUS" or "JPT".
2. Set the primary population that the user wishes to have ancestry prediction statements related to, as the "AUS" sample.

6.2 Future Changes to DNA-MAP

The following list contains a list of future directions/changes that will be made to DNA-MAP before being made publicly available" to improve its efficiency, user friendliness, and reduce the occurrence of computational errors.

Suggested Changes:

- An additional output which provides the user with a list of the inputs they have selected, that is, "The user selected a desired confidence interval level of *x*, a number of *y* iterations of 10-fold cross-validation…".
- If the user's inputted "*Training Data Containing Samples Collected from the Populations of Interest*" file does not contain any columns with the string name value "POP", DNA-MAP will opt to rename the last column on the population data file to this value, to ensure an outcome column is available. A warning prompt will be given to the user if this replacement occurs.
- Removing the "*Name of Downloadable Supplementary File*" input and having DNA-MAP instead name the output file to match the "*Unknown Sample for Ancestry Prediction*" file name with the addition of "…results".
- The inclusion of a progress bar, so that the user can determine the remaining run time for the analysis.

- The removal of unnecessary output estimates from the downloadable .csv supplementary file to avoid overwhelming the user.
- Rewrite DNA-MAP's coding to allow the user to nominate the populations they wish to utilise, removing the necessity to use the "AUS" and "JPT" population tags.
    - In addition, this change will allow DNA-MAP to provide better personalised statements.
- Inclusion of additional warnings and suggestions to the user based on evolutionary prototyping.
- Currently, when DNA-MAP's algorithm errors out, the Shiny server will crash, prompting the user to reload the server. This will later be changed to instead prompt the user with a personalised error message directing the user to which aspect of the analysis failed.
- Hide the "*Download*" action button until after the user has performed a successful analysis.
- The addition of various "data-cleansing" functions to reduce the possibility of clerical errors, these include.
- Provide an optional output to the user of a "power ranking" table of all SNPs provided. The table would rank all SNPs based on various informatic values commonly used in forensics such as δ and F_{ST}. Additionally, it would provide a binary response for whether a SNP was included or excluded in the *p*LMT models.
- Provide a calculator for the user to input the ratio of their given populations to estimate the prior probability.

Noted Bugs as of 04/02/2021

As with the future changes, below is a list of bugs that the creator is aware of and is currently fixing to improve DNA-MAP's efficiency.

- DNA-MAP has been observed to crash if the user sets both the "*Max Number of pLMT Models Before Algorithm Stops*" and "*Number of 10-fold Cross-Validation Iterations per Model*" inputs to one.
- The "*Download*" action button, when pressed prior to a successful analysis being performed, will instead download the Shiny page as a HTML file.

Appendix 11 documents the proposed structure of DNA-MAP's operational manual which will be subsequently released.

Chapter 7 – Discussion & Conclusion

7.1 Discussion

Accurate prediction of BGA has numerous real-world applications within the forensic science discipline such as corroboration of eyewitness accounts (Phillips, 2015), counterterrorism (Phillips et al., 2009), missing persons, Disaster Victim Identification (DVI), and criminal cases. However, BGA prediction is a complex process that requires multiple biological and statistical considerations. This thesis focuses on the statistical aspect of the process. To assign an individual to a population of origin based on an inferential process using evidence requires a decision-making framework. In this thesis, the case was developed that a KBDSS is one approach that could be utilised to support the decision-making process for inferring BGA. The developed KBDSS, named DNA-MAP, would collectively analyse and combine all pertinent information provided by the user regarding an unknown person's possible ancestry and output a concise BGA prediction report. The creation of DNA-MAP consisted of addressing two primary aspects: (*i*) the KBDSS itself, including what factors needed to be considered when creating a decision-support system (inputs and outputs), and (*ii*) the methodology used to predict BGA, involving the statistical aspects of the process which are required to ensure accurate inference. The following sections are structured as follows: a discussion regarding the KBDSS aspect of DNA-MAP, followed by the BGA prediction methods utilised and their various statistical aspects including assumptions and limitations, and finally conclusions drawn from this research.

7.1.1 Developing DNA-MAP

A review of published KBDSSs identified that although each system is addressing a unique problem, there are a number of "aspects" which were common in all cases. During the conceptualisation of DNA-MAP, two criteria were considered: (*i*) the general structure of a KBDSS along with the factors that should be considered during its construction and evaluation, and (*ii*) what additional factors/considerations would be required for a KBDSS specific for BGA prediction, to accommodate possible additional needs of forensic scientists. To address the former, a literature review was performed on KBDSSs from various disciplines, with four systems being selected from the literature for a detailed examination. Outlined below are the key concepts for a KBDSS which were identified in the literature, together with the approaches taken to develop them in DNA-MAP:

(i) Acknowledgement of an Issue. There must first be a need to address a pressing problem/issue in the discipline which has not yet been otherwise appropriately addressed at the time of the KBDSS's development. For example: (*i*) ineffective safety measures on oil and gas drilling sites – resulting in

serious injuries or death (Asad et al., 2019a), (*ii*) inefficient state of dairy farm conditions for milk production – resulting in a loss of profits due to less milk production than what could optimally be achieved (Kerr et al., 1999a, 1999b), and (*iii*) determining a suitable material for a pressure vessel – resulting in poor selection which could require material replacement (earlier than necessary) or risk of malfunction leading to workforce accidents (Yurdakul et al., 2020). In this thesis, the complex issue is inferring BGA for a set of unidentified remains obtained from a battlefield for UWC-A casework, with the repercussions being a soldier's remains being given to the wrong country, for example, an Australian soldier being sent to a Japanese War Cemetery.

(ii) An Extensive Knowledge Base. As the name implies a KBDSS's foundation is an extensive underlying knowledge base which the system uses during modelling. The knowledge base should define the issue that the KBDSS was designed for, and ultimately, allow comparisons and recommendations to the user (Mysiak et al., 2005). For example, in Kerr et al. (1999a)'s DAIRYPRO, the objective of the KBDSS was to inform the user where improvements could be made to dairy farm conditions to increase the average milk production to an optimal level. This optimal level was defined based on lengthy interviews with dairy farm experts with many years' experience in the field; it detailed the optimal conditions that a dairy farmer should adhere to in order to achieve maximum milk production. A KBDSS's knowledge base acts as a comparative tool, which the system treats as an optimal environment and to which comparisons can be made by the user's inputted values with suggestions. For DNA-MAP's equivalent, this knowledge base is the default settings (predefined in the R Shiny application and described in the user manual) that have been set by the experts (forensic scientist and statistician). These predefined settings allow a new user to operate the KBDSS without the need for investigative analyses themselves, however, they may wish to do so separately to inform their own knowledge as to what settings are suitable for their case.

(iii) Consultation with Experts and End Users in the Discipline.

The creation of DNA-MAP was a multidisciplinary project that expanded beyond the forensic science literature. Ultimately, the development included the use of statistics, machine learning, forensic biology and knowledge incorporated from other disciplines such as ecology, all with respect to the topic of classifying an unknown sample between two populations of interest. While the initial development stages of DNA-MAP were completed through a statistical modelling perspective, the final product could not have been readily achieved without consultation and input from the disciplinary expert. As stated by Kerr et al. (1999a), having an expert from the discipline involved in the KBDSS's construction can greatly improve the applicability of the system for users. For DNA-

MAP's construction, a forensic biologist with experience in ancestry analysis of unidentified remains (obtained during their work with historical military remains, missing persons and DVI units) was part of the supervisory team and had continuous input in the project development. During the early stages of DNA-MAP's development the expert was able to clearly define what inputs would be required in a typical BGA prediction setting; indicating that the historic information input may be necessary in certain cases. When DNA-MAP's development reached the stage of statistical modelling, the expert was able to provide the necessary information to ensure that any incorporated methodology adhered to current forensic science standards.

Originally, it was planned to consult end users (UWC-A members) during the final stages of DNA-MAP's development, so that evolutionary prototyping could occur, a common process during KBDSS development (Kerr et al., 1999a). In KBDSSs where evolutionary prototyping takes place, this consultation is performed to address questions that may arise during development, and to identify where clarification is required for given inputs or outputs. While direct evolutionary prototyping did not occur between the UWC-A due to time limitations, DNA-MAP and its research was presented at multiple international forensic and statistical conferences. From these presentations, multiple discussions occurred with other scientists and statisticians who were able to provide different insights and suggestions which were either incorporated into DNA-MAP's development or taken into consideration.

7.1.2 Influencing Factors of BGA Prediction

Cheung et al. (2017) described accurate BGA prediction as relying on three factors: (*i*) the selected SNPs used in the panel, (*ii*) the training data from the populations of interest, and (*iii*) the employed classifier. Based on an examination of previous BGA prediction studies from the literature, additional factors were identified in this thesis, which can affect the accuracy of results. The extended list of factors is provided below with their respective discussion.

Admixture. Admixture, in the context of genetic ancestry, can be defined as the combination of multiple divergent genetic lineages into a single gene flow as a result of geographic contact and interbreeding (Rius and Darling, 2014). The persistence of admixture in modern-day populations is a well-known issue that is prevalent during BGA prediction (Cheung et al., 2018a, 2018b; Phillips et al., 2014; Phillips, 2015). In current research, interest lies beyond simple classification of individuals into a single population, but rather to infer an individual's membership proportions to two or more populations. However, accurate estimation of an individual's admixture is extremely difficult and requires large samples from multiple populations to adequately train the utilised classifier. To

demonstrate with an example, consider an individual whose BGA membership proportions are estimated to be 50% Greek and 50% Romanian. Based on this, the ancestry of this individual could be inferred to be:

1) The individual has one parent from Greece and one parent from Romania;
2) The individual is from a country that is a mixture of these two populations;
3) The individual is from an entirely different country that was not sampled in the training data.

Base on genetic data alone, the above alternatives cannot be distinguished. Any difference between them is based on personal identity, namely, which population the individual declares as their origin. While it is relatively easy to select a group of biological markers to discriminate between two very different populations like Australian and Japanese, distinguishing between two populations with more similar genetic backgrounds would be much more complex. The historic background of two populations is another factor which can introduce ambiguity to the admixture issue. Two populations, which may consider themselves as two distinct populations based on differences such as creed or linguistics, may in fact be difficult to distinguish between on a genetic level.

Parsimony. The aim of incorporating a parsimonious model into DNA-MAP was to create an adaptive classification scheme which could find and determine the least number of markers needed to achieve the greatest discrimination. From a statistical perspective, the benefit of utilising parsimony is the reduction of noise (Crawley, 2012). In terms of a biological perspective, parsimony also has several benefits with respect to panel creation. When creating a SNP panel for ancestry analysis, the inclusion of each marker has an additional cost. A parsimonious approach to panel development can be considered cost effective, where, given the user's possibly limited resources, only the markers which will contribute to maximising the discrimination power. By drawing on information theory, DNA marker panels can be developed to only utilise the "best" (highest discrimination power either alone or in conjunction with other markers) markers (Rosenberg et al., 2003; Tal and Tran, 2018)

Models were only accepted by the *p*LMT algorithm if the classification accuracy was equal to or higher than the user's nominated threshold, the default value for this being set as 0.99. The model undergoes multiple repetitions of 10-fold cross-validation, where the average classification accuracy of these iterations is tested against the threshold, to get robust results (Landwehr et al., 2005). In the initial phases of the statistical development when multiple iterations of the cross validation were not implemented it was noticed that the same model fitted to the same data would result in a different number of models reaching the threshold for acceptance. This discrepancy was due to minor variation

in the accuracies obtained from the ten-fold cross-validation between runs generating the same models, despite the same datasets being used. Following the implementation of multiple iterations of cross-validation, and averaging the accuracy for all runs, the same number of models was generated each time. The benefit of this application is improving the method's robustness and ensuring that the same datasets can be analysed a number of times and result in the same number of models for each run (Landwehr et al., 2005). However, it should be noted that the number of specified iterations of cross-validation directly increases the computational time of the *p*LMT algorithm. During development, the following number of 10-fold cross-validation iterations were tested on the 40 SNP panel, 1, 10 and 100, which resulted in run-times of approximately less than one minute, two – three minutes, and twenty minutes, respectively.

