Genomic Analysis of Human Population Structure

by

David Andrew Eccles

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Biomedical Science.

Victoria University of Wellington
2011

Abstract

Recent developments in technology and computation have encouraged a shift towards a whole-genome approach to genetic analysis. Two key contributors to this shift, the Human Genome Project and the HapMap project, sparked an interest in studying the genetic patterns found in particular groups of individuals. The Maori population of New Zealand is an ideal, yet untapped, model for such studies due to recent partial mixture of two distinct population groups, and a culture of good documentation of genealogical information. A previous study carried out by the author found observable genetic differences between Maori and European populations in markers of forensic significance, yet no particular genetic patterns were found that were uniquely Maori. This study extends the previous work by developing methods to determine to what scale these differences exist, as well as demonstrating that a knowledge of these differences and methods could be used to improve current practices for clinical diagnosis.

The current project began by taking a 'candidate gene' approach, studying two regions where there were known large genetic differences between Maori and European individuals: the region of Alcohol Dehydrogenase genes on Chromosome 4 (Chapter 2), and the Monoamine Oxidase A gene region on Chromosome X (Chapter 3). In both of these regions, large frequency differences were observed between Maori and non-Maori populations at both a single mutation level, and at a haplotype level.

Despite the differences that were observed, no particular combinations of mutations could be considered uniquely Maori or uniquely non-Maori, so studies were expanded to the entire genome. This epansion was made

possible due to the recent and continuing developments in genome-wide technology and advancements in computational speed and efficiency. Once it was possible to carry out a genome-wide study of genetic differences, the goal of research changed from determining whether or not Maori and European individuals were uniquely different at a genotype level, to how small a marker set could be produced while maintaining population-uniqueness at a genotype level.

A method that uses bootstrap sub-sampling and other internal validation techniques has been developed for the generation of such a *signature set* for a Maori tribe (Ngati Rakaipaaka), and the generated set has been validated in other similar populations (Chapter 4). As a consequence of producing this set, the degree of European admixture was estimated in the tribe (28.7%), with over 15% of individuals within Rakaipaaka found to have no discernible European genomic ancestry.

In a validation of the *signature set* generation method itself, the marker selection procedure was repeated for Type 1 Diabetes, a disease with high heritability. An analysis of case and control individuals using this signature set found that the generated set is able to perform better than a genome-wide reference set of mutations known to be associated with Type 1 Diabetes. This validation study, other potential uses, and a more detailed discussion of the signature set generation method are presented in Chapter 5.

Acknowledgements

Friends and Family

My PhD study began in 2005, and I was tasked with finding suitable distractions and motivations to help me get through to the end. My parents and friends have been supportive of my research right through, even in the tricky (and very long) "still writing" stage of my project.

I was introduced to Jessica Campbell at the Interface computer club sometime near the beginning of my research, and found in her an energetic nature that would keep me going through the tough times. I proposed to her about three years after the start of this PhD project (not entirely coincidental), having decided that my research was essentially complete and finishing off the remainder of the writing wouldn't take all that long. We were married at the end of January 2009 under the name Eccles (a name from both our families back 5 generations), and now have a son, Peter Peregrin Matthew Eccles. Family life has been a welcome distraction, and has probably allowed me to retain much of the sanity I had before PhD study began.

Work Colleagues

I would like to specially thank Collette Bromhead from the Medical Laboratory (now Aotea Pathology), for sending me down the track of PhD study in the first place – a chat with Collette convinced me that research at a PhD level would be essential for my future work and research prospects. Through working as her molecular biology laboratory cleaner during the final stages of her PhD research, I gained some understanding of the suffering and rewards associated with doctoral study.

Thanks also go out to staff at ESR, especially members of the Population and Environmental Health group, who have kept me grounded in reality with respect to the financial and clinical side of genomic research.

Supervisors

My supervisors, Geoff Chambers and Rod Lea, provided me with the opportunity to attend the 11thInternational Congress of Human Genetics towards the end of the first year of my PhD project. As a fresh, starry-eyed graduate researcher, I took in as much as possible, and was rewarded with an excellent grounding in current genetic research. The opportunity to present posters at the conference also made me aware of how other fellow researchers might perceive my work.

Rod Lea's frequent requests for me to carry out more bioinformatics work allowed me to try out my research in the "real world", and enabled me to debug and test my theories – an opportunity that seems to be fairly rare in PhD research.

Geoff Chambers' help with academic style and grace has improved my presentation skills considerably. I've progressed from a student who got tongue-tied and frozen half-way through a talk on ADH genetics, to a researcher who has managed to self-publish his first book.

The Grapevine

Shortly before I began my PhD project, I asked a few people if they would be interested to receive updates on my progress. While I've only sent off a few emails to these people (via a mailing list that I named "The PhD Grapevine"), it's been wonderful to have these people around to listen to my tangential thoughts: Collette Bromhead, Te Runanga A Rangitane O Wairau, Gael Price, Graham and Elaine Langton, Daniel Briggs, Donald Gordon, Melanie Gibson, Liz Richardson, Adele Whyte, Laura Feasey, Helena Woods, Michelle Hunt, Claire Swain, John Beal, Jessica Eccles, Kirsten McEwen, my parents Noel and Faye Hall, and my nana Joan Hall.

Te Iwi o Rakaipaaka

A significant component of my research has involved the analysis of genetic and ancestry data from Te Iwi o Rakaipaaka. I am very grateful that they have entrusted their data to me, and hope that the results and insights from my research can give them a bit more understanding of what this genetic stuff all means.

Free and Open Source Software

Free and Open Source Software (FOSS) has been used extensively in this project, both in creating this thesis, and in carrying out analysis of genetic data. A number of FOSS programs in particular have been the mainstay of my research tools:

Emacs Editing of text documents, program code

LaTeX Thesis, Presentations, Publications

Scribus Posters

Inkscape Illustrations

GIMP Picture/Photo editing

Perl Data conversion / aggregation

R Graphing, statistical analysis

Openoffice.org Writing 6-month reports, spreadsheet summaries of data

Plink Genome-wide data analysis

Haploview Haplotype block diagrams

Proofreading

I also thank my wonderful proofreaders, who have commented on my thesis during its creation, provided constructive criticism, and lent me an extra set of eyes: Geoff Chambers, Rod Lea, Jessica Eccles, Murray Darroch, and Faye Hall.

Funding

Finally, as always, I thank the financial contributors to my study. I hope that I have been able to give back a little to the community in exchange for the financial support of my study, and look forward to future community presentations of my research:

Environmental Science and Research Ltd. (ESR) Population and Environmental Health Scholarship, travel and conference costs

Victoria University of Wellington (VUW) Computer equipment, laboratory reagents, travel and conference costs

Studylink Student allowance

Wellington Medical Research Foundation (WMRF) Travel and conference costs

Contents

List of Tables

List of Figures

Terminology and Abbreviations

Terminology

New Zealand English is a language that contains no macrons, hence common New Zealand English words borrowed from Maori will also contain no macrons in this thesis – the native New Zealand population will be referred to as 'Maori' rather than 'Māori'.

Advice has been received via email from *Te Puni Kōkiri* (TPK), the New Zealand Ministry of Maori Development, regarding the use of the term 'Maori' in this thesis:

> In the past the Statistics New Zealand Census differentiated Māori descendents by the two terms 'New Zealand Māori' and 'Cook Island Māori. More recently the 'New Zealand tag has become redundant because it is commonly understood that the term 'Māori' refers to Māori descendents from New Zealand. A point of differentiation is maintained for Cook Island Māori in the Statistics New Zealand Census, and general use.
>
> If you also discuss Cook Island Māori within your thesis I agree that clarification may be necessary. However, if you only discuss Māori, the fact that they are from New Zealand is inherent.
>
> *Hollie Smith, Te Puni Kōkiri, Wellington*

In this thesis, the term 'Maori' on its own refers to Maori descendents from New Zealand, consistent with the advice of TPK. References to Cook Island Maori will be stated with the fully qualified name, 'Cook Island Maori'.

Abbreviations

ADH Alcohol Dehydrogenase

AGRF Australian Genome Research Facility

ALAC Alcohol Advisory Council of New Zealand

ALFRED The Allele Frequencies Database

ALDH Aldehyde Dehydrogenase

CNV Copy Number Variant

DNA Deoxyribose Nucleic Acid

EHH Extended Haplotype Homozygosity

ESR Institute of Environmental Science and Research

GABA gamma-aminobutyric acid

GWAS Genome-wide Association Study

HWE Hardy-Weinberg Equilibrium

HLA Human Leukocyte Antigen

HUGO Human Genome Organisation

ISEA Island South-East Asia

LD Linkage Disequilibrium

MAO Monoamine Oxidase

MAOA Monoamine Oxidase A

MAOB Monoamine Oxidase B

MHC Major Histocompatibility Complex

NBS National Blood Service

NCBI National Center for Biotechnology Information

SD Standard Deviation

SNP Single Nucleotide Polymorphism

RNA Ribose Nucleic Acid

RHAS Rakaipaaka Health and Ancestry Study

ROC Receiver Operating Characteristics

T1D Type 1 Diabetes

TPK Te Puni Kokiri

TIORI Te Iwi o Rakaipaaka

VNTR Variable Number Tandem Repeat

WTCCC Wellcome Trust Case Control Consortium

Chapter 1

Introduction

This PhD project began in 2005, after the completion of the Human Genome Project, and near the time that the HapMap project released their first set of public data. Due to the nature of techniques and technologies explored in this thesis, a number of different investigations have been carried out, each sharing a common theme: the genomic analysis of genetic variation in human populations.

This is not a thesis on population genetics, nor is it a thesis on health science, nor computational biology, nor molecular biology. Rather, it synthesises a number of different areas of research to make new discoveries that cannot be found by study of a single area alone. This *bioinformatics* approach creates more hypotheses than can be tested in the short course of a PhD project, and as such provides plenty of opportunities for further exploration by other researchers.

Chapters in this thesis primarily concentrate on analyses of genetic data. It is expected that individuals who are not specialists in genetic research will read this thesis (in particular, members of Te Iwi o Rakaipaaka), as components of the thesis apply to areas outside genetics. With this in mind, the introduction begins with material that introduces some concepts relevant to the study of this thesis. The concepts presented here are not an exhaustive

review of their respective subject areas, but should provide readers with a grounding to understand key ideas required for the interpretation of results in subsequent chapters.

1.1 An Introduction to Bioinformatics

> "Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful."
>
> *NIH Working Definition Of Bioinformatics And Computational Biology, 2000*

Bioinformatics research, with a study of human populations as a component, is a continuous process of research and refinement. Researchers are challenged to race waves of new biological information to the shore of public knowledge, finding new insights about that information along the way. The general-purpose biology researcher, or bioinformatician, will become increasingly important as a consequence of large-scale collaborative efforts between many researchers (and members of the public) with vastly different educational backgrounds.

A researcher needs to understand the current studies carried out in their research domain, so that they can communicate their ideas to other colleagues within their research community. The main challenge of a bioinformatics researcher is twofold: they must be able to locate relevant research from many different areas, and understand research well enough to translate from scientific terminology into a language understood by the general public.

This research project demonstrates the challenges of bioinformatics. New genetic data has flooded in at a rate far greater than any ability to study all of it, requiring a critical eye to observe the available data and evaluate which data warrants further analysis.

1.2 Genetic Concepts

1.2.1 DNA

Deoxyribose nucleic acid (DNA) is a sequence of many single and double-ringed compounds (nitrogenous bases) that are bonded to a repeating sugar-phosphate backbone.[†] Each base in the sequence can be one of four chemicals: adenine, cytosine, guanine or thymine, usually abbreviated as the first letter of the base name (i.e. A, C, G, and T respectively). The unit that is a combination of the base, sugar, and phosphate is called a *nucleotide*.

DNA is typically double-stranded, with each strand being the complementary opposite of the other, following a base-pairing rule (see figure 1.1) where adenine and thymine are paired together, as are cytosine and guanine. This allows for exact copies to be made via splitting of strands and attachment of complementary bases in a process known as DNA replication (see Berdis, 2009). Lengths of double-stranded DNA are typically measured in *base pairs* (bp), with one base pair being a base that has been paired up with its complementary partner.

A small fraction of this DNA (about 1%) consists of *genes*. These are a sequence of bases that describe a particular sequence of amino acids. There are 20 amino acids that can be used in this sequence and only four bases in DNA, so a combination of three bases (e.g. ATG), or a *codon*, is sufficient to describe each amino acid within the sequence (see Bollenbach et al., 2007). The amino acids join together in the prescribed sequence by a structure called a ribosome, then folded and packed (see Cooper et al., 2010). This packed amino acid sequence is called a *protein*.[‡] Only a portion of the DNA inside a gene (*protein coding* DNA) is converted (or *translated*) into amino

[†]The sugar is ribose, with one fewer oxygen than a typical ribose sugar, hence *deoxyribose*.

[‡]Some proteins are composed of a few of these packed sequences. The components are then referred to as protein *subunits*

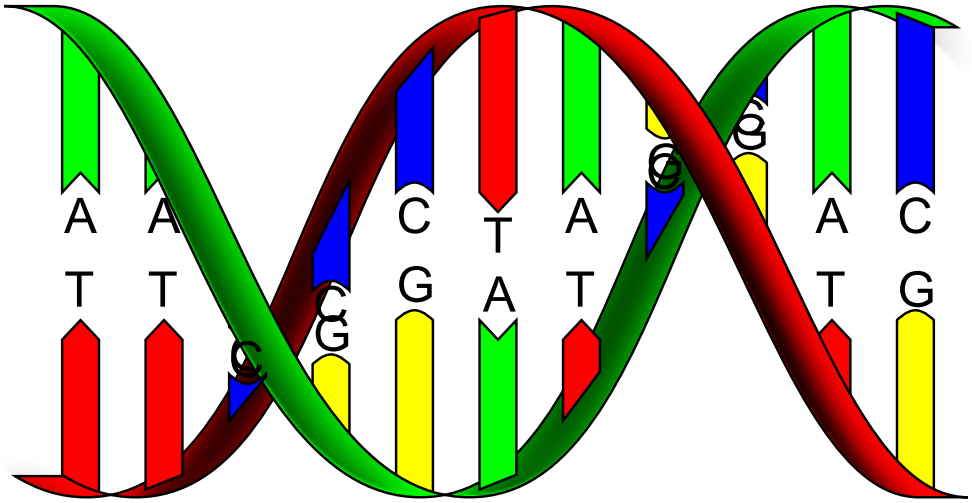

Figure 1.1: A depiction of DNA: a double-helical structure with four basic building blocks (**A**denine, **C**ytosine, **G**uanine and **T**hymine, shown as green, blue, yellow and red bands respectively), attached to a sugar-phosphate backbone (shown here as the red and green ribbons).

acids, via an intermediate molecule with a structure similar to DNA called Ribonucleic Acid (RNA). A contiguous sequence of DNA that is converted into an amino acid (or *expressed*) is known as an *exon*; the non-expressed sequences of protein coding DNA are called *introns*.