Classifiers. Examining the BGA prediction literature provided several classifiers which had previously been tested and compared for their efficiency at inferring ancestry (such as Cheung et al. (2017, 2018a), McNevin et al. (2013) and Phillips (2015)). It was decided that these previous classifiers (STRUCTURE, Generic Bayesian, GDA, and MLR, see Cheung et al. (2017)) were not suitable for integration into DNA-MAP's system due to factors discussed in Cheung et al. (2017) such as extensive run times (STRUCTURE), relatively weaker performance (GDA and MLR), made use of assumptions which may not be true in a realistic scenario (Generic Bayesian and STRUCTURE), or make it difficult for the average user to determine which SNPs and what model have been utilised by the classifier (STRUCTURE), see Section 3.2.3. A new classifier was implemented for DNA-MAP's primary analysis, the LMT, which was ultimately adjusted to incorporate parsimonious modelling, resulting in the *p*LMT algorithm. To determine how effective the *p*LMT approach was, the Generic Bayesian and STRUCTURE classifiers were also utilised in subsequent analyses so that comparisons between the two could be made.

Comparing the three classifiers on complete panel profiles showed that for all three methods, no direct errors were observed. To compare this result to what has been previously observed in the literature, the "*0% Genotypes Missing*" section of Table 1 from Cheung et al. (2017, p.905) is used. For complete profiles, Cheung et al. (2017) observed their Generic Bayesian classifier and STRUCTURE to have a perfect 100% accuracy, based on AUROC curves. The comparable metric of the three classifiers (*p*LMT, GB and STRUCTURE) tested in this thesis, is the percentage of individuals with all Australian ancestors (Simulation Group 2, Scenario 1) classified as Australian. Compared to Cheung et al. (2017)'s result of 100% accuracy for STRUCTURE and the GB classifiers, the results in this thesis show an accuracy of 99.5% (*p*LMT), 100% (GB) and 100% (STRUCTURE), depicting the same outcome as observed in Cheung et al. (2017).

It is possible in real-world cases of BGA prediction for historical military remains that a full panel profile will not be obtained; markers may be missing due to degradation. It is imperative, therefore, that the classifier utilised by DNA-MAP can still readily infer ancestry for remains with incomplete panel profiles. To represent degraded samples, a second independent sample of WWII era Australians (n = 75) with varying degrees of missing data was utilised. For the 75 individuals, 100% were categorised as Australian using both the *p*LMT classifier and STRUCTURE, while only ≈53% (40 individuals) were categorised as Australian with the Generic Bayesian classifier, the remaining 35 individuals being classed as ambiguous. This degradation experiment demonstrates that STRUCTURE and the *p*LMT are still able to infer ancestry accurately even for heavily degraded samples, while the Generic Bayesian classifier was limited. However, the Cheung et al. (2017, Table 1, p.905) study, which also tested the Generic Bayesian's accuracy on degraded samples, observed the classifier to still retain a perfect 100% accuracy for samples with 90% SNPs removed (14 out of 142 remaining). The difference in results between the results of this thesis and Cheung et al. (2017) could potentially be due to two factors:

1. Panel Difference: The ability to classify degraded DNA samples relies on which markers are remaining, if markers with a high discrimination power remain, classification may still be possible with missing data. Therefore, Cheung et al. (2017)'s utilised 142 SNP panel may have contained a higher number of highly discriminative SNPs as opposed to the 40 SNP GPSP utilised in this research;
2. Strictness of Classification: For the degradation experiment performed in this thesis, degraded individuals were classified based on thresholds previously established on individuals with complete DNA panel profiles. It is possible that the Generic Bayesian's ability to classify degraded remains may have increased if different thresholds were used. However, establishing classification thresholds on a case-by-case basis dependent on which SNPs are available is a time-consuming process, and risks incorrect assignment due to fulfilling expected outcomes (confirmation bias).

Note that another similarity observed in this research is the minimum number of SNPs utilised in the SNP removal experiment performed in both this thesis and in Cheung et al. (2017)'s study. The results of both experiments determined that the study's highest-accuracy classifier, this theses' *p*LMT and Cheung et al. (2017)'s STRUCTURE, achieved high accuracy with as few as 10 SNPs.

Sample Size and Rare Event.

Sample size is critical for the efficiency of DNA-MAP, as is the case with any statistical modelling, and there are two instances where sample size has a direct effect on DNA-MAP's output. These are:

1) Development of models;

The series of *p*LMT models are generated based on training data that the user has uploaded. Models, and the SNPs incorporated into these models, are selected based on an algorithm (C4.5, refer Section 3.2.3) which measures the interactions between a set of one or more SNPs and determines those which are significantly informative in the process of separating the two known population groups (thus allowing classification to occur). Relative genotype frequencies obtained from a small sample size may not accurately reflect the true population (Chakraborty, 1992), which may lead to models being included or excluded when the opposite action may have been taken had the user provided a larger sample size.

2) Estimation of confidence intervals;

To estimate confidence intervals to the resulting probabilities, both the GMAMP of the probability of Australian BGA and the posterior probability of Australian BGA, the Delta method was used. Sample size is incorporated into the Delta method by decreasing (larger sample size) or increasing (lower sample size) the width of the confidence interval. The benefit of tighter confidence intervals is that it provides greater reassurance to the user that the applied intervals do in fact cover the true value (Brown et al., 2001).

Prior Probability.

As discussed by Budowle et al. (2011), estimating a value for the prior odds can be a complex process where multiple variables to be considered. Ultimately, the value selected by the user should be determined with reasonable justification. When using DNA-MAP, the user can incorporate their selected prior odds into the analysis, by combining the prior odds with a likelihood ratio of conditional probabilities to output a posterior probability of population membership (as per a "normal" Bayesian approach). Incorporating the prior probability through ECDF curves provided an intuitive method for calculating the posterior probability, which ensures the effect of the sample size is taken into account through the use of rare events (See Equation 4.8).

As expected, and confirmed through a sensitivity analysis, both the prior odds ratio and the sample size, had a dramatic effect on the resulting posterior probability of population membership. This is similar to what has been previously reported in the forensic literature. For example, Budowle et al. (2011) performed a sensitivity analysis of the posterior probability, in the context of missing persons, varying the above two factors, namely the prior odds and the sample size of the data in hand arriving at a similar conclusion as in this thesis.

7.2 Advantages of DNA-MAP

DNA-MAP provides the first KBDSS designed for informative ancestry prediction in the form of a user-friendly tool. The user interface at the inputs phase at the beginning of DNA-MAP's process allows the user to customise various measures to a certain degree based on their particular case scenario, without being involved in subsequent statistical analytical steps. This user-friendly front-end ensures that the tool is accessible to more users and does not require advanced knowledge of statistics or computer programming. Numerical inputs, such as thresholds and desired confidence level, have placeholder values to ensure that users can opt for the default classification settings in situations where knowledge regarding the optimal values for a given case are unknown. These default settings are based on the experimentation for the differentiation of the Australian and Japanese populations utilised in this thesis. However, the user is always encouraged to use informed choices for their dataset and to perform sensitivity analyses by using different input values to assess the robustness of the output. For example, of such analyses, see Chapter 5. In its current state, DNA-MAP is able to accommodate any marker panel available and is not specific to the markers used in the GPSP.

DNA-MAP, and its utilised *p*LMT algorithm, presents several new features of a BGA prediction tool that have not been previously demonstrated in the literature. The removal of SNPs in the relative population datasets based on which SNPs are missing for the unknown sample is an approach not readily performed by other BGA classifiers, rather, the user must perform this task beforehand. This removal process ensures that the unknown sample is the driving force of the classification system and that subsequent models are derived based on the available DNA from the unknown individual, as opposed to a generic model based on DNA present in the training datasets. Additionally, the proposed methodology for incorporating prior information allows DNA-MAP to provide the user with an outputted probability of predicted ancestry which is based on information obtained from (*i*) genetic information (the DNA panel using information theory), (*ii*) historic knowledge (prior odds ratio using Bayes' theorem) and (*iii*) sample size and the possibility of rare events. While other tools do incorporate some of these features, such as STRUCTURE's ability to incorporate a prior, DNA-

MAP is the first tool to combine all these features within the one software. Finally, DNA-MAP provides the reported probabilities of ancestry in simple statements for the user, unlike other tools such as STRUCTURE which only provide values which the user must then interpret themselves.

7.3 Future Directions

A number of future directions have been identified throughout this research and these are briefly outlined below.

1. Forensic Standard Validations

The scope of DNA-MAP's development in this thesis did not include validation to the level of forensic standards, namely, ensuring the software is consistent with guidelines such as ISO17025. Therefore, a future direction identified is to carry out testing to ensure the software and methods meet the standards of the forensic science community.

2. Alternative Pooling Methods

Utilising the *p*LMT classifier resulted in multiple independent estimates of the same measure, the probability of Australian ancestry given the respective model. To ensure reports were user friendly, these multiple estimates were pooled together into a single value using the geometric mean as a method of averaging (Manikandan, 2011). The benefits of using the geometric mean are that the method is simple, easily performed and is minimally affected by outliers. However, there is merit in having a pooling method which incorporates some form of weighting based on information theory (Shannon, 1948, Rosenberg et al., 2003). For example, combining multiple estimates which are all indicative of the same population should ideally increase the likelihood that the overall result should be the suggested population, regardless of any averaging process. This ideal stems from the information theory's philosophy of how many sections of the message are required before the full message becomes apparent. While other pooling methods were considered during this research, at this stage, only the geometric mean will be available during DNA-MAP's early stages. A future direction would be to investigate further pooling methods to determine which, if any, are appropriate for DNA-MAP to improve the overall discrimination power and reduce the number of Ambiguous classifications.

3. Expand to Encompass More Than Two Populations

Expanding DNA-MAP to perform ancestry prediction using more than two populations of interest will allow the application to have greater utility in forensic casework, such as the other possible applications suggested in Section 7.2. This expansion may allow DNA-MAP to be an alternative

classifier to STRUCTURE for more complex cases with several populations. However, to incorporate a multinomial classification scheme into DNA-MAP, an alternative algorithm may be used in place of the *p*LMT. For example, the R package "RWeka" (Hornik et al., 2009) through which the original LMT algorithm was obtained, provides a list of various machine-learning classification algorithms, some of which may be pertinent for future research.

4. Inclusion of Other Genetic Markers

While the research performed in this thesis used solely autosomal SNPs, there are mitochondrial DNA and Y-chromosome markers that can also provide information regarding BGA prediction. The UWC-A has already performed initial research regarding these other marker types (Barden, 2014; Best, 2014; Ghaiyed, 2016; Poulsen, 2015), see Ghaiyed (2020) for a detailed history of the UWC-A's use of lineage markers. However, there still may be utility in the future for these markers to be incorporated into DNA-MAP's modelling process.

7.4 Conclusion

The primary objective of the research undertaken in this thesis was to create a KBDSS that could assist investigators to predict BGA for unknown individuals. A case study involving historical military remains was used as proof of concept. The role of the software is to act as an information hub, where the user can upload data, which are analysed through several statistical analyses hidden behind a user-friendly interface. While BGA prediction is the prime directive of DNA-MAP, the software has the additional benefit of providing the user with various suggestions related to casework figures to inform the user where further resources can be directed to increase the rate of classifiable remains. For example, as shown in Chapter 5, the ability to confidently assign BGA to a set of remains becomes increasingly difficult where extremely low prior odd ratios (less than 0.1) is used. Through sensitivity testing in the form of "what-if" scenarios, the user can prioritise areas where it is expected that the prior odds ratio will have a smaller effect on the resulting GMAMP. By utilising a profile known to belong to an individual who originated from the smaller size population, the user can analyse the sample under question under multiple prior odd values to determine "if" these were realistic prior odd ratios, "what" would be the estimated posterior probability?

In addition to creating a KBDSS, a comparison of three classifiers was performed: the Generic Bayesian, which has been previously utilised in the forensic discipline, STRUCTURE, a commonly utilised program for BGA prediction, and the Parsimonious Logistic Model Tree, a classifier that was adapted in this thesis as a potential BGA predictor. It was found that while the three classifiers performed similarly in their ability to correctly infer ancestry for individuals with a complete panel

profile, the Generic Bayesian's ability to correctly classify a sample was reduced for degraded samples. As missing data are expected in forensic casework, a classifier which can readily predict ancestry for individuals who may not have all SNPs available is highly desired. While DNA-MAP has only been utilised on a single case study (the UWC-A's investigations in the South-East Asia Pacific), it is expected that in the future it will be tested on other databases and expand the software to be applied in other casework and be developed beyond the binary-approach.

DNA-MAP introduces to the field of BGA prediction several key functions which are either not currently present or are not readily available within a single tool. These include: (*i*) the unknown sample driving the classification model, (*ii*) the incorporation of prior information to estimate a posterior probability, (*iii*) accounting for the possibility of a rare event occurring, (*iv*) a parsimonious selection method for utilising only the most informative SNPs, (*v*) accounting for sampling error and its propagation across a function, (*vi*) stylised report generations with several informed feedback prompts to the user. All these functions are contained within a single user-friendly tool which does not require any training, has minimal file formatting and provides a clear statement of BGA with its associated margin of error. The outcome of this thesis is expected to be both the basis for future work as outlined above and provide forensic scientists with the support they need in ancestral decision making.

References

Adams, C. R. and Clarkson, J. A. (1934), "Properties of functions f (x, y) of bounded variation", *Transactions of the American Mathematical Society*, **36** (4), p.711-730.

Aitken, C. G. G. (1999), "Sampling – how big a sample?", *Journal of Forensic Science*, **44** (4), p.750 – 760.

Asad, M. M., Hassan, R. B., Ibrahim, N. H., Sherwani, F,. and Soomro, Q. M. (2018a), "Induction of decision making through accident prevention resources among drilling crew at oil and gas industries: a quantitative survey", *Paper presented at the Journal of Physics: Conference Series.*

Asad, M. M., Hassan, R. B., Sherwani, F., Ibrahim, N., and Soomro, Q. (2018b), "Level of satisfcation for occupational safety and health training activities: a broad spectrum industrial survey", *Paper presented at the Journal of Physics: Conference Series*.

Asad, M. M., Hassan, R. B., Sherwani, F., Aamir, M., Soomro, Q. M., and Sohu, S. (2019a), "Design and development of a novel knowledge-based decision support system for industrial safety management at drilling process", *Journal of Engineering, Design and Technology*.

Asad, M. M., Hassan, R. B., Sherwani, F., Abbas, Z., Shahbaz, M. S., and Soomro, Q. M. (2019b), "Identification of effective safety risk mitigating factors for well control drilling operation", *Journal of Engineering, Design and Technology*.

Australian Bureau of Statistics (1933), Census of the Commonwealth of Australia, 30th June 1933, *Census Bulletin No. 15*, Retrieved from https://www.ausstats.abs.gov.au/ausstats/free.nsf/0/9137363BE03BF19ACA2578EE001DE324/$File/1933%20Census%20-%20Bulletin%20No%2015.pdf

Australian Bureau of Statistics (2017), Census of the Commonwealth of Australia, *2016 Census Quickstats*, Retrieved from https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/036?opendocument

Australian Government – Department of Veteran Affairs (2009), "About the Kokoda Track: 1942 and today", Retrieved from https://anzacportal.dva.gov.au/wars-and-missions/kokoda-track-1942-1943/kokoda-track/about-kokoda-track-1942-and-today

Australian Government (1909), *Defence Act*, Retrieved from https://www.legislation.gov.au/Details/C1909A00015

Australian War Memorial (n.d.), "Australians at War" Retrieved from https://www.awm.gov.au/articles/atwar

Ballantyne, K., Bunford, J., Found, B., Neville, D., Taylor, D., Wevers, G., and Catoggio, D. (2017), "An introductory guide to evaluative reporting", *National Institute of Forensic Science, Australia New Zealand*.

Bardan, F. 2014, 'Establishing mtDNA haplogroup distributions to assist with ancestry identification of historical military remains', School of Biomolecular and Physical Sciences, Griffith University, Bachelor of Forensic Science with Honours.

Bateson, W., and Mendel, G. (1913), "Mendel's principles of heredity", University Press.

Beauthier, J. P., Valck, E., Lefevre, P., and Winnie, J. D. (2009), "Mass disaster victim identification: The tsunami experience", *The Open Forensic Science Journal*, **2** (1).

Beechey, F. S. (1876), "Electro-telegraphy", Published London, E. & F. N. Spon.

Berger, C. E., Buckleton, J., Champod, C., Evett, I, W., and Jackson, G. (2011), "Evidence evaluation: a response to the court of appeal judgement in R v T", *Science and Justice*, **51** (2), 43 – 49.

Best, M. 2014, 'Establishing an Australian mitochondrial DNA population database for identification of historical military remains', School of Biomolecular and Physical Sciences, Griffith University, Bachelor of Forensic Science with Honours.

Black, P. and Stockton, T. (2009), "Basic steps for the development of decision support systems", *Decision support systems for risk-based management of contaminated sites*, p. 1 – 27.

Bongers, J. L., Nakatsuka, N., O'Shea, C., Harper, T. K., Tantaleán, H., Stanish, C., and Fehren-Schmitz, L. (2020), "Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast", *Proceedings of the National Academy of Sciences*, **117** (31), p.18359 – 18368.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), "Classification and regression trees", *Wadsworth & Brooks*. Cole Statistics/Probability Series.

Brenner, C. H. (2010), "Fundamental problem of forensic mathematics – the evidential value of a rare haplotype", *Forensic Science International: Genetics*, **4** (5), p.281 – 291.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval estimation for a binomial proportion", *Statistical science*, p.101 – 117.

Budowle, B., Giusti, A. M., Waye, J. S., Baechtel, F. S., Fourney, R. M., Adams, D. E., Presley, L. A., Deadman, H. A. and Monson, K. L., (1991) "Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons", *American journal of human genetics*, **48** (5), p.841

Budowle, B., Ge, J., Chakraborty, R., and Gill-King, H. (2011), "Use of prior odds for missing persons identifications", *Investigative Genetics*, **2** (1), p.15.

Budowle, B., Monson, K., and Chakraborty, R. (1996), "Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci", *International Journal of Legal Medicine*, **108** (4), p.173 – 176.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., and Yang, J. (2015), "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies", *Nature Genetics*, **47** (3), p.291 – 295.

Cavalli-Sforza, L. L., and Edwards, A. W. (1967), "Phylogenetic analysis: models and estimation procedures", *Evolution*, **21** (3), p.550 – 570.

Cereda, G. (2017), "Bayesian approach to LR assessment in case of rare type match", *Statistics Neerlandica*, **71** (2), p.141 – 164.

Cereda, G., Gill, R. D., and Taroni, F. (2018), "A solution for the rare type match problem when using the DIP-STR marker system", *Forensic Science International: Genetics*, **34**, p.88 – 96.

Cereda, G. and Gill, R. D. (2020), "A nonparametric Bayesian approach to the rare type match problem, *Entropy*, **22** (4), p.439.

Chakraborty, R. (1992), "Sample size requirements for addressing the population genetic issues of forensic use of DNA typing", *Human biology*, p.141 – 159.

Chakraborty, R., Srinivasan, M. R., and Daiger, S. P. (1993), "Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics", *American journal of human genetics*, **52** (1), p.60.

Cheung, E. Y., Gahan, M. E., and McNevin, D. (2017), "Prediction of biogeographical ancestry from genotype: a comparison of classifiers", *International Journal of Legal Medicine*, **131** (4), p.901 – 912.

Cheung, E. Y., Gahan, M. E., and McNevin, D. (2018a), "Prediction of biogeographical ancestry in admixed individuals", *Forensic Science International: Genetics*, **36**, p.104 – 111.

Cheung, E. Y., Gahan, M. E., and McNevin, D. (2018b), "Predictive DNA analysis for biogeographical ancestry" *Australian Journal of Forensic Sciences*, **50** (6), p.651 – 658.

Cheung, E. Y., Phillips, C., Eduardoff, M., Lareu, M. V., and McNevin, D. (2019), "Performance of ancestry-informative SNP and microhaplotype markers", *Forensic Science International: Genetics*, **43**, p.102141.

Chow, S.-C., Wang, H., and Shao, J. (2007), "Sample size calculations in clinical research", *CRC press*, Cambridge.
Crawley, M. J. (2012), "The R book", *John Wiley & Sons*.
Curran, J., Buckleton, J., Triggs, C., and Weir, B. (2002), "Assessing uncertainty in DNA evidence caused by sampling effects", *Science and Justice*, **42** (1), p.29 – 37.
Essen-Möller, E. (1938), "Die beweiskraft der ahnlichkert in Vaterschaftsnachweis; theoretische grandlagen", *Mitt Anthr Ges*, Vienna, **68**, p.9 – 53.
Falush, D., Stephens, M., and Pritchard, J. K. (2003), "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies", *Genetics*, **164** (4), p.1567 – 1587.
Falush, D., Stephens, M., and Pritchard, J. K. (2007), "Inference of population structure using multilocus genotype data: dominant markers and null alleles", *Molecular Ecology Notes*, **7** (4), p.574 – 578.
Genomes Project Consortium (2015), "A global reference for human genetic variation", *Nature*, **526** (7571), p.68 – 74.
Gettings, K. B., Borsuk, L. A., Steffen, C. R., Kiesler, K. M., and Vallone, P. M. (2018), "Sequence-based US population data for 27 autosomal STR loci", *Forensic Science International: Genetics*, **37**, p.106 – 115.
Ghaiyed, A. (2016), "Estimation of genetic admixture in the Australian WWII era population to assist in repatriation of Australian soldiers", Bachelor of Forensic Science with Honours, Griffith University.
Ghaiyed, A. (2020), "A Genomic Ancestry Panel for Australian and Japanese WWII Military Remains Recovered in the Asia-Pacific", Doctor of Philisophy, Queensland University of Technology.
Gojanovic, T. (2007), "Zero defect sampling", *Quality Progress*, **22**, p.72.
González-Ferrer, A., Seara, G., Cháfer, J., and Mayol, J. (2018), "Generating Big Data Sets from Knowledge-based Decision Support Systems to Pursue Value-based Healthcare", *International Journal of Interactive Multimedia & Artificial Intelligence*, **4** (7).
Good, I. J. (1979), "Studies in the history of probability and statistics: XXXVII A. M. Turing's statistical work in World War II", *Biometrika*, p.393 – 396.
Graydon, M., Cholette, F., and Ng, L. K. (2009), "Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits", *Forensic Science International: Genetics*, **3** (4), p.251 – 254.
Green, R. H., and Young, R. C. (1993), "Sampling to Detect Rare Species" *Ecological Applications*, **3** (2), p.351 – 356.
Griffith, K. (2019), "DNA testing company that charges up to $298 to tell customers their 'superpowers' and give 'personalized' health advice is accused of fabricating results and offering generic tips to 'eat kale' and 'wear sunscreen'", *Dailymail.com*, Retrieved from https://www.dailymail.co.uk/news/article-7453175/DNA-testing-company-Orig3n-accused-fabricating-results.html
Hackshaw, A. (2008), "Small studies: strengths and limitations", *European Respiratory Society*, p. 1141 – 1143.
Harvard Medical School. (2019), "Largest-ever ancient-DNA study illuminates millennia of South and Central Asian prehistory" Retrieved from https://www.sciencedaily.com/releases/2019/09/190905145348.htm#:~:text=2-,Largest%2Dever%20ancient%2DDNA%20study%20illuminates%20millennia%20of,South%20and%20Central%20Asian%20prehistory&text=Summary%3A,the%20ancient%20Indus%20Valley%20Civilization.