While genes and the proteins derived from genes have typically been the target for most DNA research in the past, it is now known that a high proportion of non-protein coding DNA also has functional significance; the DNA is transcribed into RNA which is involved in the regulation of cellular processes (see Mattick, 2007). For example, several loci associated with Crohn's disease were found in regions with no known genes or transcripts (Mathew, 2008). Therefore, it is useful to dispense with the classical view of DNA as something that generates proteins, and consider the possibility that *every* region of DNA (not just protein coding regions) may contribute to the diversity and complexity of our species.

1.2.2 Mutation

Variation can be introduced into DNA through *mutation*. A mutation is an alteration of DNA sequence or structure to something different from the original sequence. It can be induced by chemicals interacting with DNA, by electromagnetic radiation, and by random errors in the replication or repair process. Each DNA variant is known as an *allele*, with common variants being referred to as *major* alleles, and rare variants being referred to as *minor* alleles.

1.2.2.1 Single Nucleotide Polymorphism

A Single Nucleotide Polymorphism (SNP), is a DNA sequence variation that describes a change of a single base of DNA, from one base to another (see Figure 1.2). Most discovered SNPs have two alleles, in which case they are called *dimorphic* SNPs. Some definitions require SNPs to have a minimum rare variant frequency and/or reside within genes, but the use of the term in this thesis has none of these additional restrictions.

As some SNPs reside within genes, those SNPs can also be classified in terms of the change in the amino acid that the base (and its neighbours) codes for (see Russell, 1998, Chapter 19, pp. 619-621). The redundancy of the genetic code (i.e. the codons, having 64 different possible base combinations, only code for 20 amino acids) means that the presence of a SNP does not always affect the protein product of a gene. A *synonymous* mutation keeps the same amino acid, but recent research demonstrates that in some cases a change at the DNA level can still influence the folding and binding properties of the final protein product (Kimchi-Sarfaty et al., 2007). *Missense* (or *non-synonymous*) mutations are those that result in an amino acid substitution (i.e. a codon for one amino acid becomes a codon that codes for another amino acid). *Nonsense* mutations are those that result in the substitution of an amino acid for a signal to stop further addition of

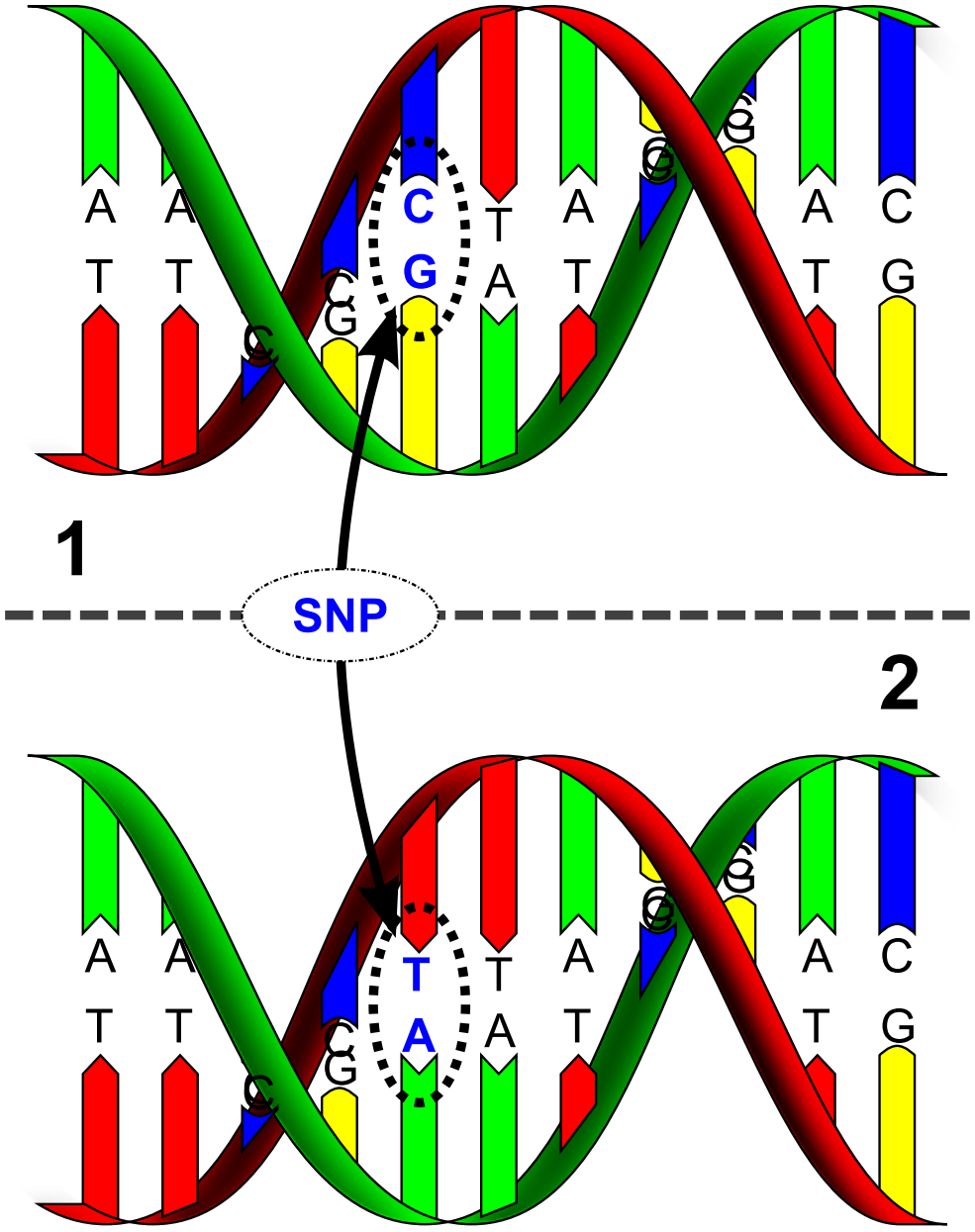

Figure 1.2: A single nucleotide polymorphism found within the Monoamine Oxidase A (MAO-A) gene. DNA strand 1 differs from DNA strand 2 at a single base-pair location (a C/T dimorphism).

amino acids (a *termination* codon), which prevents correct construction of the final protein product.

An Infinite Allele Model of neutral theory (see Kimura, 1991) predicts that nucleotide substitutions should occur at a constant rate based on time, rather than on the number of generations. Kumar and Subramanian (2002) carried out an investigation of nucleotide substitution mutation rate in mammalian genomes, following a suggestion that mutation did not seem to be generation-based as previously suspected. They compared 17,208 protein-coding DNA sequences from 326 mammals, and discovered that substitution mutations occur at a rate of 2.22×10^{-9} substitutions per locus (site) per year. Taking a human genome sequence size of 3.23609×10^{9} base pairs[†], this works out to just over 7 mutations per person per year. Alternatively, assuming a generation time of 20 years, this would result in 4.44×10^{-8} substitutions per locus per generation, or about 145 substitutions per generation across the entire genome.

1.2.2.2 Tandemly Repeated DNA

Tandemly Repeated DNA is a sequence of DNA that repeats many times with no break between repeats. The length of the repeated sequence can vary greatly, from a single base-pair to parts of genes, or even entire genes. A number of common subclasses of tandemly repeated DNA exist (ranges of repeat length vary depending on reference) :

CNV Copy number variant – repeated sequence $>$ 1 kb in length (see Freeman et al., 2006)

VNTR Variable number of tandem repeat – repeated sequence 15-100 bp (see Russell, 1998, Chapter 15, p. 487)

[†] `http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=3`

Minisatellite repeated sequence with 15-70 bp 'core' (see Shriver et al., 1993)

STR Short tandem repeat – repeated sequence 3-5 bp (see Shriver et al., 1993)

Microsatellites repeated sequence 1-6 bp (see Eckert and Hile, 2009)

1.2.2.3 Differences between SNPs and Tandem Repeats

The most common SNPs are dimorphic and have low mutation rates (approximately 4×10^{-8} per locus per generation, see Section 1.2.2.1). Tandem repeats have many different possible variant states and a much higher (but more variable) mutation rate, between 10^{-6} and 10^{-2} per locus per generation (see Eckert and Hile, 2009). These two factors mean that a single SNP is likely to be less informative than a single tandem repeat. However, the large number of SNPs in the human genome makes up for the lower information content in SNPs. As an estimate of how many SNPs exist in the human genome, the dbSNP database of NCBI (NCBI dbSNP Build 132) indicates that in October 2010, there were 30,442,771 recorded RefSNP clusters[†]. However, typing large numbers of SNPs can be extremely costly, meaning that it is useful to be able to select SNPs that have the highest information content possible.

1.2.3 Chromosomes and Inheritance

DNA is arranged in structures called *chromosomes*. In the nucleus of human cells, the chromosomes are linear and range in size from around 50 million base pairs (chromosome 21) to 250 million base pairs (chromosome 1). There are 23 pairs of chromosomes that have a similar structure (homologous pairs), making up 46 nuclear chromosomes. Outside the nucleus, other

[†]`http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi`

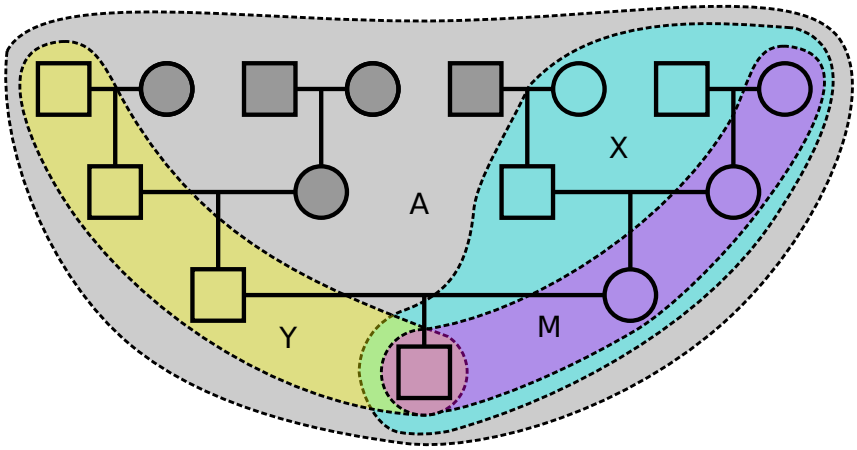

Figure 1.3: An indication of the contribution and mixture of each type of chromosome to an individual, transferred through the generations. Types shown are autosomal (A, shaded light grey), X-chromosomal (X, shaded cyan), Y-chromosomal (Y, shaded yellow) and mitochondrial (M, shaded magenta) DNA. The DNA from the individuals shaded in dark grey is only passed to their common descendant through autosomal DNA.

cell structures called mitochondria carry circular chromosomes of DNA, containing around 17 thousand base pairs (17 kb).

There are four different inheritance patterns found in Human chromosomes, dependant on the parental source from which the chromosomes are derived (see Figure 1.3). The sex chromosomes, X and Y, are transferred (or not) from parents to children depending on the sex of the child: males inherit an X chromosome from their mother, and the Y chromosome from their father; females inherit an X chromosome from both parents. The non-sex chromosomes (or *autosomes*) are inherited from both parents; one set of 22 autosomes is inherited from the mother, and another matching set is inherited from the father. Mitochondrial chromosomes are inherited from the mother, and therefore provide a female inheritance pattern that mirrors the male inheritance pattern of the Y chromosome. Genetic information from some ancestors of an individual will only be found in the *autosomal* DNA of that descendant individual (shaded in Figure 1.3).

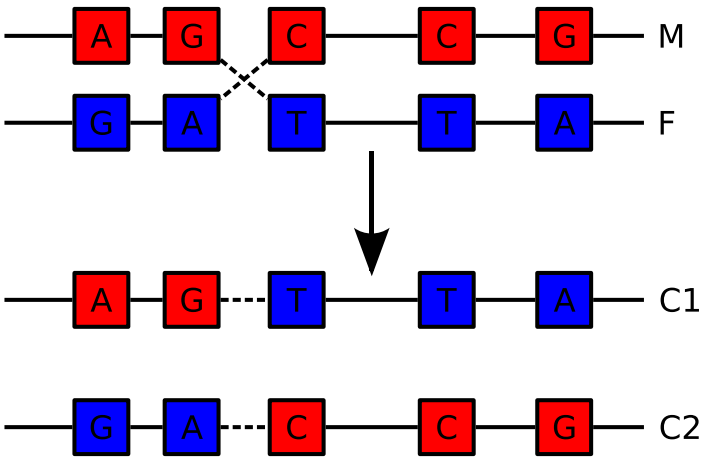

Figure 1.4: Recombination involves the breakage and rejoining of parental chromosomes. This process produces recombinant chromosomes that look similar to, but not the same as, both of the parental chromosomes. The letters in the above figure represent genetic variants (or *markers*) that differ in the two parental chromosomes; the remainder of genetic sequence is identical in both parental chromosomes.

1.2.4 Recombination

In the context of human populations, chromosomes represent a molecular genealogy, written in an extremely ancient genetic language. Most of the individual variation of DNA is introduced through a process known as *recombination*. During this process, which happens each generation, homologous pairs of chromosomes bind together and exchange large segments of DNA, with the exchanges happening via breakage and rejoining of the DNA backbone at a number of *recombination points* (see Figure 1.4). The original chromosomes are referred to as *parental* chromosomes, while the recombined chromosomes are referred to as *recombinant* chromosomes. The effect of this process is a shuffling of the DNA in a way that preserves a significant amount of the genetic structure from the two parents as contiguous blocks.[†]

The process of recombination breaks a sequence of DNA into two parts.

[†]This differs from mutation, where the mutation creates genetic sequence different to previous generations of DNA.

Recombination likelihood is typically specified in *centimorgans* (cM). This statistic indicates the probability that two genetic regions on the same chromosome will segregate and not be passed down together to the next generation. If two points are one centimorgan apart, then there is a 1% chance that a recombination event will cause those two points to segregate during the process of DNA recombination. The nature of recombination[†] means that even when a recombination event *always* happens between two points, the chance of segregation cannot exceed 50%. Hence any points that are greater than 50cM apart are considered *unlinked*, as the variant at one point cannot predict the variant at the other point – this is the same outcome as when the two genetic regions are on separate chromosomes.

Figure 1.5 demonstrates the effect of this recombination process by tracking the recombination of a single autosome through four successive generations (see Appendix F for details on the computer code used to generate this figure). The process as shown in Figure 1.5 begins with eight different ancestral chromosomes: black, yellow, magenta, blue, cyan, green, red, and white, with segments of different colours indicating a chromosomal region that has been inherited from a different ancestral line. Each pair indicates a set of two homologous chromosomes for a single individual. Between the first and second generation, 3-4 recombination points form in each chromosome, resulting in shuffling of DNA of the two parental chromosomes and a recombinant chromosome that appears different to – yet shares some structure with – its parental chromosomes. The process continues on similarly to the next generation, with recombinant chromosomes still sharing some portion of DNA from the original eight chromosomes. In the final generation, the random construction of recombination points has resulted in the DNA from the green and red chromosomes being lost, so the final recombinant chromosome only has DNA from six of the original eight chromosomes.

[†]Recombination only happens in two of four gametes at meiosis (see Russell, 1998, pg. 143)

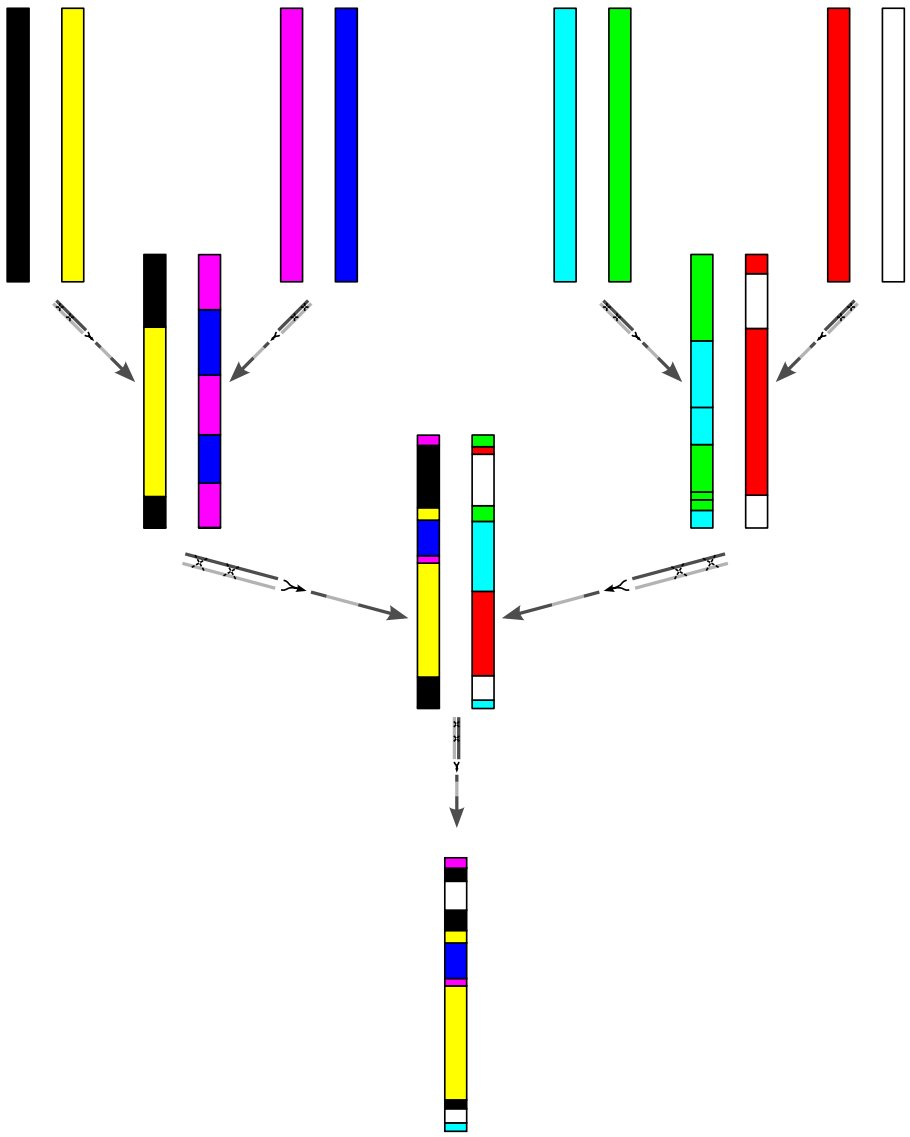

Figure 1.5: The genetic ancestry of a single chromosome is complex, the result of multiple recombination events that happen at each generation. In this figure of simulated recombination, black lines indicate recombination points (see Figure 1.4). The final chromosome shown in this figure contains a genetic history that is derived from six of the original eight ancestral chromosomes.

1.3 Population Genetics Concepts

A *haplotype* is the entire DNA sequence from a single chromosome, but the use of the term in this thesis is also used to describe particular subsequences of the full haplotype (e.g. the haplotype for the *MAOA* gene). A single human has 46 haplotypes from 46 nuclear chromosomes (see Section 1.2.3) which each describe the genetic ancestry of that individual. However, genetic studies do not usually concentrate on a single individual, but on groups of individuals (or populations). In a population context, genetic variants have frequencies within populations, and these frequencies can differ between different populations. In aggregation, genetic data can provide information on the way genetic structure changes over time, and demonstrate whether particular genetic sequences are preserved despite substantial variation throughout the genome.

1.3.1 Linkage Disequilibrium

When considering recombination at a population level, it is evident that recombination does not always occur at the same location within a particular chromosome. The chance of a recombination point forming within two regions of a chromosome increases with base-pair distance in a more or less predictable and linear fashion, with 1 centimorgan per megabase ($1cMMb^{-1}$) being a reasonable estimate. Mean recombination rates in the human population vary between about $0.3cMMb^{-1}$ and $1.9cMMb^{-1}$, although local regions of high recombination (*recombination hotspots*) are found all over the genome (International HapMap Consortium, 2007; Ke et al., 2004).

Genetic sequences from different regions are *linked* when they are usually transferred to the next generation together in a population. Two genetic variants (or *markers*) are said to be in *Linkage Disequilibrium* (LD) if recombination rarely occurs between those regions, i.e. recombination points are

unlikely to appear between the markers. The LD structure of the human genome appears to have functional significance (Hinds et al., 2005), so an understanding of LD patterns aids investigations into the effects of genetics on phenotype.

1.3.1.1 Calculating Marker Linkage

Calculations of linkage disequilibrium compare genotypes at two markers and represent the correlation between genotypes at those two markers (see Hedrick and Kumar, 2001; Du et al., 2007). Consider two markers with alleles m/M for marker 1, and n/N for marker 2. It is assumed that M/N and m/n correspond to the major and minor alleles of the two markers respectively. The correlation between these markers (D) can be determined by frequencies in a 2x2 table:

	n	N
m	f(mn)	f(mN)
M	f(Mn)	f(MN)

Where $f(mn)$, $f(mN)$, $f(Mn)$, and $f(MN)$ are the frequencies of the four possible genotype combinations of major and minor alleles at each marker.

The statistic, D, is calculated as the difference between linked alleles (major allele at both sites, or minor allele at both sites) and unlinked alleles (major allele at one site, minor allele at the other site):

$$D = f(mn) \times f(MN) - f(mN) \times f(Mn) \tag{1.1}$$

However, this statistic suffers from the issue that different combinations of allele frequencies will have different statistical distributions – the maximum value for D (0.25) is only possible when both alleles have a frequency

of 0.5. Two modifications to D are suggested by Hedrick and Kumar (2001) in an attempt to remove the dependency on allele frequency, namely D' (D-prime) and r^2 (R-squared):

$$D' = \frac{D}{min(f(m) * f(N), f(M) * f(n))} \tag{1.2}$$

$$r^2 = \frac{D^2}{f(m) * f(n) * f(M) * f(N)} \tag{1.3}$$

Where $f(M)$, $f(N)$, $f(m)$, and $f(n)$ are the frequencies of major and minor *alleles* of each marker, rather than genotype frequencies as used in the calculation of D. The D' statistic is an adjusted form of D such that the result for all allele frequency combinations lies within the range of -1 to 1, adjusted by scaling by the maximum possible value of D for the particular allele frequency combination. The r^2 statistic scales D such that it fits into the range of 0 to 1 and is equivalent to the square of the Pearson's correlation r for the 2x2 table of genotype frequencies (see Du et al., 2007), but only reaches its maximum value when the frequencies of each marker are the same (Hedrick and Kumar, 2001).

1.3.2 The Haplotype Block Theory

An analysis of linkage disequilibrium throughout the genome indicates that human chromosomes exhibit a "block-like" structure built from discrete segments of DNA that form characteristic patterns in populations (Hurles et al., 2002; Gabriel et al., 2002; Wall and Pritchard, 2003). The observation of large regions of low recombination and small regions of high recombination has led to the development of the haplotype block theory of DNA. A *haplotype block* is a sequence of DNA that has persisted in a population through successive generations with minimal (or no) recombination occurring within that region. The haplotype block arrangement in human

groups is influenced by factors such as geographic isolation, founder effect and admixture and therefore varies substantially among populations with different demographic histories (Walsh et al., 2003).

The haplotype block model has been reviewed with respect to the human genome, and while the human genome appears to only be moderately block-like, such a model is more consistent with the picture we have of the genome than a model of uniform recombination (Wall and Pritchard, 2003). It was observed that mutations that are within 20-30kb of each other are tightly linked, and linkage within haplotype blocks breaks down outside this distance. It was also noticed that if mutations were part of one block on one side, and another block on the other side, then it was highly likely that both sides were actually resides in the same haplotype block. An analysis by Ke et al. (2004) determined that the average haplotype block length in European populations is 11.1kb, but only 3.3kb in the African American samples typed in that study. Another earlier study of block sizes determined that the mean haplotype block size in European populations is 22kb, and 11kb in Yoruban and African-American populations (Gabriel et al., 2002).

1.3.3 Admixture

When a number of genetically diverse populations combine, the combined group will contain some genetic features from both of the ancestral populations, a situation known as *admixture*. A population is considered *admixed* if there is substantial genetic contribution (through recombination and re-assortment) from more than one ancestral population among the individuals in that population. This definition also extends to an individual level. A population-based description of admixture should be treated as a statistical average; it can not be used to infer the degree of admixture for particular individuals within that population.

Admixture can be calculated by observing the frequency of particular genetic variants in ancestral populations, then comparing those frequencies to the frequencies of those same variants in the population under test. This can then be expressed as a percentage, e.g. "The Eurvedio population has 58% Apalete admixture, with the remainder from the Daitar and Cralnus populations."[†]

1.4 Genomic and Bioinformatic Concepts

Studies related to population variation at a genetic level have been helped by international collaboration and public sharing of data.

1.4.1 Human Genome Project

The Human Genome Project (HGP) was a research project with a goal to generate a human genome reference sequence in about ten years. The project was launched in 1990, and the first draft genome sequence was completed in late 2000 (International Human Genome Sequencing Consortium, 2001). The sequence was essentially complete in April 2003, although some regions still remain (particularly long lengths of repeated DNA) that are very difficult to unambiguously determine the sequence for using current technology. Between July 2000 and October 2009, there have been 18 revisions of the reference sequence;[‡] data is publicly available on the US National Institute of Health (NIH) website.[§]

The study of genome variation at the molecular level (genomics) did not start with the human genome project; cytogeneticists have studied

[†]Population names are fictitious and do not represent any real-world populations.

[‡]`http://www.ncbi.nlm.nih.gov/mapview/stats/BuildList.cgi#Homosapiens`

[§]`http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606`

chromosome variation since the 1900s (see Pearson, 2006). However, the completion of the HGP represented a shift from single-lab research to large international collaborative efforts, and demonstrates the difference that collaboration can have for genomic research.

1.4.2 HapMap

The International HapMap Project[†] is an ongoing attempt to describe *common* genetic variation in Human populations. Whereas the Human Genome Project has produced a single reference sequence for human DNA, the HapMap Project investigates how DNA sequence changes in different populations. The initial phase of the project (Phase I), completed in 2005, was the genotyping of at least one common (occurring in at least 5% of the population) SNP every 5kb in 270 individuals from four different global populations (International HapMap Consortium, 2005):

- Yoruba in Ibadan, Nigeria (YRI)
- Japanese in Tokyo, Japan (JPT)
- Han Chinese in Beijing, China (CHB)
- CEPH – Utah residents with ancestry from northern and western Europe (CEU)

Phase II of the project, completed in 2007, continued this genotyping process in the four populations, with the aim of effectively capturing all common SNP variation either directly, or via linkage ($r^2 > 0.9$) with an already genotyped SNP (International HapMap Consortium, 2007). The HapMap Project has encouraged a substantial body of research that has contributed to the process of large-scale genotyping, as well as the use of genome-wide data for disease association studies.

[†] `http://hapmap.org`

1.4.2.1 SNPchips

The analysis of haplotype blocks in the four HapMap populations led to the discovery of sets of SNPs that were able to type a large proportion of the common haplotype block variants in the human population.

SNPchips are a new technology that attempt to capture as much of the genome-wide SNP variation as possible in a single assay, while only typing a proportion of that total variation. Commercially available SNPchips began with around a hundred thousand mutations being typed on a single chip, and have since increased in capacity to over a million mutations[†‡]. They are a tool that can be used to obtain hundreds of thousands of SNP genotypes out of one DNA sample in a single assay. They end up being incredibly cheap per SNP (about 0.1 cents), but due to the sheer number of SNPs being typed, the cost per sample is still quite high (around $400 NZD at the end of 2007).