Hssina, B., Merbouha, A., Ezzikouri, H., and Erritali, M. (2014), "A comparative study of decision tree ID3 and C4. 5", *International Journal of Advanced Computer Science and Applications*, **4** (2), p.13 – 19.
Holzworth, D. P., Huth, I. N., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., Chenu, K. et al. (2014), "APSIM – Evolution towards a New Generation of Agricultural Systems Simulation.", *Environmental Modelling & Software* **62**, p.327–350.
Hornik, K., Buchta, C., Zeileis, A. (2009), "Open-Source Machine Learning: R Meets Weka.", *Computational Statistics*, **24** (2), p.225-232. doi: 10.1007/s00180-008-0119-7.
Humphry, R. W., Cameron, A., and Gunn, G. J. (2004), "A practical approach to calculate sample size for herd prevalence surveys", *Preventive Veterinary Medicine*, **65** (3 – 4), 173 – 188.
Hystad, G., Downs, R. T., and Hazen, R. M. (2015), "Mineral species frequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals", *Mathematical Geosciences*, **47** (6), p.647 – 661.
Jackson, K. (2019), "When World War II came to PNG: The 10 key battles of 1942", Retrieved from https://www.pngattitude.com/2019/04/when-world-war-ii-came-to-png-the-10-key-battles-of-1942.html
Jin, S., Chase, M., Henry, M., Alderson, G., Morrow, J. M., Malik, S., Ballard, D., McGrory, J., Fernandopulle, N., Millman, J., and Laird, J. (2018), "Implementing a biogeographic ancestry inference service for forensic casework", *Electrophoresis*, **39** (21), p.2757 – 2765.
Jo, J., Park, J., Ji, H., Yang, Y., and Lim, H. (2016), "A study on factor analysis to support knowledge based decisions for a smart class", *Information Technology and Management*, **17** (1), p.43 – 56.
Jovanovic, B. D., and Levy, P. S. (1997), "A Look at the Rule of Three", *The American Statistician*, **51** (2), p.137 – 139.
Jung, H., and Chung, K. (2016), "Knowledge-based dietary nutrition recommendation for obese management", *Information Technology and Management*, **17** (1), p.29 – 42.
Kalinowski, S. T. (2011), "The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure", *Heredity*, **106** (4), p.625 – 632.
Kamangar, F., and Islami, F. (2013), "Sample size calculation for epidemiologic studies: principles and methods", *Archives of Iranian Medicine (AIM)*, **16** (5).
Kass, G. V. (1980), "An exploratory technique for investigating large quantities of categorical data", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **29** (2), p.119 – 127.
Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N., Meinke, H., Hochman, Z. and McLean, G. (2003), "An overview of APSIM, a model designed for farming systems simulation.", *European journal of agronomy*, **18** (3-4), p.267-288.
Kennedy, D. (2019), "SimAdmixtR: Simulates a set of admixed diploid genotypes based on SNP allele frequencies", Retrieved from https://github.com/danwkenn/SimAdmixtR
Kerr, D., Cowan, R., and Chaseling, J. (1999a), "DAIRYPRO—a knowledge-based decision support system for strategic planning on sub-tropical dairy farms. I. System description", *Agricultural Systems*, **59** (3), p.245 – 255.
Kerr, D., Chaseling, J., Chopping, G., and Cowan, R. (1999b), "DAIRYPRO—a knowledge-based decision support system for strategic planning on sub-tropical dairy farms. II. Validation", *Agricultural Systems*, **59** (3), p.257 – 266.
Khrameeva, E. E., Bozek, K., He, L., Yan, Z., Jiang, X., Wei, Y., Tang, K., Gelfand, M. S., Prufer, K., Kelso, J., and Paabo, S. (2014), "Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans" *Nature Communications*, **5** (1), p.1 – 8.

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., and Kidd, K. K. (2011), "Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples", *Investigative Genetics*, **2** (1), p.1.
Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R., and Kidd, J. R. (2014), "Progress toward an efficient panel of SNPs for ancestry inference", *Forensic Science International: Genetics*, **10**, p.23 – 32.
Kidd, K. K., Soundararajan, U., Rajeevan, H., Pakstis, A. J., Moore, K. N. and Ropero-Miller, J. D., (2018), "The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb", *Forensic Science International: Genetics*, **33**, p.33 – 37.
King, G., and Zeng, L. (2001), "Explaining rare events in international relations", *International Organization*, **55** (3), p.693 – 715.
Kotz, S. (2006), "Geometric Distribution", *S. Kotz, N. Balakrishnan, C. B. Read, B. Vidakovic, & N. L. Johnson Eds*.
Krebs, C. J. (1989), "Ecological methodology", *Harper & Row New York*.
LaFramboise, T. (2009), "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances." *Nucleic acids research*, **37**, (13) p.4181-4193.
Landwehr, N., Hall, M., and Frank, E. (2005), "Logistic model trees", *Machine Learning*, **59** (1-2), p.161 – 205.
Leni, S., Supriadi, L., Pheng, S., Evelyn, A., and Hwang, B. (2013), "A knowledge based decision support system (KBDSS) for Indonesian contractors to implement business continuity management (BCM)", Paper presented at the Conference: *CIB World Building Congress 2013*.
Letzer, R. (2018), "I Took 9 Different Commercial DNA Tests and Got 6 Different Results", Retrieved from https://www.livescience.com/63997-dna-ancestry-test-results-explained.html
Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., and Myers, R. M. (2008), "Worldwide human relationships inferred from genome-wide patterns of variation", *Science*, **319** (5866), p.1100 – 1104.
Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C. J., and Kayser, M. (2009), "Eye color and the prediction of complex phenotypes from genotypes", *Current Biology*, **19** (5), p.192 – 193.
Lombardi, O., Holik, F., and Vanni, L. (2016), "What is Shannon information?", *Synthese*, **193** (7), p.1983 – 2012.
Lowe, A. L., Urquhart, A., Foreman, L. A., and Evett, I. W. (2001), "Inferring ethnic origin by means of an STR profile", *Forensic Science International*, **119** (1), p.17 – 22.
Lucid Chart. (n.d.), "What is a decision tree diagram", *Website*, Retrieved from https://www.lucidchart.com/pages/decision-tree
Makuch, R. W. (2006), "Detecting rare adverse events in postmarketing studies: sample size considerations", *Drug information journal: DIJ/Drug Information Association*, **40** (1), p.89 – 98.
Manikandan, S. (2011), "Measures of central tendency: The mean", *Journal of Pharmacology and Pharmacotherapeutics*, **2** (2), p.140.
Marano, L. A., and Fridman, C. (2019), "DNA phenotyping: current application in forensic science", *Research and Reports in Forensic Medical Science*, **9**, p.1 – 8.
Marin, G. (2008), "Decision support systems", *Journal of Information Systems & Operations Management*, **2** (2), p.513 – 520.
Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F., and Hicks, T. (2016), "Discussion on how to implement a verbal scale in a

forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings", *Science and Justice*, **56** (5), p.364 – 370.
McCown, R. L., Hammer, G. L., Hargreaves, J. N. G., Holzworth, D. and Huth, N. I. (1995), "APSIM: an agricultural production system simulation model for operational research.", *Mathematics and computers in simulation*, **39** (3-4), p.225-231.
McNevin, D., Santos, C., Gómez-Tato, A., Álvarez-Dios, J., de Cal, M. C., Daniel, R., Phillips, C., and Lareu, M. (2013), "An assessment of Bayesian and multinomial logistic regression classification systems to analyse admixed individuals", *Forensic Science International: Genetics Supplement Series*, **4** (1), p.63 – 64.
Miller, F., Zohar, S., Stallard, N., Madan, J., Posch, M., Hee, S. W., Pearce, M., Vagero, M., and Day, S. (2018), "Approaches to sample size calculation for clinical trials in rare diseases", *Pharmaceutical statistics*, **17** (3), p.214 – 230.
Mingers, J. (1989), "An empirical comparison of pruning methods for decision tree induction", *Machine Learning*, **4** (2), p.227 – 243.
Mogensen, H. S., Tvedebrink, T., Børsting, C., Pereira, V., and Morling, N. (2020), "Ancestry prediction efficiency of the software GenoGeographer using a z-score method and the ancestry informative markers in the Precision ID Ancestry Panel", *Forensic Science International: Genetics*, **44**, p.102154.
Morvan, G., Jolly, D., Dupont, D., and Kubiak, P. (2007), "A decision support system for forensic entomology", *EUROSIM*, p.311.
Mysiak, J., Giupponi, C., and Rosato, P. (2005), "Towards the development of a decision support system for water resource management", *Environmental Modelling and Software*, **20** (2), p.203 – 214.
Nassar, M., Khamis, S., and Radwan, S. (2011), "On Bayesian sample size determination", *Journal of Applied Statistics*, **38** (5), p.1045 – 1054.
National Museum of Australia (n.d.), "1901: Inauguration of the Commonwealth of Australia", Retrieved from https://www.nma.gov.au/defining-moments/resources/federation
National Research Council (1996), "The evaluation of forensic DNA evidence", *National Academies Press*.
Nei, M. (1972), "Genetic distance between populations", *The American Naturalist*, **106** (949), p.283 – 292.
Noor, N. M. M., Ghazali, A. F., Saman, M. Y. M., Zainuddin, Z., Harun, M. I. H., and Abdullah, M. C. (2014), "Evolutionary Framework of a Decision Support System for Forensic DNA Analysis", *Lecture Notes on Software Engineering*, **2** (2), p.150.
Oatley, G., Ewart, B., and Zeleznikow, J. (2006), "Decision support systems for police: Lessons from the application of data mining techniques to "soft" forensic evidence", *Artificial Intelligence and Law*, **14** (1-2), p.35 – 100.
Oehlert, G. W. (1992), "A note on the Delta method", *The American Statistician*, **46**, p.27 – 29.
Ogden, R., Dawnay, N., and McEwing, R. (2009), "Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement", *Endangered Species Research*, **9** (3), p.179 – 195.
Openstax (n.d.), "The Genetic Code", Retrieved from https://cnx.org/contents/GFy_h8cu@9.87:QEibhJMi@8/The-Genetic-Code
Pajnič, I. Z., Pogorelc, B. G., and Balažic, J. (2010), "Molecular genetic identification of skeletal remains from the Second World War Konfin I mass grave in Slovenia", *International Journal of Legal Medicine*, **124** (4), p.307 – 317.
Pardo-Seco, J., Martinón-Torres, F., and Salas, A. (2014), "Evaluating the accuracy of AIM panels at quantifying genome ancestry", *BMC Genomics*, **15** (1), p.543.