There are two main competing companies who develop and license SNPchip technology, Affymetrix and Illumina. Each company has a different process for genotyping, but the basic concept is essentially the same (see Figure 1.6 and Kim and Misra, 2007). Affymetrix SNPchips are created by attaching 25 base-pair synthetic DNA probes (with a fluorescent tag) to specific locations on a static surface using photolithography. The genomic sample DNA is amplified, hybridised to the SNPchip, and washed to remove unbound sample DNA. Fluorescence intensity is then measured for each probe to indicate which genetic variants have been found in the target DNA. The Illumina SNPchip technology uses small beads as reaction substrates for sample DNA hybridisation. The probes are constructed by combining two fluorescent-tagged allele-specific DNA sequences with a third locus-specific sequence that is used to identify the location of the

[†] http://www.illumina.com/products/human1m_duo_dna_analysis_beadchip_kits.ilmn

[‡] http://www.affymetrix.com/estore/browse/products.jsp?productId=131533

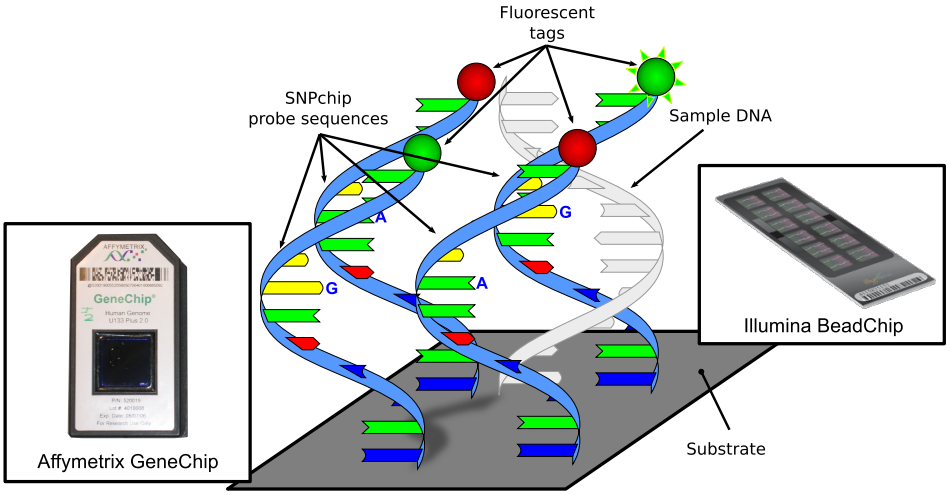

Figure 1.6: A symbolic depiction of a portion of a generic SNPchip. The red-labelled DNA detects the 'T' variant of a mutation (complementary to 'A'), while the green-labelled DNA detects the 'C' variant (complementary to 'G'). The inset images are examples of SNPchips from two main competing companies, Affymetrix (left) and Illumina (right).

variant. Amplified target DNA is hybridised to probe sequence on the beads, which fill small wells and are identified by a DNA barcode specific to each bead. Allele-specific fluorescent signals are combined with the specific barcodes to indicate the location and nature of each genetic variant.

Various analyses and filtering techniques are carried out on the measured fluorescence intensities, eventually producing a list of SNP IDs together with the assayed genotype (and probability of error) at that location.

1.4.3 Genome-Wide Association Studies

A genome-wide association study (GWAS) is a hypothesis-generating approach for identifying genetic risk factors for a particular trait, using SNPchips or similar large-scale genome-wide assays. No initial assumptions are made regarding where in the genome an association may be, allowing for the discovery of unexpected links between genes and disease. Recent

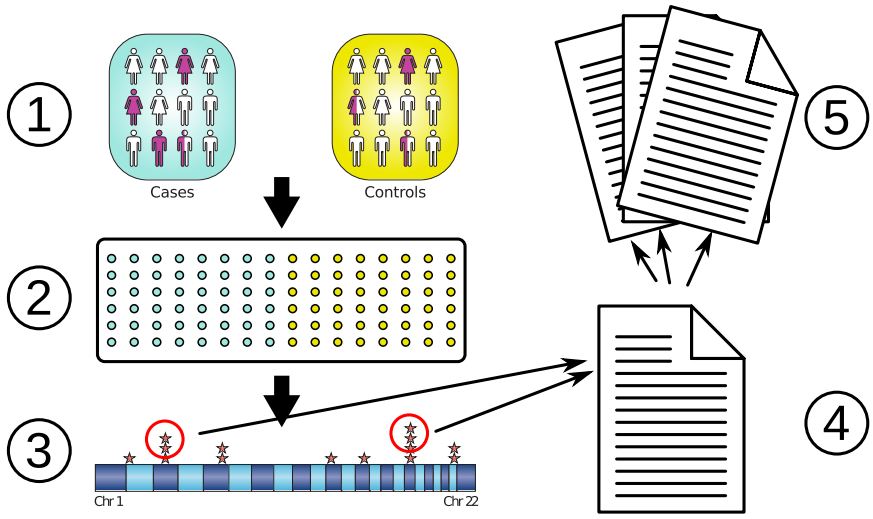

Figure 1.7: The processes involved in a standard genome-wide association study include case/control recruitment (1), genotyping (2), calculation of association statistics (3), reporting of highest associations (4), and validation in independent populations (5). This figure is adapted from Figure 1 of Mathew (2008).

GWAS test thousands of individuals, and involve researchers from all over the world.

The general process for a GWAS is as follows (see Figure 1.7):

1. Recruit case and control individuals from the same source population
2. Genotype individuals at a large number of loci across the genome
3. Compare genotypes for cases and controls at each locus using a relevant association statistic
4. Report loci and genomic regions with the highest association for the case phenotype
5. Confirm candidate associations by repeating association tests for reported loci in a separate population

A benefit of genome-wide association studies is that they allow investigators to screen out genetic variants that are not associated with a particular trait or disease, and so reduce the cost of further genotyping in a larger group of people.

1.4.3.1 Wellcome Trust Case Control Consortium

The technology of the SNPchip introduced the world to relatively cheap genome-wide genotyping, and consequently association studies that covered the entire genome. The Wellcome Trust Case Control Consortium (WTCCC) was set up soon after the introduction of the SNPchip, with a goal to use this technology to identify novel genetic variants associated with common diseases (Wellcome Trust Case Control Consortium, 2007).[†] The initial study investigated genetic associations for seven common diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes).

Individual-level genotype data from this study is available by application to the WTCCC data access committee, and provides an abundance of information that can be used for exploratory research of disease susceptibility through genetic variation.

The success of the initial WTCCC study has allowed collaborators to extend their work into a second phase (WTCCC2), with a goal of typing over 60,000 participants at over 600,000 genome-wide locations for 13 more common disease conditions.

1.4.3.2 Purpose of Association Studies

The purpose of a genome-wide association study is to discover *associations* between particular markers and the disease of interest. There is an

[†] `http://www.wtccc.org.uk`

expectation that these associations may provide an insight into causes of disease, but this will not happen in all situations because a correlation (or association) will not always indicate a causative link. Furthermore, if a marker is not a causative genetic variant, but lies near a causative variant, then risk will not necessarily be the same in different populations due to recombination or mutation. Also, similar allele frequencies for a given marker in different populations may have different associated risks due to that marker having a different haplotype background in each population (see Neale and Sham, 2004).

1.4.4 Common Statistics used in Association Studies

An integral part of association studies is the calculation of statistics for demonstrating association (or lack thereof). No particular statistic is the best option for testing association in all cases. A number of statistics are used in this thesis when testing for association, and testing for differences at a population level. These statistics are described in the next few sections (Sections 1.4.4.1 to 1.4.5.1).

1.4.4.1 Delta

The simplest calculation to determine genotypic differences between populations is Delta (Δ), the absolute allele frequency difference between two populations. Assuming possible alleles denoted $a_1, a_2 \ldots a_n$ with the frequency of allele k in population 1 and population 2 being $f_{k,1}$ and $f_{k,2}$ respectively, then Δ can be calculated for a particular marker as follows:

$$\Delta = \sum_{k=1}^{n} \frac{|f_{k,1} - f_{k,2}|}{2} \tag{1.4}$$

For a dimorphic SNP, this formula reduces to the absolute difference in frequency of any variant in both populations:

$$\Delta = |f_1 - f_2| \tag{1.5}$$

Delta is used to give an indication of the degree to which all the possible variants at a genetic locus differ in frequency between two populations. Delta is easy to compute, providing a quick overview of differences between populations. However, the averaging nature of the statistic means that large differences in frequency of one variant can be masked by many small differences in frequency of other variants.

Rosenberg et al. (2003) discuss a statistic (Informativeness for Assignment, I_n) that may perform better than Delta in some cases. In a two-population model testing markers with the same Delta, this statistic increases as allele frequencies within each population deviate from 0.5. This relationship makes sense, as a variant that has a high frequency in a particular population will predict membership for that population better than a variant that is similar in frequency to other variants for that marker. A limitation of this statistic is that it has upwardly biased estimates of informativeness for small population samples. However, this statistic may be useful for association testing as this upward bias would be expected to affect all markers in a similar manner, so markers could at least all be evaluated with respect to each other within particular populations.

1.4.4.2 Chi Squared

Another statistic that can be used in genetic association studies is Chi Squared (χ^2), typically used to determine if differences in counts between groups are due to variation that would normally occur through a random selection of members of those groups. The Chi Squared model compares

observed counts with expected counts, and a distribution (the χ^2 distribution for given degrees of freedom) can be used to ascribe a probability value to the differences between the observed and expected counts. This probability (p) value is interpreted as the probability that the differences are due to chance; it is common to conclude that for p values greater than 0.05, chance effects cannot be ruled out as a factor that explains differences.

The χ^2 statistic compares observed and expected *counts* (not frequencies), so individual-level genotype data is necessary to calculate this statistic. The result of a χ^2 calculation is a sum of the square of differences between observed (o) and expected (e) counts for all cells (c) in a table, scaled by the number of expected counts:

$$\chi^2 = \sum_{c=1}^{n} \frac{(o_c - e_c)^2}{e_c} \tag{1.6}$$

The most common χ^2 calculation used in association studies for a dimorphic SNP considers the observed counts to be the counts for each allele (or each genotype) in both case and control (or affected and unaffected) individuals. The expected counts are mean counts for all individuals, scaling counts so that expected row totals are the same as the observed row totals.[†] It is also possible to calculate χ^2 values using observed counts as counts for case individuals, and expected counts as counts for control individuals.

1.4.5 Contingency Table Analysis

After a genetic association between groups has been found, it is useful to know how good that genetic variant is at predicting group membership. Additional association statistics can be calculated from a *contingency table*[‡], a 2×2 table that shows error rates for test outcomes if a particular model is

[†]This is the calculation used by the computer program *Plink*.
[‡]also known as a *confusion matrix* (see Fawcett, 2006)

assumed. Columns of the table are categories (indicating the true value of a particular phenotype), rows are test outcomes (indicating the predicted value of the phenotype), and cells are filled with counts of categories with each outcome (assuming an *enriched* control group that contains no cases):

	Case	Control
Positive	(case,+)	(control,+)
Negative	(case,-)	(control,-)

Cases that produce a positive result are called *true positives*, cases that produce a negative result are called *false negatives*. Conversely, controls that produce a positive or negative result are called *false positives* or *true negatives* respectively. False positive and false negative results (also referred to as Type I errors or Type II errors respectively) are undesirable outcomes, and tests are often modified in an attempt to reduce the frequency of these false results.

1.4.5.1 Quantitative Test Results

In a situation where continuous quantitative data are available for each individual, there is frequently no ideal cutoff value for distinguishing groups (see Figure 1.8). The choice of cutoff value will change depending on the use of the test.

1.4.5.2 Positive and Negative Predictive Value

Two statistics commonly used for the purpose of evaluating the effectiveness of a test from a quantitative distribution are sensitivity (proportion of cases that are correctly classified as positive) and specificity (proportion of controls that are correctly classified as negative). A sensitivity of 100% means that a test will produce a positive result for all cases (i.e. no false

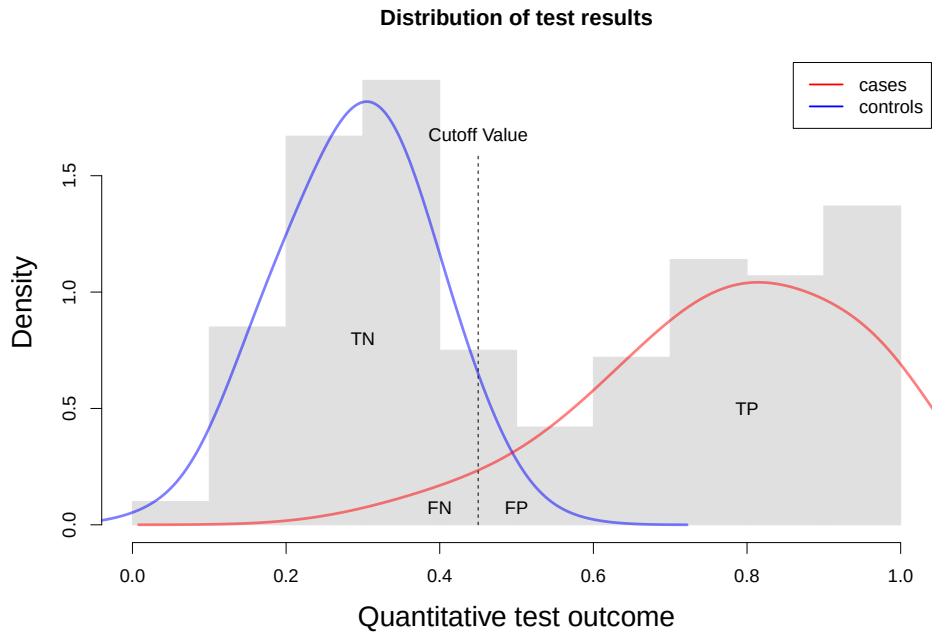

Figure 1.8: A simulated distribution of quantitative test results, with results sampled from two groups. The distribution of the two groups overlap, such that for any test cutoff that is picked over the range of results, there will be at least some false negative (FN) results and/or some false positive (FP) results. If the cutoff value is increased, then the number of true negative (TN) results increases, but the number of false negative results also increases. When the cutoff value is decreased, true positive (TP) results are increased as well as false positive values.

negative results), while a specificity of 100% means that no false positive results will be generated for any controls. Sensitivity and specificity are related to each other – as one is increased (e.g. by adjusting the cutoff value of a test), the other will decrease (but not necessarily at the same rate). Clinicians use sensitivity and specificity to help determine the appropriate use of a test, e.g. screening a population, testing an at-risk individual, or validating the result of a previous test.

Positive and negative predictive values are statistics that are also used for investigating the usefulness of a test. The *positive predictive value* (PPV)

indicates the likelihood that a positive test result is a true positive result. The *negative predictive value* (NPV) indicates the likelihood that a negative test result is a true negative result. A positive predictive value of 20% means that only 20% of positive test outcomes are likely to be correct, and likewise for negative outcomes with a negative predictive value of 20%. A correct calculation of these statistics requires knowledge of population prevalence for each class. For example, when a test is used in a population that has a lower prevalence of cases than the original tested population, the positive predictive value of the test will be reduced.

1.4.5.3 ROC Graphs

A receiver operating characteristics (ROC) graph can be produced to demonstrate how the choice of cutoff values influences the outcome of a test (see Figure 1.9). This graph is generated by plotting false positive rate versus true positive rate at all possible cutoff values for the quantitative statistic. The *Area Under the Curve* (AUC) of this graph indicates how well the quantitative test is able to predict the category of an individual across the range of cutoff values, based on the likelihood of producing a positive result for a case individual rather than a control individual (see Fawcett, 2006). The AUC represents the probability that a randomly selected indivdual from the case group (i.e. those individuals that should be reported as positive) will have a greater test value than a randomly selected individual from the control group (see Zweig and Campbell, 1993).

1.4.6 Population Sampling and Statistical Uncertainty

Genotyping entire human populations is a near-impossible exercise: populations are dynamic due to migration, birth, and death; it is unlikely that all individuals in a population will even agree to genotyping; and carrying out whole-genome genotyping on an entire population is prohibitively

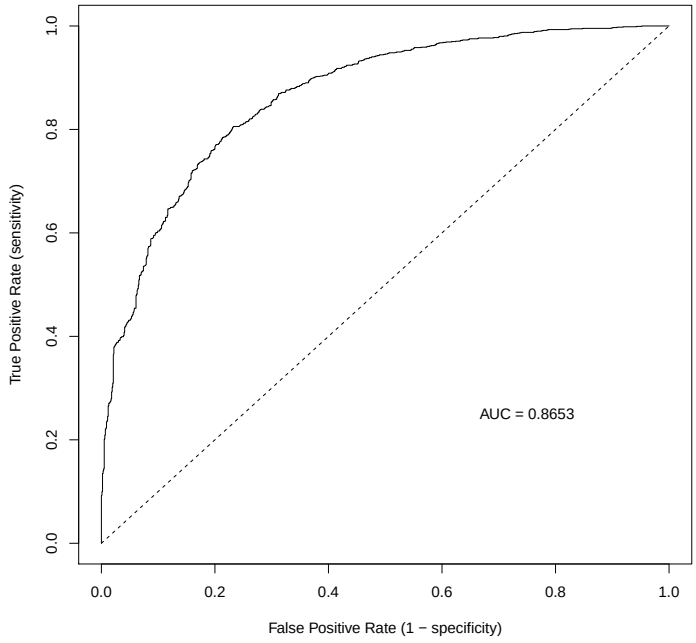

Figure 1.9: An example receiver operating characteristics (ROC) graph – a line graph of true positive rate vs. false positive rate at all possible cutoff values for a quantitative test. The Area Under the Curve (AUC) of this graph is 0.8653.

expensive. Therefore, any particular recruitment exercise for genome-wide genotyping can only type a small *sample* of a particular population from which some inferences about the population as a whole can be made. This process of sampling the population introduces some error (or statistical uncertainty), as genotype frequencies within samples differ from the frequencies in the total population (see Weir and Cockerham, 1984). As the proportion of the population that is genotyped increases, this sampling error will reduce.

A single sample of a population (as used in most GWAS) cannot be used to estimate the sampling error of a descriptive statistic. This is because there is no variance in any frequencies calculated from the sample – genotype frequencies will stay the same for the same group of people no matter how

many times the calculation is made.[†] In order to determine sampling error, multiple samples are required; the variation observed across samples can be used to estimate the accuracy of genotype sampling. This is an expensive process – setup costs are minimal compared to the cost of recruitment and genotyping, so experiment costs for n times the number of samples will be approximately n times the cost for a single sample.

1.4.6.1 Bootstrapping by Population Sub-Sampling

Bootstrapping is a term to describe a method that uses available data to generate more information about that data, so called because it is like pulling yourself up by your own boot straps (or shoe laces). Computers use a bootstrap process (known as booting) to start up; a computer initially has no idea about the state of any of its components, and runs a sequence of steps to get all the components into a known state.

In biology, the term *bootstrapping* is most commonly used in phylogeny, where the probability of particular tree branches is estimated by removing small sections of the aligned sequence in all individuals then recalculating the most likely tree (see Campbell and Heyer, 2002, Chapter 2.1, p. 45). Each removal alters the available information very slightly, which can cause some different branches to be preferred over branches present in the original tree. The bootstrap value generated for these trees indicates the confidence that the compared genetic sequence produces a particular branch structure.

Bootstrapping can also be used to determine the reliability of genotype frequencies in a population, by removing small numbers of individuals from the sampled population and then recalculating frequencies. This *bootstrap sub-sampling* can help to estimate sampling error by simulating the process of taking multiple samples to estimate this error (Jain et al., 1987).

[†]assuming no sampling error.

Bootstrap sub-sampling can be used to estimate relative error, i.e. when allele p has a lower than average frequency in a population sub-sample, allele q has a higher than average frequency.

1.4.7 Computer Programs used in This Thesis

Over the course of this thesis, over sixty short programs (or scripts) have been written by the author to supplement the research undertaken for this thesis, including a program that can carry out a bootstrap sub-sampling procedure on genetic data. It is expected that the scripts created for this research will be of use to other people carrying out similar work, and they have been written with the expectation that the script will be adapted by other people in the future.

Most commonly, the purpose of these scripts is to convert data from one file format to another – data is often received in different formats from different researchers, and will usually need to be converted into another format in order to work with a particular program. Scripts have also been written for producing many of the figures seen in this thesis, as well as simple summary statistics. Where more complex calculations (in an external program) have been required to be done repeatedly, a script has been created to automate that process.

Documentation for programs developed over the course of this research project can be found on the supplementary CD. The CD also contains the source code for those programs.

1.4.7.1 Structure

Falush et al. (2003) created a clustering program called *structure* which is used to determine population structure using a set of unlinked genotypes. The expected use case for structure is in the analysis of *population* structure,

but is also used in this thesis for determining the utility of a given set of markers for categorising (or quantifying) a given trait.

The main benefit of using *structure* over other methods is that it can derive estimated ancestry (or group membership) coefficients from the data, rather than a discrete yes/no for each individual. This allows for the use of cutoff values for the assignment of individuals (useful for tweaking false positive and false negative rates in diagnostic tests), and a quantifiable value where a classification is not useful or does not make sense (as in the case of a continuous trait with high heritability such as height).

The program begins by assigning each individual to a given population. This can be either given in the input data as a 'popinfo' flag, or determined randomly by *structure*. After this, the program constructs a probability distribution for the data, considering the population assignments given, and uses that to estimate population allele frequencies. The program then constructs another probability distribution for these data, and uses that to estimate the population that each individual is likely to have originated from. This process is repeated many times (such a process is known as a *random walk*), and the underlying theory suggests that the progression of population assignments through the random walk process can be used as an approximation of the actual populations of origin for each individual.

The documentation included with *structure* describes the general algorithm for each iteration[†] as follows:

1. Sample (population allele frequencies for iteration m) from the probability of (allele frequency) given (genotypes of sampled individuals) and (population of origin of individuals for iteration $m-1$).

2. Sample (population of origin of individuals for iteration m) from the probability of (population of origin) given (genotypes of sampled individuals) and (allele frequencies for iteration m).

[†]typical *structure* runs have about 10,000-100,000 iterations.

In more simple terms, the steps are similar to the following statements:

1. these populations probably have *this* allele frequency distribution therefore...
2. these individuals probably came from *this* population

The *structure* program is able to accommodate admixture into its model by considering the population of origin as a probability, rather than a single discrete value.

The *structure* program produces a text file with individual identifiers and predicted ancestry coefficients (Q values). While the program does provide means to display these data graphically, the visual representation of data (including sorting) is better controlled using an external program. The most commonly used program is Noah Rosenberg's *distruct* (Rosenberg, 2004), a program that allows colour customisation and group re-ordering. For the purpose of this thesis, a custom R script has been designed (`snpchip2structure.R`) which also allows selection of different sorting methods, a scatter plot for two populations, error bars, and a few additional features (see figure 1.10). More details about this script can be found in the `thesis/programs` directory of the supplementary CD for this thesis.

1.4.7.2 Haploview

The computer program *Haploview* can be used to locate and visualise haplotype block patterns within a population (Barrett et al., 2005). The most common visualisation generated by Haploview is the Linkage Disequilibrium (LD) triangle plot, indicating the degree of LD between markers within a particular genomic region (see Figure 1.11). Once haplotype blocks have been identified, *Haploview* determines frequencies of common haplotypes that are present within those blocks, and the correlation between

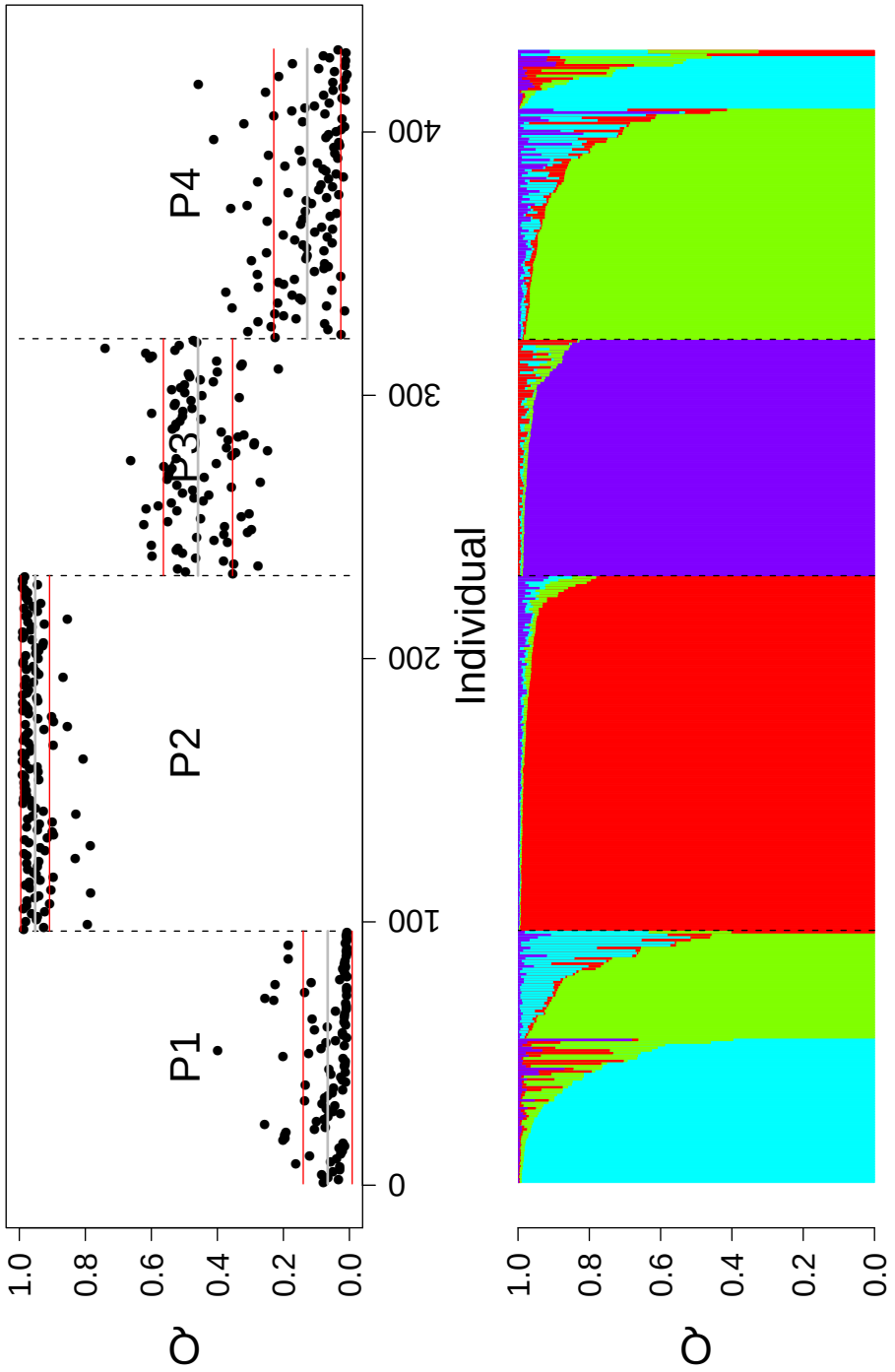

Figure 1.10: Example output from the structure program assuming K=2 populations (above) and K=4 populations (below), converted into graphical representations using `structure2pdf.r`, a custom R script. The top graph indicates mean (grey) and SD (red) values for each of the four input populations: P1, P2, P3 and P4. Individuals within each input population have been sorted by Q value in the bottom graph.