Parliament of Australia. (n.d.), "Asian Immigration", Retrieved from https://www.aph.gov.au/sitecore/content/Home/About_Parliament/Parliamentary_Departments/Parliamentary_Library/Publications_Archive/CIB/CIB9697/97cib16

Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002), "An introduction to logistic regression analysis and reporting", *The Journal of Educational Research*, **96** (1), p.3 – 14.

Perner, P. (2015), "Decision tree induction methods and their application to big data", *Modeling and Processing for Next-Generation Big-Data Technologies*¸ Springer, p. 57 – 88.

Phillips, C. (2015), "Forensic genetic analysis of bio-geographical ancestry", *Forensic Science International: Genetics*, **18**, p.49 – 65.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Borsting, C., Johansen, P., Fondevila, M., and Morling, N. (2014), "Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set", *Forensic Science International: Genetics*, **11**, p.13 – 25.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brion, M., Montesino, M., and Carracedo, A. (2009), "Ancestry analysis in the 11-M Madrid bomb attack investigation". *PloS one*, **4** (8).

Phillips, C., Salas, A., Sanchez, J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M. C., Ballard, D., Lareu, M., and Carracedo, A. (2007), "Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs," *Forensic Science International: Genetics*, **1** (3-4), p.273 – 280.

Pick, R. A. (2008), "Benefits of decision support systems. In Handbook on Decision Support Systems 1", *Springer*, p. 719 – 730.

Pick, R. A., and Weatherholt, N. (2013), "A review on evaluation and benefits of decision support systems", *Review of Business Information Systems (RBIS)*, **17** (1), p.7 – 20.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., and Lareu, M. (2013), "An overview of STRUCTURE: applications, parameter settings, and supporting software", *Frontiers in Genetics*, **4**, p.98.

Poulsen, F. (2015), "Construction of an Australian Y-haplogroup database to assist with ancestry identification of historical military remains", Bachelor of Forensic Science with Honours, *Griffith University*,

President's Council of Advisors on Science and Technology (2016), "Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods", *Executive Office of the President of the United States.*

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000), "Inference of population structure using multilocus genotype data", *Genetics*, **155** (2), p.945 – 959.

Pritchard, J. K., Wen, X., and Falush, D. (2009), "Documentation for structure software: Version 2.3", *Department of Human Genetics University of Chicago*. Retrieved from http://pritch.bsd.uchicago.edu/structure.html

Quinlan, J. (1986), "Induction of Decision Trees.", *Machine Learning*, **1** (1), p.81 – 106.

Quinlan, J. R. (2014), "C4. 5: programs for machine learning", *Elsevier*. California.

Quinlan, J. R. (1992), "Learning with continuous classes", *Paper presented at the 5th Australian joint conference on artificial intelligence, Hobart.*

R Core Team. (2019), "R: A language and environment for statistical computing", *R Foundation for Statistical Computing Vienna Austria: R Foundation for Statistical Computing Vienna Austria*, Retrieved from https://www.R-project.org/

Rajeevan, H., Soundararajan, U., Kidd, J. R., Pakstis, A. J., & Kidd, K. K. (2012), "ALFRED: an allele frequency resource for research and teaching," *Nucleic Acids Research*, **40**, p.1010 – 1015.

Rakow, T., Demes, K. A., and Newell, B. R. (2008), "Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice", *Organizational Behavior and Human Decision Processes*, **106** (2), p.168 – 179.
Ramos, E., Doumatey, A., Elkahloun, A., Shriner, D., Huang, H., Chen, G., Zhou, J., McLeod, H., Adeyemo, A., and Rotimi, C. (2014), "Pharmacogenomics, ancestry and clinical decision making for global populations", *The Pharmacogenomics Journal*, **14** (3), p.217.
Regalado, A. (2019), "More than 26 million people have taken an at-home ancestry test", Retrieved from https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/
Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983), "Estimation of the coancestry coefficient: basis for a short-term genetic distance", *Genetics*, **105** (3), p.767 – 779.
Rishishwar, L., Conley, A. B., Vidakovic, B., and Jordan, I. K. (2015), "A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians", *Gene*, **574** (2), p.345 – 351.
Ritchie, S. G. (1990), "A knowledge-based decision support architecture for advanced traffic management", *Transportation Research Part A: General*, **24** (1), p.27 – 37.
Rius, M., and Darling, J. A. (2014), "How important is intraspecific genetic admixture to the success of colonising populations?" *Trends in Ecology and Evolution*, **29** (4), p.233 – 242.
Robinson, O. J., Ruiz-Gutierrez, V., and Fink, D. (2018), "Correcting for bias in distribution modelling for rare species using citizen science data", *Diversity and Distributions*, **24** (4), p.460 – 472.
Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003), "Informativeness of genetic markers for inference of ancestry", *The American Journal of Human Genetics*, **73** (6), p.1402 – 1422.
Royal, C. D., Novembre, J., Fullerton, S. M., Goldstein, D. B., Long, J. C., Bamshad, M. J., and Clark, A. G. (2010), "Inferring genetic ancestry: opportunities, challenges, and implications", *The American Journal of Human Genetics*, **86** (5), p.661 – 673.
Sergeant, E. S. G. (2018), "Epitools Epidemiological Calculator", Ausvet, Retrieved from http://epitools.ausvet.com.au.
Shannon, C. E. (1948), "A mathematical theory of communication", *Bell System Technical Journal*, **27** (3), p.379 – 423.
Shen, Q., Keppens, J., Aitken, C., Schafer, B., and Lee, M. (2006), "A scenario-driven decision support system for serious crime investigation", *Law, Probability and Risk*, **5** (2), p.87 – 117.
Singh, S., and Gupta, P. (2014), "Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey", *International Journal of Advanced Information Science and Technology (IJAIST)*, **27** (27), p.97 – 103.
Slatkin, M. (2016), "Statistical methods for analyzing ancient DNA from hominins", *Current Opinion in Genetics and Development*, **41**, p.72 – 76.
Song, Y. Y., and Ying, L. (2015), "Decision tree methods: applications for classification and prediction", *Shanghai Archives of Psychiatry*, **27** (2), p.130.
Szabolcsi, Z., Farkas, Z., Borbély, A., Bárány, G., Varga, D., Heinrich, A., Volgyi, A., and Pamjav, H. (2015), "Statistical and population genetics issues of two Hungarian datasets from the aspect of DNA evidence interpretation", *Forensic Science International: Genetics*, **19**, p.18 – 21.
Tal, O., and Tran, T. D. (2018), "New perspectives on multilocus ancestry informativeness", *Mathematical Biosciences*, **306**, p.60 – 81.

Tehan, D. (2017), "World War I Soldier Honoured", Retrieved from https://www.minister.defence.gov.au/minister/dan-tehan/media-releases/world-war-i-soldier-honoured

Themudo, G. E., Mogensen, H. S., Børsting, C., and Morling, N. (2016), "Frequencies of HID-ion ampliseq ancestry panel markers among Greenlanders", *Forensic Science International: Genetics*, **24**, p.60 – 64.

Theofilatos, A., Yannis, G., Kopelias, P., and Papadimitriou, F. (2016), "Predicting road accidents: a rare-events modeling approach", *Transportation Research Procedia*, **14**, p.3399 – 3405.

Tvedebrink, T., and Eriksen, P. S. (2019), "Inference of admixed ancestry with Ancestry Informative Markers", *Forensic Science International: Genetics*, **42**, p.147 – 153.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2017), "GenoGeographer–A tool for genogeographic inference", *Forensic Science International: Genetics Supplement Series*, **6**, p.463 – 465.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2018), "Weight of the evidence of genetic investigations of ancestry informative markers", *Theoretical Population Biology*, **120**, p.1 – 10.

Unrecovered War Casualties - Army (2012), "Unrecovered War Casualties - Army: What we do", Retrieved from http://www.army.gov.au/Our-work/Unrecovered-War-CasualtiesArmy

Unrecovered War Casualties - Army (n.d.), "World War Two: Papua and New Guinea", Retrieved from https://www.army.gov.au/our-work/unrecovered-war-casualties/world-war-two-papua-and-new-guinea

Uricchio, V. F., Giordano, R., and Lopez, N. (2004), "A fuzzy knowledge-based decision support system for groundwater pollution risk evaluation", *Journal of Environmental Management*, **73** (3), p.189 – 197.

Voskoboinik, L., Motro, U., and Darvasi, A. (2018), "Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes", *Forensic Science International: Genetics*, **35**, p.136 – 140.

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., and Kayser, M. (2011), "IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information", *Forensic Science International: Genetics*, **5** (3), p.170 – 180.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., and Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Science International: Genetics, 7(1), 98-115.

Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C., Sjerps, M., Lucena-Molina, J., and Zadora, G. (2015), "Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE)", *ENFSI guideline for evaluative reporting in forensic science*.

Workneh, A., Teferi, D., and Kumilachew, A. (2019), "Knowledge Based Decision Support System for Detecting and Diagnosis of Acute Abdomen Using Hybrid Approach", *Paper presented at the International Conference on Information and Communication Technology for Development for Africa*, Springer, p. 57 – 67.

Wright, J. L., Wasef, S., Heupink, T. H., Westaway, M. C., Rasmussen, S., Pardoe, C., Fourmile, G. G., Young, M., Johnson, T., Slade, J., and Kennedy, R. (2018), "Ancient nuclear genomes enable repatriation of Indigenous human remains", *Science Advances*, **4** (12), p.5064.

Yurdakul, M., Balci, A., and Ic, Y. T. (2020), "A knowledge-based material selection system for interactive pressure vessel design", *International Journal on Interactive Design and Manufacturing (IJIDeM)*, p.1 – 21.

Zouri, N., Zouri, M., Cumpăt, M. C., and Ferworn, A. (2019), “Knowledge discovery in support of health management decisional process: implementation oppurtunity”, *New Trends in Sustainable Business and Consumption*, **321**.