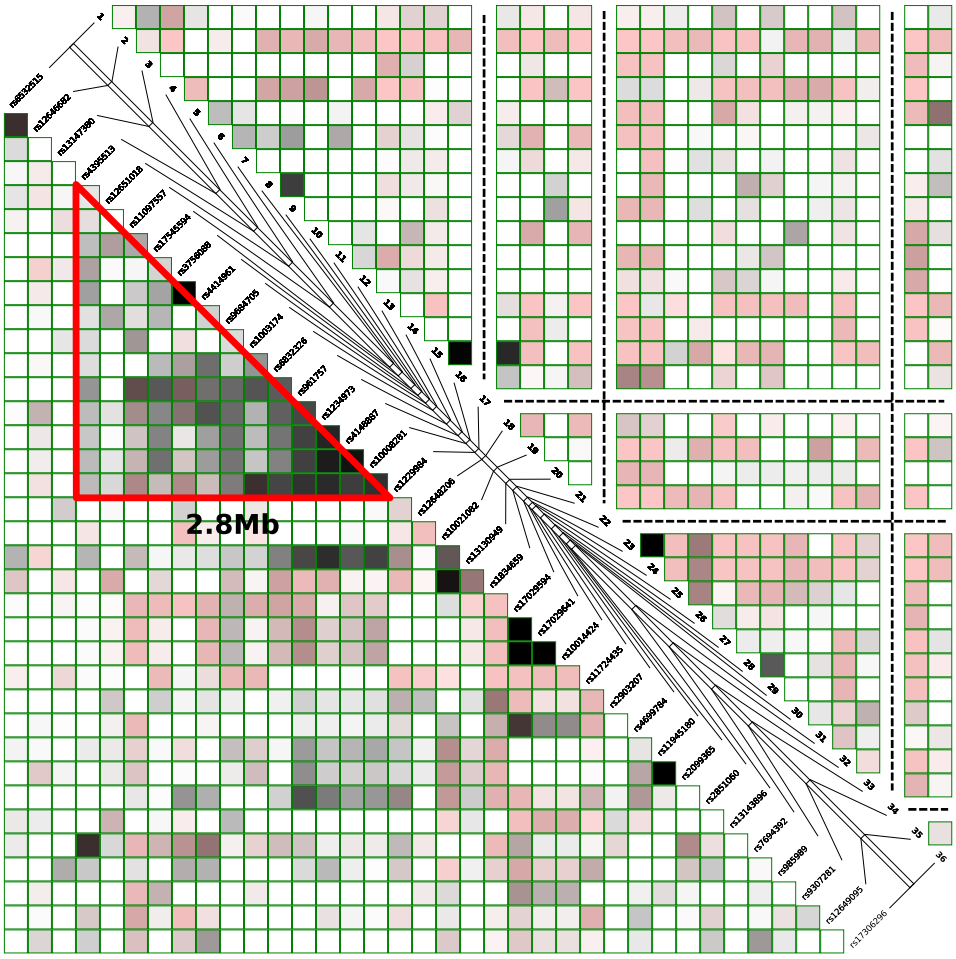

Figure 1.11: An example of the triangle plot *Haploview* output for two populations (combined into one image), showing areas of LD in black, grey and red shading. Black regions have high LD, grey regions have moderate LD, and red regions have high LD but a low p value (so are not considered statistically signifigant). Non-polymorphic sites are identified as dotted lines. This image shows a large block containing markers in high LD in the population below the diagonal (indicated by the red triangle) which is not present in the population above the diagonal.

different haplotypes in different blocks. The program can also be used to identify haplotype-tagging SNPs that can be used to describe common haplotype variation by typing a small subset of SNPs in the population.

The *Haploview* program is primarily a visual aid for finding the extent

of haplotype blocks, and new insights into alternative interpretations of LD statistics mean that the software is under continual development.

1.5 The Maori Population

The Maori settlement of New Zealand represents an end point of a series of island-hopping voyages (see Figure 1.12) throughout the South Pacific ocean – the last of the great human migrations (Murray-McIntosh et al., 1998; Underhill et al., 2001; Hurles et al., 2003; Whyte et al., 2005).

1.5.1 Polynesian Origins

Polynesia is a group of islands in the central Pacific ocean, typically defined as islands within a triangle (see Figure 1.12) with corners at Hawaii, Rapa Nui (Easter Island), and Aotearoa (New Zealand). The Polynesian population (i.e. the native settlers of Polynesia) are quite similar in terms of their culture, biology, and language (see Kayser, 2010; Addison and Matisoo-Smith, 2010).

Kayser (2010) reviewed current literature regarding the genesis of the Polynesian population, a population which has its genetic origin in two main waves of migration. The first wave was an early migration through Island Southeast Asia (ISEA) around 40,000 years ago (40 kya) to Sahul (a land mass which broke up to form Australia and Papua New Guinea around 8 kya). The second wave occurred much later, and is presumed to have originated in Taiwan around 5.5 kya, spreading through southeast Asia with some mixing in New Guinea and the Bismarck archipelago (a group of islands off the northeastern coast of New Guinea) around 3.4 kya, followed by a fast trip through the vast domain of Polynesia (times for this later migration were not specified in Kayser's review). Polynesian populations have a close genetic similarity to tribes from Island South-East Asia,

Figure 1.12: Migration history of Polynesian populations. The approximate dates and migration paths for this figure are from Figure 3 of Chambers (2008). Evidence points to an ancestral population that migrated from Taiwan around 5000 years ago, and dispersed throughout Melanesia and Polynesia. The blue and pink arrows indicated on this diagram represent two different routes of migration that merged in the area of Papua New Guinea to form the Polynesian population. One of these dispersals was a trip from the Cook Islands to begin the Maori settlement of Aotearoa (New Zealand) around 800 years ago.

building on other linguistic and cultural research that suggests a Taiwanese origin (see Friedlaender et al., 2008). Donohue and Denham (2010) suggest that the diversity of New Guinea populations makes it difficult to confirm the origin of the Polynesian migration, and the Taiwanese contribution to the genetics of ISEA populations is relatively minor. However, some of this diversity may have been post-dispersal and mask the passage of genetic signatures through ISEA, even though these signatures are present in the origin and destination populations (Spriggs, 2010).

A review by Addison and Matisoo-Smith (2010) in general supports this explanation of the two-wave origin of the Polynesian population, with early travel to Sahul around 30-40 kya, and mixing in the Bismarck archipelago around 3.3 kya. However, they also suggest a third wave of migration began from Asia between 2 and 1.5 kya that introduced new breeds of plants, animals, culture and ideas. Donohue and Denham (2010) support the idea of multiple introductions of animals and plants into ISEA. Linguistic and archaeological evidence suggests a pause of 500-1000 years between the migration to the west edge of Polynesia (from ISEA) and the subsequent permanent settlement of the remainder of Polynesia (see Hurles et al., 2003). More recent computer simulations suggest that the passage from West Polynesia (Samoa) to East Polynesia (Cook Islands and beyond) would have been a significant challenge for sail and canoe voyages, a plausible hypothesis for this long pause (Di Piazza et al., 2007). Given this pause in time between the arrival in the Bismarcks and expansion into Polynesia, it seems reasonable to suggest that the introduction of new technology from a third migration provided the impetus needed to settle the remainder of the Polynesian islands.

Genetic studies on Y-chromosomal, mitochondrial, and autosomal data are a component of evidence that supports the hypothesis of a dual-wave origin for the Polynesian population. A study by Underhill et al. (2001) identified 9 Y-chromosome haplotypes in the Polynesian ancestral population. In particular, three main lineages were identified that had very low

diversity across the samples in the study, suggesting a recent colonisation of the islands in Polynesia. Kayser et al. (2006a) found that almost all (94%) Polynesian mitochondrial DNA was of relatively recent Asian origin, with 6% derived from New Guinea populations, but the contribution was somewhat reversed in Y-chromosomal DNA (28% Asian origin, 66% New Guinea origin). An analysis of autosomal SNP data by Kimura et al. (2008) found a mixture closer to, but less extreme than, the mitochondrial data (around 70% Asian origin, 28% New Guinea origin). South-East Asian societies were historically matrilocal, i.e. females remain in their local village and males move in with their wives' family (Jordan et al., 2009), and Kayser suggest that the discrepancy between Y-chromosomal and mitochondrial data is due to this matrilocal culture.

1.5.2 Settlement of New Zealand

New Zealand is a geographically isolated island country at the southern edge of the Pacific ocean. The country is composed of three main islands, the North Island (Te Ika a Maui), the South Island (Te Wai Pounamu), and Stewart Island. The native Maori population of New Zealand is believed to descend from island-hopping adventurers from Eastern Polynesia. A restricted group of these people travelled to New Zealand, founding the Maori population in waves around 600-800 years ago (see Marshall et al., 2005; Anderson, 1991).

The series of settlements and radiation through the Pacific ocean created multiple bottleneck and founding effects as populations travelled between islands in Polynesia. The result of such effects would suggest that the Maori population is likely to be fairly homogeneous, with genetic variants having unpredictable frequencies. Historically, the New Zealand Maori were extremely adventurous risk takers (Vayda, 1970), particularly considering the hazards involved in making long journeys across vast stretches of ocean.

The settlement of New Zealand by Maori must have been deliberate. Over 3,000km separate the Cook Islands and New Zealand, requiring a journey of about a month (see McKinnon, 2003, plate 10) – it is hard to imagine that a trip to New Zealand was the result of a shorter expedition blown off course. Also, the remoteness of New Zealand makes it more likely that returning exploratory voyages preceded an intentional migration voyage to New Zealand (Irwin et al., 1990). It is therefore reasonable to conclude that Maori who arrived in New Zealand were selected to be there. Hence social selection, and probably some genetic selection as well, occurred during the establishment of the Maori population.

1.5.3 European Colonisation

The first recording of a European sighting of New Zealand was by Abel Tasman in 1642, when Maori appeared to be well established across the country (see McLauchlan, 1984, p. 527). However, Abel Tasman did not land in New Zealand at that time due to Maori hostilities in Golden Bay (at the North end of the South Island) and Three Kings (North of the top of the North Island).

The next recorded European visit to New Zealand was by James Cook, who arrived in New Zealand in 1769 and eventually managed to forge close associations with local Maori (see McLauchlan, 1984, p. 119). European migration to New Zealand was largely through traders and missionaries through the late 1830s, with large European settlements appearing in the 1840s (see McKinnon, 2003, plate 30).

1.5.4 Population Genetic Insights into Maori Migration

The founding population size for Maori settlement of New Zealand was first estimated at 50-100 females using mitochondrial DNA (mtDNA) sequences,

and the observed variability in mtDNA sequences support a fast settlement over 30 generations (Murray-McIntosh et al., 1998). More recent estimates of the number of females in the initial Maori settlement of New Zealand are higher (170-230), and may be even higher, depending on the speed of population expansion (see Marshall et al., 2005; Whyte et al., 2005).

1.5.4.1 Maori-European Admixture

Variation in genetic sequence has been used to determine the process of change that has occurred in the Maori population over time. A 2003 study determined that self-declaration for individuals with a mixed Maori / European background correlates well with genetic admixture estimated from New Zealand DNA databases (Walsh et al., 2003). An analysis of the most recent New Zealand census data that included ancestral information (1976)[†] found that the Maori population had 37.4% European ancestry by self-report. It is expected that this fraction of European ancestry has increased and is now around 40-50% of the Maori gene pool (see Lea and Chambers, 2007b).

A comparison of Y-chromosomal and mitochondrial DNA (mtDNA) by Underhill et al. (2001) found complex genetic histories in the Maori population. The diversity of Y-chromosome DNA was greater than the diversity of mtDNA in Maori; three core non-European Y-chromosome haplotypes were identified, but only one non-European mitochondrial haplotype was found. Their results support a history of mixing of multiple Austronesian populations before colonisation of Polynesia, and multiple migrations of Maori ancestors to New Zealand.

A more recent study of genetic data from 687 microsatellites found that Maori and other Polynesian populations are most similar to Taiwanese aborigines (and then East Asians), with a small amount of European ancestry,

[†]The New Zealand census no longer collects ancestral information

probably due to admixture following European colonisation. This supports a hypothesis that the Polynesian islands were populated by voyagers starting from the vicinity of Taiwan, and is consistent with evidence from Y-chromosomal and mitochondrial DNA (Friedlaender et al., 2008).

Genetic studies of the Maori population have concentrated on rare diseases that are prevalent in large Maori families, identifying causal mutations for diseases including gastric cancer and malignant hyperthermia. Recent studies have also been carried out investigating the genetics of nicotine, alcohol, and other drug metabolism that suggest treatment strategies for Maori may be different than strategies for other New Zealand populations (see Lea and Chambers, 2007b).

Genetic variation is lower in Polynesian individuals than in either European or Asian individuals, and the Maori population have even lower genetic variation (see Marshall et al., 2005).

The preliminary analyses of Maori genetics carried out in the author's Honours thesis (Hall, 2004) have provided supporting evidence for genomic uniqueness and reduced diversity of the Maori population. In particular, Maori and European *populations* were easily distinguishable using data from a genome-wide panel of forensic markers. However, no particular combinations of forensic marker variants were identified in Maori that were unique to the Maori population, so *individuals* could not reliably be assigned to either population.

1.5.5 Maori Health

There are many diseases for which Polynesian populations (with Maori as a subset) have a significantly higher frequency in comparison to European populations. These diseases include coronary heart disease, low cholesterol, asthma, gout, measles and iron deficiency. Other diseases have a

reduced frequency of occurrence among the Polynesian ancestral population: rheumatoid arthritis, melanoma, Crohn's disease, and depression. Many (if not all) of these diseases have possible genetic influences that affect disease prevalence in the population (see Abbott et al., 2001). More recent studies have been carried out that demonstrate differences in disease and disease-associated trait frequency between Maori (specifically) and non-Maori populations (e.g. Rossaak and Pitto, 2005; Shand et al., 2007; Sundborn et al., 2007). Genetic links to disease mean that studies carried out to identify genetic patterns within a population may also improve the diagnosis and study of disease.

1.6 Hypothesis and Key Questions

An understanding of differences in health outcomes for Maori and European populations in New Zealand, combined with evidence that suggests the Maori population is genetically unique, led to the following core hypothesis to be tested:

> The Maori population has distinct and unique genomic and related disease patterns that can be identified through a combination of DNA polymorphism, bioinformatics, and medical analysis.

This hypothesis suggests a number of key questions that should be answered in order to evaluate whether the hypothesis is valid:

1. Maori are less genetically diverse than European populations, but previous studies have not found any combinations of markers are unique to the Maori populations. Can genetic marker combinations be found that describe the Maori population uniquely?

2. Considering the lower diversity of the Maori population in comparison the non-Maori populations, is the haplotype block model more appropriate for the Maori population. If so, will it be possible to find haplotype blocks with low marker density?

3. How does reported ancestral information compare to observed genetic information?

4. Given the exponential increase in available information in most areas of science, can developing technologies, combined with technologies from other areas, be used to gain new insights on old data?

Lea and Chambers (2007b) reviewed the current knowledge of disease prevalence within the Maori population, and have identified smoking addiction and alcohol dependence as two phenotypes that have different expression profiles in the Maori and European populations. The next chapter of this thesis will explore what insights can be gained from a study of alcohol-metabolising genes in the Maori population.

Chapter 2

The Genetic Structure of the Alcohol Dehydrogenase Genes in the Maori Population

2.1 Overview

This study of the Alcohol Dehydrogenase (*ADH*) gene cluster on chromosome 4 is an extension of the previous work of Chambers et al. (2002b). It represents a further attempt to identify linked polymorphisms within the ADH loci that may also have utility as markers for identifying susceptibility to alcoholism. It is argued that the unique and recent migration history of the Maori population will help to explore gene interactions within this region for genetic variants that are not common in other populations. This chapter is an extension of a study by this author and others that has already been published in the Journal of Human Genetics (Hall et al., 2007) – see Appendix A.

2.2 Background

Alcohol is the most commonly used behaviour-altering recreational drug in New Zealand (Sarfati and Scott, 1999). In 2003, approximately 81% of all adults (18+) reported that they were current drinkers (McMillen et al., 2004), and about 10 litres of alcohol per adult (including non-drinkers) is produced and available for consumption each year (Alcohol Advisory Council of New Zealand, 2005).[†] Alcohol consumption has high costs to society due to crime, reduced work output, hospitalisations and other diverted resources. The social cost of harmful alcohol use in the 2005/2006 year was estimated to be 4.8 billion dollars (Slack et al., 2009).

2.2.1 Historical Maori Drinking Patterns

Hutt (2003) provides a valuable account of the history of the use of alcohol in the Maori population. According to the author, alcoholic beverages were not present within the Maori population prior to the arrival of European settlers in New Zealand. The Maori word for alcohol is *waipiro*, literally 'stinking water', which suggests that at its introduction, alcohol was not palatable by many in the Maori population.

The New Zealand government imposed many laws and restrictions on the Maori population regarding access to alcohol – the most significant of these being the Ordinance to Prohibit the Sale of Spirits to Natives in 1847.[‡] The creation of these laws suggests that Maori were expected to abuse alcohol, despite their initial distaste for the substance. Having observed the effects of alcohol consumption, Maori were well aware of the behavioural traits associated with drinking alcohol. A number of prominent Maori leaders went beyond these laws from 1850 onwards, restricting access further within their own people. Discriminatory legislation was removed

[†] http://www.alac.org.nz/NZStatistic.aspx?PostingID=4346
[‡] http://www.nzlii.org/nz/legis/hist_act/sostna184711v1847n3430/

after 1948 (see p. 73 of Hutt, 2003), and refusal to serve alcohol to patrons on the basis of race was made an offence in 1962 by the Sale of Liquor Act.[†]

Alcohol was very rare in trade between Maori and Europeans before 1840, but by the 1860s it started to be associated with the formation of political alliances, so became more desired as an item of trade. From around 1870 to 1920, consumption of *waipiro* at important occasions became more evident, and acceptance of drinking anyway was more common. While the drinking habits varied greatly within the Maori population, there was still a much lower consumption compared with Europeans. In the post-war period of the 1950s, brewers began targeting advertising for a Maori market, and there was an increase in the consumption of alcohol within the Maori community as a whole.

2.2.2 Recent Maori Drinking Patterns

According to the 1996/97 Health Survey (Sarfati and Scott, 1999), 27.4% of Maori reported that they had not had a drink in the last year, compared to 12.9% for Europeans. Despite this difference in the proportion of drinkers, the mean quantity of alcohol consumed is similar between the two groups, because Maori drinkers typically consume more alcohol per drinking session.

In 2000, a detailed survey was carried out to identify Maori drinking patterns and alcohol-related problems (Barnes et al., 2003), but did not include non-Maori populations for a comparative study. Among Maori, 20% abstain from alcohol altogether, but those who do drink consume a large quantity of alcohol annually, much higher than the national average (22 litres for Maori males, 8 litres for Maori females).[‡] Maori drinkers consume alcohol around once every three days, with about half of the

[†] `http://www.nzlii.org/nz/legis/hist_act/sola19621962n139186/`

[‡] Alcohol consumption figures are standardised to pure alcohol. A 330ml bottle of 5% beer contains 16.5ml of pure alcohol

alcohol being drunk at residential dwellings. The report produced for the Alcohol Advisory Council of New Zealand (ALAC) by McMillen et al. (2004) compared the current trends of alcohol consumption behaviour between Maori and European populations. The proportion of regular drinkers (at least one drink per week) in the Maori population (39%) is less than that of the European population (56%), but these individuals will typically drink more per session than drinkers in the European population. A statistic that demonstrates this contrast is the amount consumed in the last drinking session: 22% of adult Maori drinkers consumed more than 10 standard drinks of alcohol, compared to 8% for the European population.

2.2.3 The Classification of Alcoholism

While *alcohol consumption* is fairly easy to characterise and quantify (e.g. quantity of alcohol consumed per week / session), *alcoholism* is not. Alcoholism refers to an entire class of psycho-social disorders that are based upon the misuse of alcohol. Prior to 1980, alcoholism was considered as a single disorder. The introduction of the *Diagnostic and Statistical Manual of Mental Disorders, Third Edition* represented a change in how alcoholism was classified, using specific diagnostic criteria and separating it into alcohol dependence and alcohol abuse (see Hasin, 2003).

Cloninger (1987) also split alcoholism into two types: type 1, which has a late onset (> 25 years) and an emphasis on psychological dependence and guilt, and type 2, with an early onset and emphasis on aggressive behaviour and an inability to abstain.

There are three main sets of diagnostic criteria for alcohol dependence / abuse classifications that are in current use, with each set having slightly different definitions (see Hasin, 2003). Hence, individuals classified as dependent using one set of definitions will not necessarily be put into the same class using one another diagnostic measure. The *Diagnostic and*

Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R) classification is still used by researchers due to historical publications that also use these criteria, even though an updated version (DSM-IV) is used by clinicians, and a different classification (ICD-10) is used by the World Health Organisation (WHO).

It is tempting to try to place patients into distinct groups based on what type of alcoholism they have. For instance, Moss et al. (2007) attempted to subdivide alcoholism even further, identifying no fewer than five classes of alcoholism. Their rigid approach does not seem to fit the complex nature of the disease. A better viewpoint may be to see the two behaviours (abuse and dependence) as two interacting dimensions of alcoholism (with the possibility of more unknown dimensions being present), with a continuum of severity possible for both types. This view is shared by Helzer et al. (2006), who have written a review that discusses the benefits and disadvantages of this multidimensional approach to alcoholism.

2.2.4 Genetic and Environmental Contributions to Traits

A common public misconception relating to genetic traits is that a trait with a genetic component is inescapably influenced by genetic factors – this is referred to as genetic determinism (see Condit, 2007), portrayed as the common world view in the film *GATTACA* (see Kirby, 2000). However, every human behavioural trait will always have some greater or lesser genetic component and some environmental component that modifies its expression.

Heritability is a commonly used statistic that describes the proportion of variation of a given trait contributed by genetic factors. In the case of alcoholism, heritability estimates are generally in the range of 40% to 60% (Messas and Filho, 2004; Dick and Foroud, 2003), although these vary considerably depending on what particular facet of alcoholism is being looked at (e.g. Liu et al., 2004; Prescott et al., 2005; Saccone et al., 2000).

One demonstration of an environmental effect influencing alcohol dependence is the recent study by Guéguen et al. (2008) on the association between sound level and alcohol consumption (one of the typical components of diagnostic criteria for alcohol dependence). Researchers recorded the consumption (number of drinks, gulp size, time to finish) of 250ml glasses of draft beer by patrons of two bars in France. The study found that loud (88dB) music resulted in an increase in drink consumption rate when compared to background-level (72dB) music.

2.2.5 Genetic Contributions to Alcoholism Risk

Alcoholism is a complex disease that has several primary and secondary contributory factors throughout the genome – it is clearly not a single-gene disorder (see Devor, 1993). Recent reviews of genetic influences on alcoholism have been written by Schuckit (2009) and Nurnberger and Bierut (2007).

There are two main classes of ethanol-metabolising enzymes, Aldehyde Dehydrogenase (*ALDH*) and *ADH* (see Figure 2.1); particular variants of the genes encoding these enzymes have demonstrated variation in the response to ingested ethanol. A particular variant of the ALDH genes that is common in Asian and Polynesian populations (see Chambers et al., 2002b) has been linked to a flushing reaction and some discomfort after the consumption of alcohol. The combination of this variant and a slow-metabolising *ADH* gene variant substantially decreases the risk of alcoholism (see Nurnberger and Bierut, 2007), likely due to increased levels of aldehyde remaining in the blood for an extended period of time.

Schuckit (2009) suggests that a reduced level of response to alcohol (i.e. a need to consume more alcohol in order to obtain a pleasurable effect) seems to increase the risk of alcoholism. It has been hypothesised that people who need to drink more to enjoy feeling a pleasurable effect from alcohol

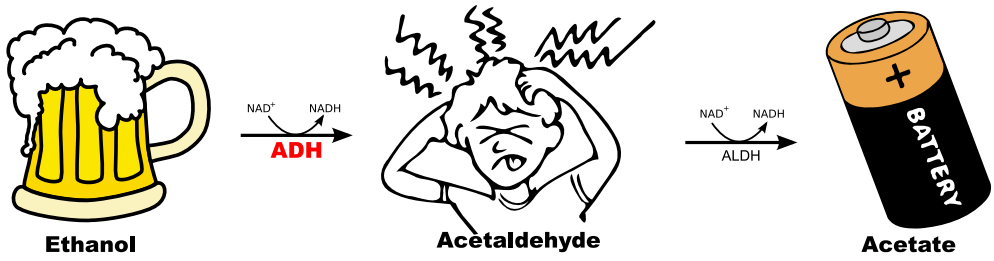

Figure 2.1: Ethanol (CH_3CH_2OH) is converted in the liver into acetaldehyde (CH_3CHO), by the *ADH* enzymes, before being converted into acetate (CH_3COO^-) by *ALDH* detoxification. This end product of alcohol metabolism can then be used by the body as an energy source.

will be more likely to drink more per session, and in turn, associate with other heavy drinkers (reinforcing this behaviour of increased consumption). Supporting this hypothesis, Schuckit (2009) notes that variants of genes in the gamma-aminobutyric acid (*GABA*) receptor gene cluster are associated with reduced sensitivity to the effects of alcohol, and also that reduced serotonin levels (e.g. due to increased re-uptake) are suspected to be related to a reduced response to alcohol.

Risk of alcoholism is also modified in variants of genes that alter concentrations of other neurotransmitters. For example, the *CHRM2* gene encodes the M2 muscarinic acetylcholine receptor, part of a family of receptors that are involved in learning, memory and cognition (see Wang et al., 2004). Individuals with a common T-T-T haplotype within intron 4 of this gene have a reduced risk of alcoholism (Wang et al., 2004), such as dopamine (e.g. a *DRD4* exon 3 polymorphism, Bau et al., 2001) and acetylcholine (e.g. a TTT haplotype in intron 4 of *CHRM2*).

While there have been many hypotheses generated about genetic predisposition to alcoholism (including linkage studies such as in Edenberg et al., 2006 and Djoussé et al., 2005), validation of these hypotheses has been difficult. The complex interactions between different genes for disorders such as alcoholism may mean that a validated genetic interaction that works for all situations is an unobtainable target.

The study presented here is an analysis of the *ADH* gene cluster, which has a role in the metabolism of alcohol and is a good candidate for a region that has an influence on alcohol dependence.

2.2.6 Summary of the *ADH* Genes

2.2.6.1 Enzyme Function & Activity

The hepatic alcohol dehydrogenases (*ADH*, EC 1.1.1.1) are a family of catabolic enzymes that oxidise alcohols, including ethanol, to form acetaldehyde (see Crabb et al., 2004). These proteins are dimeric, composed of a combination of five different classes of *ADH* subunits, α, β, γ, π, χ and μ. Although the $\alpha - \alpha$ homodimer is the most common form, multiple homodimeric and heterodimeric configurations are characteristically observed in liver extracts (e.g. Bosron et al., 1983). These subunits are encoded by a cassette of linked genes (recently renamed as *ADH1A*, *ADH1B*, *ADH1C*, *ADH4*, *ADH5*, *ADH6*, and *ADH7* respectively)[†] positioned in order on the long arm of chromosome 4 (4q21-4q25). Previous studies have shown that the liver activity of *ADH* varies among individuals and between human geographic sub-populations, and may influence metabolic response to ingested alcohol and susceptibility to abuse behaviour (e.g. see Chambers et al., 2002a; Lee et al., 2004). Variation in enzyme activity is due in part to variation within the genes that encode the enzymes. This variation can increase the complexity of interactions between *ADH* dimers, as variants of the same class of *ADH* enzymes can have differing kinetic parameters (Lee et al., 2004; Crabb et al., 2004).

[†] *ADH1* $\rightarrow$ *ADH1A*, *ADH1B* $\rightarrow$ *ADH2*, *ADH1C* $\rightarrow$ *ADH3*

2.2.6.2 Previous Research on *ADH* Polymorphisms

A variant that is known to alter enzyme activity is a single nucleotide polymorphism (SNP) within the *ADH1B* gene, rs1229984. It is an exonic SNP, coding for an Arginine to Histidine amino acid change in the β subunit polypeptide of the human *ADH* protein (see Lee et al., 2004). The rare variant of this mutation has been widely associated with reducing the risk of alcohol dependence (Higuchi et al., 2004; Osier et al., 2002; Chen et al., 1997), and has been found to have a high prevalence in Asian and Pacific populations (Chambers et al., 2002b; Chen et al., 1997). The proposed mechanism of protection is based upon metabolic properties that result from this genetic variant. It has been observed that an *ADH1B* protein with the *ADH1B*47His* variant has a higher affinity for ethanol than an *ADH1B* protein without this variant (Higuchi et al., 2004). The metabolic product, aldehyde, is quite toxic and produces a number of unpleasant physiological effects (such as nausea and headaches). Due to these unpleasant effects, it is expected that this genetic variant could make a person less likely to consume alcohol in large quantities.

Edenberg et al. (2006) have carried out a systematic analysis of the seven genes within the *ADH* region, looking at Linkage Disequilibrium (LD) patterns and association with alcoholism in families from the Collaborative Study on the Genetics of Alcoholism (COGA).[†] They found the strongest evidence for association with alcohol dependence around the *ADH4* gene, but evidence for association in the region around *ADH1B* only reached statistical significance when a broader definition of dependence was used. They also discovered that LD is high within genes and lower in regions between genes.

[†] `http://zork.wustl.edu/niaaa/`

2.2.6.3 Frequency Differences Within and Between Populations

Chambers et al. (2002b) looked at genotype and allele frequency differences at three loci within *ADH* and *ALDH* genes. They typed alcohol dependent (DSM-III-R) Maori and non-Maori males, as well as control subjects with European, Asian, and Polynesian ancestry (including New Zealand Maori, Cook Island Maori, and Samoan ancestry).

It was found that the *ADH2*2* (rare rs1229984 variant) frequency in Maori alcoholics (0.15) was significantly less $(p < 0.01)$ than that in Maori controls (0.42), and the difference between alcoholics and controls was greater than that observed in any other group for which the variant has been typed. The reported frequency of this variant in most European populations is very low (around 0.03), high in Asian populations (around 0.76), but has similar frequencies in Maori and other Polynesian populations (0.42-0.46). This result was interesting especially considering that Polynesian population more closely resembled the European population than the Asian population at a SNP found in the ALDH2 gene. While being able to show the protective effect of the rare rs1229984 variant in Polynesians (irrespective of other known protective variants), Chambers et al. (2002b) were unable to determine if this protective effect was also present in European populations, due to low frequencies and small sample sizes in the study. Edenberg et al. (2006) have also failed to validate this protective effect in European populations, but also reported similarly low allele frequencies (0.034) of the rare rs122984 variant.