Appendix

A.1 Allele Frequencies

	rs1426654	rs9809818	rs28777	rs12913832	rs4683510	rs820371	rs4749305	rs6494411
Allele	**A**	**A**	**A**	**A**	**C**	**C**	**A**	**C**
AUS	1	0.9279	0.9252	0.1961	0.8925	0.7381	0.8411	0.1085
JPT	0.0048	0.0625	0.1635	1	0.0913	0.0385	0.0288	0.9183
Allele	**G**	**C**	**C**	**G**	**T**	**T**	**G**	**T**
AUS	0	0.0721	0.0748	0.8039	0.1075	0.2619	0.1589	0.8915
JPT	0.9952	0.9375	0.8365	0	0.9087	0.9615	0.9712	0.0817
	rs7997709	**rs1448485**	**rs730570**	**rs1876482**	**rs722869**	**rs1250233**	**rs9286879**	**rs2196051**
Allele	**C**	**G**	**A**	**A**	**C**	**A**	**A**	**A**
AUS	0.028	0.8878	0.8774	0.0787	0.8785	0.217	0.7684	0.6989
JPT	0.8077	0.1058	0.101	0.7981	0.1971	0.9663	0.0288	0
Allele	**T**	**T**	**G**	**G**	**G**	**G**	**G**	**G**
AUS	0.972	0.1122	0.1226	0.9213	0.1215	0.783	0.2316	0.3011
JPT	0.1923	0.8942	0.899	0.2019	0.8029	0.0337	0.9712	1
	rs6754311	**rs10455681**	**rs1366220**	**rs3811801**	**rs10496971**	**rs2758988**	**rs9319336**	**rs192655**
Allele	**C**	**A**	**A**	**A**	**G**	**A**	**C**	**A**
AUS	0.2731	0.8524	0.215	0	0.0566	0.1963	0.0514	0.8785
JPT	1	0.1346	0.9471	0.7019	0.7788	0.899	0.7163	0.2548
Allele	**T**	**G**	**G**	**G**	**T**	**T**	**T**	**G**
AUS	0.7269	0.1476	0.785	1	0.9434	0.8037	0.9486	0.1215
JPT	0	0.8654	0.0529	0.2981	0.2212	0.101	0.2837	0.7452
	rs984654	**rs11725412**	**rs4918664**	**rs4463276**	**rs683**	**rs4787040**	**rs4781011**	**rs1471939**
Allele	**C**	**A**	**A**	**A**	**A**	**A**	**G**	**C**
AUS	0.2067	0.081	0.8935	0.2576	0.7423	0.3287	0.7664	0.225
JPT	0.899	0.7163	0.2163	0.8942	0	0.9038	0.1538	0.7788
Allele	**T**	**G**	**G**	**G**	**C**	**T**	**T**	**T**
AUS	0.7933	0.919	0.1065	0.7424	0.2577	0.6713	0.2336	0.775
JPT	0.101	0.2837	0.7837	0.1058	1	0.0962	0.8462	0.2212
	rs1950993	**rs4984913**	**rs4833103**	**rs3907047**	**rs2357442**	**rs1393350**	**rs12203592**	**rs4959270**
Allele	**G**	**A**	**A**	**C**	**A**	**A**	**C**	**C**
AUS	0.6731	0.7689	0.4811	0.0602	0.9	0.3095	0.8056	0.5142
JPT	0.101	0.2019	0	0.5817	0.3173	0	1	0.6058
Allele	**T**	**G**	**C**	**T**	**C**	**G**	**T**	**C**
AUS	0.3269	0.2311	0.5189	0.9398	0.1	0.6905	0.1944	0.4858
JPT	0.899	0.7981	1	0.4183	0.6827	1	0	0.3942

A.2 Genotype Frequencies

	Genotype	AUS	JPT	Genotype	AUS	JPT	Genotype	AUS	JPT
rs1426654	AA	1	0	AG	0	0.0096	GG	0	0.9904
rs9809818	AA	0.8654	0	AC	0.125	0.125	CC	0.0096	0.875
rs28777	AA	0.8598	0.0288	AC	0.1308	0.2692	CC	0.0093	0.7019
rs12913832	AA	0.0588	1	AG	0.2745	0	GG	0.6667	0
rs4683510	CC	0.8037	0	CT	0.1776	0.1827	TT	0.0187	0.8173
rs820371	CC	0.5524	0	CT	0.3714	0.0769	TT	0.0762	0.9231
rs4749305	AA	0.7009	0	AG	0.2804	0.0577	GG	0.0187	0.9423
rs6494411	CC	0.0094	0.8365	CT	0.1981	0.1635	TT	0.7925	0
rs7997709	CC	0	0.6827	CT	0.0561	0.25	TT	0.9439	0.0673
rs1448485	GG	0.7857	0.0288	GT	0.2041	0.1538	TT	0.0102	0.8173
rs730570	AA	0.783	0.0192	AG	0.1887	0.1635	GG	0.0283	0.8173
rs1876482	AA	0	0.6635	AG	0.1574	0.2692	GG	0.8426	0.0673
rs722869	CC	0.757	0.0481	CG	0.243	0.2981	GG	0	0.6538
rs1250233	AA	0.0755	0.9423	AG	0.283	0.0481	GG	0.6415	0.0096
rs9286879	AA	0.6	0	AG	0.3368	0.0577	GG	0.0632	0.9423
rs2196051	AA	0.4946	0	AG	0.4086	0	GG	0.0968	1
rs6754311	CC	0.1111	1	CT	0.3241	0	TT	0.5648	0
rs10455681	AA	0.7143	0.0096	AG	0.2762	0.25	GG	0.0095	0.7404
rs1366220	AA	0.0187	0.9038	AG	0.3925	0.0865	GG	0.5888	0.0096
rs3811801	AA	0	0.5821	AG	0	0.3731	GG	1	0.0448
rs10496971	GG	0	0.5962	GT	0.1132	0.3654	TT	0.8868	0.0385
rs2758988	AA	0.028	0.8077	AT	0.3364	0.1827	TT	0.6355	0.0096
rs9319336	CC	0	0.5096	CT	0.1028	0.4135	TT	0.8972	0.0769
rs192655	AA	0.757	0.0769	AG	0.243	0.3558	GG	0	0.5673
rs984654	CC	0.0673	0.8077	CT	0.2788	0.1827	TT	0.6538	0.0096
rs11725412	AA	0	0.6418	AC	0.1619	0.2761	GG	0.8381	0.0821
rs4918664	AA	0.7963	0.0481	AC	0.1944	0.3365	GG	0.0093	0.6154
rs4463276	AA	0.0909	0.8077	AC	0.3333	0.1731	GG	0.5758	0.0192
rs683	AA	0.5773	0	AC	0.3299	0	CC	0.0928	1
rs4787040	AA	0.0741	0.8077	AT	0.5093	0.1923	TT	0.4167	0
rs4781011	GG	0.5607	0.0288	GT	0.4112	0.25	TT	0.028	0.7212
rs1471939	CC	0.08	0.6058	CT	0.29	0.3462	TT	0.63	0.0481
rs1950993	GG	0.4423	0	GT	0.4615	0.2019	TT	0.0962	0.7981
rs4984913	AA	0.6132	0.0385	AG	0.3113	0.3269	GG	0.0755	0.6346
rs4833103	AA	0.1792	0	AC	0.6038	0	CC	0.217	1
rs3907047	CC	0	0.3462	CT	0.1204	0.4712	TT	0.8796	0.1827
rs2357442	AA	0.8	0.1058	AC	0.2	0.4231	CC	0	0.4712
rs1393350	AA	0.0571	0	AG	0.5048	0	GG	0.4381	1
rs12203592	CC	0.6759	1	CT	0.2593	0	TT	0.0648	0
rs4959270	CC	0.2642	0.1442	AC	0.5	0.5	CC	0.2358	0.3558

A.3 Files for SimAdmixtR

1) *Allele frequencies data file* – a spreadsheet with the allele frequencies of each SNP for each of the two populations (example of 3 SNPs given in Table A.1);

Table A.1: Allele Frequency Data File
Example of allele frequency data required for the "Admixture Tool", detailing the SNP's ID, alleles present, and the allele frequency in both populations for the allele that occurs first based on lexicographic order.

SNP ID	*Alleles*	*AUS*	*JPT*
rs1426654	A/G	1	0.0048
rs9809818	A/C	0.9279	0.0625
rs28777	A/C	0.9252	0.1635

2) *Simulation details* – the nominated base-generation (great-grandparents) to describe the intended admixture to be simulated and the number of simulations required, that is, the number of individuals to be independently simulated based on the specified base-generation (Figure A.1);

Base-Generation	Number of Simulations
11111112	10

Figure A.1: Nominated Simulation Details File
An example of the admixed pedigree to be simulated by describing the base-generation where 1 = Australian and 2 = Japanese. The string 11111112 is a pedigree of 7 Australian great-grandparents and one Japanese great-grandparent, giving the soldier 1/8th Japanese ancestry. Ten individuals have been requested to be simulated independently with using the nominated base-generation string. For the scenarios used in this thesis (Table 4.3) the number of simulations was set to 10000.

3) *Example STRUCTURE Input* – In its current state, the Admixture Simulation tool outputs the simulated individuals in the format of a file that can readily inputted into STRUCTURE; as such, the tool requires an example STRUCTURE file (Figure A.2).

SAMPLE	POP	rs1426654		rs9809818		rs28777		rs12913832	
HG00096	1	1	1	1	1	1	1	4	4
HG00097	1	1	1	1	1	1	1	4	4
HG00099	1	1	1	1	1	1	1	4	4
NA18943	2	4	4	2	2	2	2	1	1
NA18944	2	4	4	2	2	2	2	1	1
NA18945	2	4	4	2	2	2	2	1	1

Figure A.2: Formatting File Example
Each row represents a different individual with their respective Sample ID number described in the SAMPLE column. The POP column details which population the sample belongs to, and the successive columns represents the given SNPs with each SNP consisting of two sub-columns to detail the two alleles for that given individual. The numbers represent the four possible bases (or alleles): where 1 = A, 2 = C, 3 = T, and 4 = G.

If the user does not wish for the outputted file to be in the format of a STRUCTURE file, additional steps must be taken by the user to transform the file into the format they desire. For the testing

performed in this chapter, input files were needed detailed an individual's genotype. Therefore, the following steps were taken to transform the output file (identical to the input file shown in Figure A.2) to the desired format:

1) The "*POP*" column is removed;
2) Alleles are converted from numbers back to their respective letters using the following key: 1 = A, 2 = C, 3 = T, 4 = G.
3) The two alleles at a given SNP, shown in Figure 4.6 to be spread across two cells, are combined into a single cell to create a genotype.
4) Genotypes are converted into lexicographic order, ensuring that only the following genotypes are present: AA, CC, GG, TT, AC, AG, AT, CG, CT, GT.

With the files uploaded, the simulation tool then creates the pedigree as follows:

1) A SNP profile is created for each member of the base-generation by randomly assigning two alleles independently for each SNP based on the allele frequencies for that individual's nominated ancestry

2) After the base-generation has been simulated, the program simulates the offspring in the next generation using the laws of Mendelian inheritance (Bateson and Mendel, 1913). One allele at each SNP is selected at random with equally likely probabilities, and passed down to the offspring, which is then combined with the inherited allele from the offspring's other parent to create the genotype of the offspring.

3) Step 2 occurs repeatedly until reaching the desired individual at the end of the pedigree, whose simulated genetic profile is then outputted by the Admixture Simulation tool.

A.4 Parsimonious Logistic Model Tree for the GPSP

Classification System >> SNPs not used in models: rs6754311, rs4787040, rs2357442, rs1393350, rs12203592, rs4959270,
Total SNPs in panel = 40, Total SNPs used in models = 34.

Model 1

Class AUS :
0.98 +
[rs1426654GG] * -1.98

Class JPT :
-0.98 +
[rs1426654GG] * 1.98

Model 2

Class AUS :
1.18 +
[rs12913832AA] * -1.89 +
[rs2196051GG] * -1.3 +
[rs3811801GG] * 1.21

Class JPT :
-1.18 +
[rs12913832AA] * 1.89 +
[rs2196051GG] * 1.3 +
[rs3811801GG] * -1.21

Model 3

Class AUS :
2.06 +
[rs28777AA] * 1.12 +
[rs820371TT] * -0.94 +
[rs4749305GG] * -1.85 +
[rs6494411CC] * -1 +
[rs9286879GG] * -1.32 +
[rs10496971TT] * 0.93 +
[rs683CC] * -1.22

Class JPT :
-2.06 +
[rs28777AA] * -1.12 +
[rs820371TT] * 0.94 +
[rs4749305GG] * 1.85 +
[rs6494411CC] * 1 +
[rs9286879GG] * 1.32 +
[rs10496971TT] * -0.93 +
[rs683CC] * 1.22

Model 4

Class AUS :
2.23 +
[rs9809818AA] * 0.89 +
[rs9809818CC] * -0.98 +
[rs4683510CC] * 0.91 +
[rs7997709TT] * 0.95 +
[rs1448485TT] * -1.27 +
[rs730570AA] * 1.1 +
[rs722869GG] * -0.89 +

[rs1250233AA] * -0.85 +
[rs1366220AA] * -1.79 +
[rs2758988AA] * -0.89 +
[rs984654CC] * -0.92 +
[rs4463276AA] * -1.18 +
[rs4833103CC] * -0.87 +
[rs3907047TT] * 0.85

Class JPT :
-2.23 +
[rs9809818AA] * -0.89 +
[rs9809818CC] * 0.98 +
[rs4683510CC] * -0.91 +
[rs7997709TT] * -0.95 +
[rs1448485TT] * 1.27 +
[rs730570AA] * -1.1 +
[rs722869GG] * 0.89 +
[rs1250233AA] * 0.85 +
[rs1366220AA] * 1.79 +
[rs2758988AA] * 0.89 +
[rs984654CC] * 0.92 +
[rs4463276AA] * 1.18 +
[rs4833103CC] * 0.87 +
[rs3907047TT] * -0.85

Model 5

Class AUS :
-0.39 +
[rs1876482GG] * 0.94 +
[rs10455681AA] * 0.72 +
[rs10455681GG] * -1.11 +
[rs9319336TT] * 1.62 +
[rs192655AA] * 0.91 +
[rs11725412AA] * -0.79 +
[rs11725412GG] * 0.83 +
[rs4918664GG] * -0.95 +
[rs4781011TT] * -1.22 +
[rs1471939TT] * 0.79 +
[rs1950993TT] * -1.24 +
[rs4984913GG] * -1

Class JPT :
0.39 +
[rs1876482GG] * -0.94 +
[rs10455681AA] * -0.72 +
[rs10455681GG] * 1.11 +
[rs9319336TT] * -1.62 +
[rs192655AA] * -0.91 +
[rs11725412AA] * 0.79 +
[rs11725412GG] * -0.83 +
[rs4918664GG] * 0.95 +
[rs4781011TT] * 1.22 +
[rs1471939TT] * -0.79 +
[rs1950993TT] * 1.24 +
[rs4984913GG] * 1

A.5 Validation of SimAdmixtR

Expected genotype frequencies were estimated using Mendelian inheritance formulae and observed genotype frequencies were calculated by simply recording the genotype's proportion for a given scenario's output from the Admixture Simulation tool. The accuracy of the Admixture Simulation tool was determined by calculating the absolute difference of the expected frequency and the observed frequency for each genotype. The following thresholds were established to assess the Admixture Simulation tool's accuracy: (*i*) an absolute difference of ≥ 0.05 was highlighted as a possible error, and (*ii*) $0.01 \leq 0.05$ was highlighted as a noticeable difference but may simply have occurred by chance. Summary statistics for the ten scenarios are shown in Table A.2 where a possible error is highlighted in red, bold, italics, and a noticeable difference in yellow, bold.

Table A.2 is interpreted as follows:

- The observed genotype frequencies were obtained from the ten admixture scenarios, where each scenario contained n = 10,000 observations and simulated once;
- The absolute difference was calculated for the expected and observed genotype frequencies for each of the ten scenarios at each SNP;
- The minimum, mean, standard deviation, and maximum of the absolute difference was then estimated for each SNP across the ten scenarios.
    - For example, SNP rs1426654's genotype 1 was observed on average to have an absolute difference of 0.0013 between the expected and observed genotype frequencies across the ten scenarios.

Table A.2: Admixture Simulation Tool Verification

Summary statistics obtained from the absolute difference of expected genotype frequencies and observed genotype frequencies across ten scenarios. Genotypes 1, 2 and 3 represent the three possible genotypes at each SNP and have been ordered lexicographically at each given SNP, for example rs1426654 has alleles A and G, therefore, Genotype 1 = AA, Genotype 2 = AG, and Genotype 3 = GG.

SNP	Genotype 1				Genotype 2				Genotype 3			
	Minimum	Mean	Standard Deviation	Maximum	Minimum	Mean	Standard Deviation	Maximum	Minimum	Mean	Standard Deviation	Maximum
rs1426654	0	0.0013	0.001417	0.004	0	0.00362	0.002324	0.008	0	0.00252	0.002411	0.007
rs9809818	0	0.00316	0.003259	**0.011**	0	0.00306	0.002338	0.008	0.0006	0.00192	0.001056	0.004
rs28777	0	0.00226	0.001632	0.006	0.001	0.00311	0.002258	0.009	0	0.00215	0.001624	0.006
rs12913832	0	0.0029	0.002108	0.008	0	0.00365	0.003163	**0.011**	0	0.00195	0.002509	0.008
rs4683510	0	0.00229	0.002477	0.008	0.001	0.00432	0.002794	0.009	0	0.00203	0.00215	0.008
rs820371	0	0.00212	0.001954	0.007	0.0009	0.00412	0.002844	**0.01**	0.0004	0.00368	0.002481	0.009
rs4749305	0	0.00135	0.001468	0.005	0.0008	0.00484	0.003095	**0.011**	0	0.00372	0.002954	0.009
rs6494411	0	0.00205	0.001379	0.005	0	0.00312	0.002642	0.008	0	0.00252	0.002465	0.009
rs7997709	0.0001	0.00325	0.002726	0.0094	0.0005	0.00562	0.00294	0.009	0	0.00373	0.003071	0.009
rs1448485	0.0004	0.00451	0.003206	**0.012**	0	0.0047	0.003232	**0.011**	0.0006	0.00349	0.002383	0.009
rs730570	0	0.00309	0.002365	0.0074	0	0.00362	0.00227	0.007	0.001	0.00319	0.00213	0.007
rs1876482	0	0.00238	0.001659	0.006	0.0005	0.00348	0.003131	**0.01**	0	0.00357	0.002619	**0.01**
rs722869	0	0.00328	0.002668	0.007	0	0.00297	0.002581	0.009	0	0.00231	0.002788	0.009
rs1250233	0	0.00368	0.002206	0.008	0.001	0.00482	0.002663	**0.011**	0	0.00213	0.002168	0.008
rs9286879	0	0.00193	0.001567	0.005	0	0.00361	0.00262	0.008	0.001	0.00327	0.002008	0.008
rs2196051	0	0.00177	0.002295	0.0077	0	0.00393	0.002005	0.007	0	0.00254	0.002033	0.007
rs6754311	0	0.00359	0.003232	**0.012**	0	0.00297	0.002282	0.007	0	0.00166	0.002256	0.006
rs10455681	0	0.00212	0.001692	0.0061	0	0.00243	0.00232	0.0067	0.0005	0.00252	0.001515	0.006
rs1366220	0.001	0.00441	0.002253	0.009	0.001	0.00627	0.00358	**0.013**	0.001	0.00384	0.003276	**0.012**
rs3811801	0	0.00099	0.001027	0.003	0	0.00371	0.002252	0.007	0	0.00339	0.002151	0.008
rs10496971	0	0.00244	0.002324	0.008	0	0.00196	0.001944	0.007	0.0013	0.00341	0.001685	0.008
rs2758988	0.0001	0.003	0.002249	0.009	0.001	0.00485	0.002846	**0.011**	0.001	0.00425	0.003694	**0.014**
rs9319336	0.0007	0.00177	0.000961	0.004	0.001	0.00367	0.002178	0.008	0	0.00219	0.001903	0.007
rs192655	0	0.0026	0.00325	**0.01**	0.0006	0.00286	0.001967	0.008	0	0.00151	0.001203	0.004
rs984654	0	0.00331	0.001941	0.007	0.0015	0.00405	0.00163	0.008	0	0.00266	0.002322	0.009
rs11725412	0	0.00321	0.002963	**0.011**	0.001	0.00416	0.002388	0.009	0.0005	0.00285	0.001383	0.006

SNP	Genotype 1				Genotype 2				Genotype 3			
	Minimum	Mean	Standard Deviation	Maximum	Minimum	Mean	Standard Deviation	Maximum	Minimum	Mean	Standard Deviation	Maximum
rs4463276	0	0.00181	0.001759	0.005	0.0009	0.00391	0.002304	0.008	0.0007	0.00294	0.002165	0.007
rs683	0	0.00201	0.002347	0.007	0	0.00338	0.002666	**0.01**	0	0.00323	0.002769	**0.01**
rs4787040	0	0.00281	0.001736	0.006	0.001	0.00228	0.001068	0.005	0	0.00145	0.000912	0.003
rs4781011	0.001	0.00281	0.001759	0.008	0	0.00409	0.002463	0.008	0	0.00247	0.001574	0.0064
rs1471939	0	0.00227	0.002638	**0.01**	0.0006	0.0043	0.002911	**0.011**	0	0.00317	0.001925	0.006
rs1950993	0	0.00185	0.001367	0.005	0	0.00409	0.002913	**0.01**	0.002	0.00414	0.001634	0.007
rs4984913	0	0.00258	0.002313	0.008	0.0007	0.00317	0.001624	0.006	0	0.0026	0.00142	0.005
rs4833103	0	0.00185	0.002559	0.009	0	0.00286	0.001667	0.0056	0	0.00311	0.003342	**0.012**
rs3907047	0	0.00116	0.001644	0.006	0.001	0.0028	0.001725	0.006	0	0.00247	0.001908	0.006
rs2357442	0	0.00358	0.002278	0.007	0.001	0.00333	0.00269	**0.01**	0	0.00259	0.00177	0.005
rs1393350	0	0.00053	0.001231	0.0043	0	0.00311	0.002599	0.007	0	0.00374	0.002421	0.007
rs12203592	0	0.00272	0.001865	0.007	0	0.00229	0.001747	0.005	0	0.00077	0.000795	0.002
rs4959270	0.0051	***0.07261***	**0.043788**	***0.149***	0.001	0.00657	0.00496	**0.0149**	0.0033	***0.07852***	**0.048586**	***0.1639***

A.6 Geometric Means for Degraded Australian Sample

Table A.3: Second WWII era Australian Sample's Geometric Mean

Collected in collaboration with (Ghaiyed, 2020). The observed geometric mean for each individual in the second WWII Australian sample along with their respective sample identification code (ID) and their available number of SNPs.

Sample No.	Sample ID	Geometric Mean	No. SNPs	Sample No.	Sample ID	Geometric Mean	No. SNPs
1	8072	0.996255	24	39	A64	0.956983	23
2	8075	0.96458	33	40	A83	0.940189	20
3	8080	0.955861	20	41	GN0976	0.936424	20
4	8134	0.966901	28	42	GN13	0.964039	26
5	8152	0.968288	30	43	GN133	0.952388	20
6	8177	0.996965	23	44	GN92	0.956809	20
7	8198	0.995044	24	45	GRF217	0.967367	29
8	8201	0.96525	26	46	GRF243	0.998481	14
9	8204	0.957042	23	47	GRF29	0.956808	16
10	8247	0.956729	23	48	GRF332	0.953859	17
11	8281	0.972345	24	49	GRF37	0.956896	18
12	8286	0.996417	23	50	GRF370	0.955171	17
13	8331	0.9636	28	51	GRF508	0.968143	29
14	8338	0.996746	25	52	GRF623	0.912997	21
15	8373	0.971634	27	53	GRF689	0.869156	20
16	8377	0.935922	15	54	GRF715	0.971035	35
17	8388	0.967347	25	55	GRF732	0.954677	21
18	8393	0.971792	29	56	GRF816	0.955664	29
19	8462	0.955301	16	57	MU1324	0.996919	26
20	8494	0.957084	16	58	MU1457	0.995834	21
21	8501	0.935445	15	59	MU8007	0.945657	28
22	8534	0.953858	23	60	MU8008	0.956803	17
23	8570	0.955594	16	61	MU8016	0.957116	16
24	8571	0.87689	14	62	MU8028	0.972076	34
25	8600	0.972074	32	63	MU8029	0.956464	19
26	8606	0.994569	29	64	MU8039	0.954541	19
27	8607	0.936405	18	65	MU8049	0.952846	17
28	8611	0.956827	21	66	MU8057	0.936343	12
29	8632	0.905964	34	67	MU8086	0.95688	18
30	8726	0.956399	18	68	MU8091	0.813843	26
31	8762	0.971397	33	69	MU8145	0.956803	17
32	8794	0.934713	14	70	MU8169	0.955152	39
33	8799	0.913816	34	71	MU968	0.971608	32
34	8810	0.973962	33	72	N24	0.996521	23
35	8818	0.957007	16	73	OTH23	0.956786	18
36	8825	0.991714	23	74	OTH47	0.954857	21
37	8826	0.951224	16	75	OTH48	0.953833	20
38	8923	0.956406	16				

A.7 Natural Log of the Likelihood Ratio for Degraded Australian Sample

Table A.4: Second WWII era Australian Sample's Likelihood Ratio

Collected in collaboration with (Ghaiyed, 2020). The observed natural logs of the likelihood ratio for each individual in the second WWII Australian sample along with their respective sample identification code (ID) and their available number of SNPs.

Sample No.	Sample ID	LR	LN(LR)	No. SNPs	Sample No.	Sample ID	LR	LN(LR)	No. SNPs
1	8072	3.28E+32	74.87	24	40	A64	1.03E+22	50.69	23
2	8075	1.2E+31	71.57	33	41	A83	9.89E+18	43.74	20
3	8080	2.19E+17	39.93	20	42	GN0976	4.85E+21	49.93	20
4	8134	9.97E+29	69.07	28	43	GN13	1.14E+28	64.60	26
5	8152	1.33E+38	87.78	30	44	GN133	5.74E+16	38.59	20
6	8177	2.65E+36	83.87	23	45	GN92	3.36E+24	56.47	20
8	8198	7.42E+25	59.57	24	46	GRF217	7.63E+35	82.62	29
9	8201	4.16E+26	61.29	26	47	GRF243	9.83E+21	50.64	14
10	8204	2.15E+22	51.42	23	48	GRF29	2.45E+13	30.83	16
11	8247	7.96E+26	61.94	23	49	GRF332	9.04E+13	32.13	17
12	8281	4.51E+32	75.19	24	50	GRF37	2.72E+18	42.45	18
13	8286	9.09E+35	82.80	23	51	GRF370	1.16E+13	30.08	17
14	8331	1.81E+32	74.28	28	52	GRF508	4.17E+33	77.41	29
15	8338	2.72E+33	76.98	25	53	GRF623	1.6E+15	35.01	21
16	8373	1.66E+29	67.28	27	54	GRF689	6.61E+11	27.22	20
17	8377	1.45E+17	39.51	15	55	GRF715	6.11E+42	98.52	35
18	8388	1.05E+33	76.03	25	56	GRF732	2.07E+20	46.78	21
19	8393	2.9E+30	70.14	29	57	GRF816	6.43E+30	70.94	29
20	8462	2.34E+14	33.09	16	58	MU1324	2.98E+35	81.68	26
21	8494	4.85E+16	38.42	16	59	MU1457	1.75E+25	58.12	21
22	8501	1.22E+14	32.44	15	60	MU8007	2.43E+21	49.24	28
23	8534	2.6E+23	53.91	23	61	MU8008	6.74E+17	41.05	17
24	8570	3.39E+12	28.85	16	62	MU8016	1.13E+18	41.57	16
25	8571	5.54E+13	31.65	14	63	MU8028	4.06E+41	95.81	34
26	8600	1.96E+42	97.38	32	64	MU8029	4.63E+21	49.89	19
27	8606	1.35E+25	57.87	29	65	MU8039	2.74E+14	33.24	19
28	8607	5.86E+20	47.82	18	66	MU8049	2.39E+12	28.50	17
29	8611	1.11E+24	55.37	21	67	MU8057	3.91E+11	26.69	12
30	8632	2.31E+44	102.15	34	68	MU8086	9.29E+16	39.07	18
31	8726	6.08E+13	31.74	18	69	MU8091	1.93E+28	65.13	26
32	8762	5.47E+39	91.50	33	70	MU8145	3.95E+18	42.82	17
33	8794	3.05E+14	33.35	14	71	MU8169	2.51E+20	46.97	39
34	8799	5.61E+26	61.59	34	72	MU968	9.23E+31	73.60	32
35	8810	5.38E+35	82.27	33	73	N24	3.73E+28	65.79	23
36	8818	1.4E+16	37.18	16	74	OTH23	8.32E+21	50.47	18
37	8825	2.25E+26	60.68	23	75	OTH47	9.29E+19	45.98	21
38	8826	4.65E+10	24.56	16	76	OTH48	1.55E+17	39.58	20
39	8923	7.9E+14	34.30	16					

A.8 Australian Q Value for Degraded Australian Sample

Table A.4: Second WWII era Australian Sample's Q Values

Collected in collaboration with (Ghaiyed, 2020). The observed Australian Q values for each individual, obtained from STRUCTURE, in the second WWII Australian sample along with their respective sample identification code (ID) and their available number of SNPs.

Sample No.	Sample ID	Australian Q Value	No. SNPs	Sample No.	Sample ID	Australian Q Value	No. SNPs
1	8072	0.998	24	39	A64	0.991	23
2	8075	0.995	33	40	A83	0.985	20
3	8080	0.915	20	41	GN0976	0.998	20
4	8134	0.997	28	42	GN13	0.997	26
5	8152	0.999	30	43	GN133	0.991	20
6	8177	0.999	23	44	GN92	0.999	20
7	8198	0.994	24	45	GRF217	0.999	29
8	8201	0.997	26	46	GRF243	0.998	14
9	8204	0.998	23	47	GRF29	0.995	16
10	8247	0.999	23	48	GRF332	0.993	17
11	8281	0.998	24	49	GRF37	0.999	18
12	8286	0.999	23	50	GRF370	0.993	17
13	8331	0.998	28	51	GRF508	0.997	29
14	8338	0.999	25	52	GRF623	0.987	21
15	8373	0.991	27	53	GRF689	0.968	20
16	8377	0.998	15	54	GRF715	0.999	35
17	8388	0.999	25	55	GRF732	0.998	21
18	8393	0.998	29	56	GRF816	0.999	29
19	8462	0.997	16	57	MU1324	0.997	26
20	8494	0.998	16	58	MU1457	0.998	21
21	8501	0.998	15	59	MU8007	0.965	28
22	8534	0.996	23	60	MU8008	0.999	17
23	8570	0.964	16	61	MU8016	0.999	16
24	8571	0.998	14	62	MU8028	0.998	34
25	8600	0.999	32	63	MU8029	0.998	19
26	8606	0.98	29	64	MU8039	0.992	19
27	8607	0.999	18	65	MU8049	0.984	17
28	8611	0.999	21	66	MU8057	0.997	12
29	8632	0.999	34	67	MU8086	0.995	18
30	8726	0.994	18	68	MU8091	0.998	26
31	8762	0.998	33	69	MU8145	0.998	17
32	8794	0.995	14	70	MU8169	0.707	39
33	8799	0.972	34	71	MU968	0.998	32
34	8810	0.999	33	72	N24	0.998	23
35	8818	0.998	16	73	OTH23	0.999	18
36	8825	0.998	23	74	OTH47	0.998	21
37	8826	0.976	16	75	OTH48	0.994	20
38	8923	0.997	16				

A.9 Derivation of Delta Method

$$V\left(f\left(\frac{\hat{x}}{\hat{y}}\right)\right) \approx \left(\frac{\partial f\left(\frac{\hat{x}}{\hat{y}}\right)}{\partial(\hat{x})}\right)^2 \times V(\hat{x}) + \left(\frac{\partial f\left(\frac{\hat{x}}{\hat{y}}\right)}{\partial(\hat{y})}\right)^2 \times V(\hat{y})$$

$$V\left(f\left(\frac{\hat{x}}{\hat{y}}\right)\right) = \left(\frac{1}{\hat{y}}\right)^2 \times \frac{\hat{x}(1-\hat{x})}{n_1} + \left(-\frac{\hat{x}}{\hat{y}^2}\right)^2 \times \frac{\hat{y}(1-\hat{y})}{n_2}$$

$$V\left(f\left(\frac{\hat{x}}{\hat{y}}\right)\right) = \frac{\hat{x}(1-\hat{x})}{\hat{y}^2 \times n_1} + \frac{\hat{x}^2}{\hat{y}^4} \times \frac{\hat{y}(1-\hat{y})}{n_2}$$

A.10 Example DNA-MAP Files

The following links contain example files for the DNA-MAP inputs "*Training Data Containing Samples Collected from the Populations of Interest*", "*Unknown Sample for Ancestry Prediction*" and "*Sample Data of Known Individuals not Included in the Training Data*".

Dropbox: https://www.dropbox.com/sh/cwny7fz8zuld88y/AACMjNe2aPNm2AavjIY3j-qYa?dl=0

OneDrive: https://1drv.ms/u/s!AvNVx7NP1j7vawr_N3YbXDUlU0Y?e=1aH5Z7

A.11: DNA-MAP's Operation Manual

The following lists the planned operation manual for the Shiny application DNA-MAP which will be extending from this thesis.

1. Introduction
    a. A summary of the DNA-MAP application and the currently utilised *p*LMT algorithm it utilises.
2. Benefits/Limitations/Assumptions
    a. A list of benefits and limitations of DNA-MAP compared to other classifiers in the literature. Along with any assumptions that are assumed by the algorithm.
3. Inputs
    a. What inputs are required for DNA-MAP's operation, including any formatting required for files. This section will be carried over from the contents of Chapter 6.
4. Algorithm
    a. An outlined methodology of the *p*LMT algorithm, indicating where each input is utilised. This methodology will be an extension of the content provided in Chapters 4 and 6.
5. Outputs and Interpretations
    a. Descriptions of what outputs are available to the user, as well as suggestion interpretation guidelines based on previous research. This section will be carried over from the contents of Chapter 6.