2.2.7 Expectations of Haplotype Block Structure

In previous studies of association with alcohol dependence within the *ADH* region it has been common to analyse individual SNPs, rather than combine linked SNPs and analyse haplotypes, even when it is apparent that two or more nearby SNPs reside on the same haplotype block (e.g. Djoussé

et al., 2005; Edenberg et al., 2006). However, some later studies of the *ADH* region have now included an analysis of haplotypes. For instance, Han et al. (2007) identified haplotype block patterns that were consistent with selection in the neighbourhood of the *ADH1B* gene. They used the results from extended haplotype homozygosity (EHH) tests (Sabeti et al., 2002) as evidence in support of their hypothesis of a selection event in Japanese and Korean populations. A single *ADH1B* haplotype showed high homozygosity (> 0.6) over about 100kb in Japanese and Korean populations; this homozygosity was much higher than that observed for other Asian populations, where lower homozygosities (between 0.4 and 0.6) were maintained over a shorter distance (40-80kb).

2.3 Methods

This study is an extension of the design used by Chambers et al. (2002b). The principal aim is to identify linked mutations within the *ADH* gene cluster on chromosome 4 that can be used to determine *ADH* haplotype block variation between the Maori and European populations. The study involves a comparison of experimentally obtained genotypes from Maori individuals with comparable publicly available genetic data from European individuals. The genotypes have been analysed at both an allele frequency and haplotype frequency level.

2.3.1 Study Population

A previous study by Chambers et al. (2002b) collected DNA from 18 male and 29 female individuals (all unrelated), drawn from the general population of Wellington, New Zealand. Banked DNA samples from these individuals were used for new SNP genotyping assays covering nine SNPs within the *ADH* region. The subjects self-reported four Maori grandparents

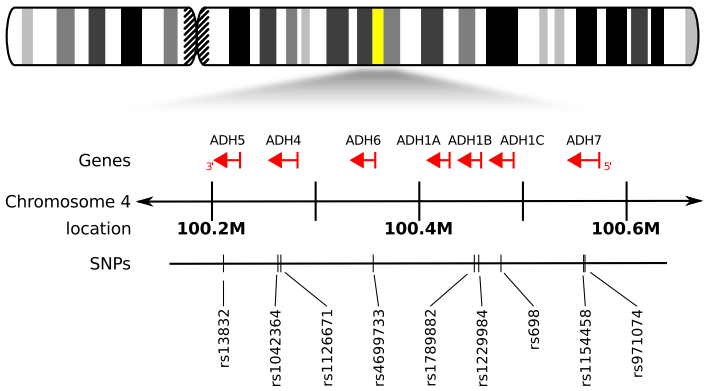

Figure 2.2: A section of chromosome 4, showing the relative positions of the *ADH* genes and SNPs typed in this study. Names (from the NCBI database) of typed SNPs are shown at the bottom.

and as such can be reliably considered as representative of the ancestral Maori population (i.e. they have zero or minimal European genetic admixture).

2.3.2 ADH SNPs Typed

The *ADH* genes are located on the long arm of chromosome 4 (region 4q23), nucleotide locations 10200kb-10600kb according to the NCBI reference assembly (figure 2.2). Nine SNPs were chosen (both intronic and exonic polymorphisms), spread across the entire cassette of *ADH* genes. The SNPs were chosen based on existing knowledge of mutations that were well characterised in the literature – no SNPs were chosen within the *ADH1A* gene because, at the time of SNP selection, no well-characterised mutations were known within that gene. Genotyping was done via a service contract with the Australian Genome Research Facility (AGRF)[†], which used the

[*] All references to locations of genes and mutations in this chapter are based on the NCBI reference assembly for *Homo sapiens*, build 36, March 21, 2006. The names of genes and mutations refer to the Gene name and refSNP ID respectively in the NCBI database as of May 2007.

[†] `http://www.agrf.org.au`

Sequenom MassArray Genotyping system (Buetow et al., 2001).

Haplotype frequencies were determined for two gene regions having two linked SNPs each:

The first haplotype, including the SNPs rs1229984 and rs698, lies within the region of class I *ADH* genes (*ADH1B* and *ADH1C*). The rs1229984 mutation corresponds with a change in amino acid 48 of the *ADH1B* product, from a Histidine (A allele) to an Arginine (G allele). The rs698 mutation corresponds with a change in amino acid 350 of the *ADH1C* product, from an Isoleucine (A allele) to a Valine (G allele). Data from European subjects for this haplotype are from haplotype frequencies for a combination of the 'Europeans, Mixed' and 'Irish' populations in ALFRED, the Allele Frequency Database (Rajeevan et al., 2003).

The second haplotype, including rs1042364 and rs1126671, lies within the region of the Class II *ADH4* gene. The rs1042364 mutation corresponds with a change at mRNA position 1238 ($3'$ untranslated region). The rs1126671 mutation corresponds with a change in amino acid 309 of the *ADH4* product, from an Valine (G allele) to a Isoleucine (A allele). Data from European subjects for this haplotype are from the HapMap dataset (CEU – Utah residents with ancestry from northern and western Europe), calculated using *Haploview* (Barrett et al., 2005).

2.3.3 Sources for Web-based Data

Genotype and haplotype frequency statistics for European subjects were obtained from the HapMap database (International HapMap Consortium, 2005). Additional genotype and haplotype frequency data that have been used came from the ALFRED database (Rajeevan et al., 2003), which contains information on SNP and haplotype frequencies for over 1300 populations at a large number of sites throughout the human genome. Allele frequencies of European subjects for five mutations (rs13832, rs1042364,

rs1126671, rs4699733, and rs1789882) were retrieved from the HapMap database. Allele frequencies of European subjects for rs1229984, rs698, rs1154458, and rs971074 were retrieved from the ALFRED database.

Only haplotype blocks that were in complete LD with high LOD scores in the Maori population were used in the analysis. Haplotype frequencies for a haplotype including the SNPs rs1229984 and rs698 for European subjects were obtained from ALFRED. European (HapMap CEU) haplotype frequencies for the haplotype including the SNPs rs1042364 and rs1126671 were determined using *Haploview*, with blocks defined by the Four Gamete rule (Wang et al., 2002).

2.3.4 Statistical Analyses of *ADH* Variants

Frequencies of haplotypes and alleles were determined for Maori and European populations. Allele frequencies were generated from the genotype data obtained from AGRF, as well as from the ALFRED and HapMap databases. To determine probabilities associated with frequency differences, χ^2 values were calculated, and probabilities were determined using the CHIDIST function in OpenOffice.org Calc (Sun Microsystems, 2009). The χ^2 values have only one degree of freedom when comparing allele frequencies between two populations. The *Haploview* program (Barrett et al., 2005) was used to identify LD (haplotype) patterns among SNPs within the Maori and European populations.

Calculation of the χ^2 values for the comparisons of haplotypes were carried out in a similar way to those for allele frequencies. They also have one degree of freedom, comparing each haplotype to all other haplotypes between the Maori and European populations.

2.4 Results

2.4.1 Comparison of Allele Frequencies in Maori and European Populations

Individual χ^2 tests were carried out for each SNP, comparing Maori and European allele frequencies (Table 2.1). All calculated probabilities of similarity were below 0.01 except at rs4699733 ($p = 0.18$) and rs971074 ($p = 0.56$). The largest difference in allele frequency was at rs1229984: the rare allele was present in 45% of the Maori population, but only 4% of the European population. Among those SNPs with significant differences in allele frequencies, only rs1229984 had a greater frequency of the rare allele in the Maori population. The rare allele frequencies for the Maori population shows a general trend of reduced prevalence in the population as the distance from rs1229984 increases ($r^2 = 0.63$). This trend is not mirrored in the European population ($r^2 = 0.17$), in which most rare allele frequencies are near $30 - 35\%$.

The smallest difference in allele frequencies was at rs971074, where frequency of its rare allele for Maori was 4%, while the rare frequency for European was 12%. Three of the nine SNPs involve non-synonymous mutations in coding sections of DNA. All of these three SNPs had significant differences in allele frequency between the Maori and European populations.

A total of six SNPs (rs13832, rs1042364, rs1126671, rs698, rs1154458, rs971074) had a rare frequency in the Maori population of less than 30%, while only three had a frequency of less than 30% in the European population (rs4699733, rs1229984, rs971074). Only one rare allele, that of rs971074, had a frequency of less than 30% in both the Maori and European populations.

RefSNP ID	Gene	Mutation	(Coding)	Maori	European
rs13832	*ADH5*	T→G	–	0.10(90)	0.38(164)
rs1042364	*ADH4*	G→A	–	0.05(94)	0.32(164)
rs1126671	*ADH4*	G→A	(V→I)	0.07(94)	0.33(164)
rs4699733	*ADH6*	C→G	–	0.31(94)	0.23(164) *
rs1789882	*ADH1B*	A→G	(I→I)	0.45(94)	0.88(164)
rs1229984	*ADH1B*	G→A	(R→H)	0.45(94)	0.04(328)
rs698	*ADH1C*	A→G	(I→V)	0.30(94)	0.44(358)
rs1154458	*ADH7*	C→G	–	0.19(94)	0.38(394)
rs971074	*ADH7*	G→A	(R→R)	0.04(92)	0.12(410) *

Table 2.1: *ADH* allele frequency statistics for the initial 9-SNP study. Values in parentheses following the frequencies are the total number of alleles observed (2N). The rs1789882 mutation was subsequently removed from the analysis due to suspected genotyping error. Amino acid changes are described using the single letter symbols from IUPAC (Dixon et al., 1984). Intronic variants, or variants outside of a gene transcript, have no associated amino acid changes.
* All calculated probabilities were below 0.01 except at rs4699733 ($p = 0.18$) and rs971074 ($p = 0.56$).

2.4.2 Linkage Disequilibrium Within the *ADH* Region

Table 2.2 shows results from the analysis that was carried out on the *ADH* gene markers in the Maori population, using the program *Haploview*. The degree of disequilibrium between pairs of SNPs (as D') were calculated, together with an estimate of the level of confidence (p values) which can be ascribed to that score.

Regions of relatively high LD were observed within the *ADH* region: one between rs13832 and rs1126671, and another between rs4699733 and rs971074. In each of these regions, there were two markers that were in complete LD with each other ($D' = 1$, $p < 0.01$). Haplotype frequencies within these regions can be found in Table 2.3 and Table 2.4 respectively. Alleles with the highest frequency in the NCBI RefSeq database are marked with '*'. Probabilities calculated are based on χ^2 (1 d.f.), comparing each haplotype to all other haplotypes between the Maori and European populations.

Gene		1	2	3	4	5	6	7	8
ADH5	**1**	–	0.76	0.67	0.36	1	0.59	0.29	0.39
ADH4	**2**	< 0.01	–	1	1	1	0.26	0.01	0.17
	3	< 0.01	< 0.01	–	0.07	1	0.34	0.12	0.19
ADH6	**4**	0.21	0.18	0.64	–	0.91	0.48	0.65	1
ADH1B	**5**	0.03	0.31	0.11	< 0.01	–	1	0.84	1
ADH1C	**6**	0.04	0.55	0.17	< 0.01	< 0.01	–	0.54	0.39
ADH7	**7**	0.13	0.78	0.59	< 0.01	< 0.01	< 0.01	–	1
ADH7	**8**	0.12	0.35	0.37	0.05	0.04	0.54	0.74	–

Table 2.2: Calculated D' scores (above diagonal) and probability values (below diagonal, derived from a bivariate Spearman's correlation test of genotype frequencies), showing Linkage Disequilibrium within the *ADH* region. Mutations are coded as follows: 1: rs13832, 2: rs1042364, 3: rs1126671, 4: rs4699733, 5: rs1229984, 6: rs698, 7: rs1154458, 8: rs971074

2.4.3 Differences in Haplotype Frequencies

Haplotype	Maori (94)	European (164)	Probability
GG*	0.926	0.667	< 0.01
AA	0.053	0.316	< 0.01
GA	0.021	0.017	0.83
AG	0.000	0.000	1

Table 2.3: *ADH4* Haplotypes – rs1042364 (on left), rs1126671 (on right). Values in parentheses in the column headings are chromosome counts.

Large differences in haplotype frequencies are apparent at both haplotype block regions identified in the Maori population (Tables 2.3 and 2.4). Over 92% of the Maori individuals that were typed had the 'GG' haplotype in the *ADH4* region, while fewer than 3% of the European individuals were

Haplotype	Maori (94)	European (310)	Probability
AA	0.447	0.023	< 0.01
GA*	0.255	0.542	< 0.01
GG	0.298	0.435	0.02
AG	0.000	0.000	1

Table 2.4: *ADH1B-ADH1C* Haplotypes – rs1229984 (on left), rs698 (on right). Values in parentheses in the column headings are chromosome counts.

observed with the 'AA' haplotype on the *ADH1B/ADH1C* region. Also in the *ADH1B/ADH1C* region, the least common haplotype in the European population is the most common haplotype in the Maori population. The frequency in Europeans for this haplotype is slightly lower than that of the rs1229984 rare allele alone, although this difference may be due to the different European populations from which the haplotype and allele frequencies were derived.

2.5 Discussion

2.5.1 Large Differences in both Allele and Haplotype Frequencies between Maori and European Populations

There were significant differences ($p < 0.01$) in allele frequencies between Maori and European populations for seven of the nine SNPs that have been typed in this study. While the underlying reasons for these differences remain unknown, it is evident from this observation, and from the associated haplotype comparisons, that the Maori and European populations do have different genetic signatures within this region. In addition to providing more genetic evidence to different alcohol responses between the populations, the differences would also be useful in distinguishing the two populations, and for understanding the genomic patterns within the Maori population. In particular, variants that are quite different in frequency between populations can be used for confirmation of the ancestry of an individual in situations where that may be uncertain. Individual variants can be attributed to any population, but more confidence can be placed in an abundance of variants with a high prevalence in a particular population. It is hoped that studies such as this will be able to be used to determine the degree to which genetic differences are related to self-reported ancestry, and then be used to work out how useful they can be for inference of ancestry

in the absence of prior genealogical information.

The rare variant of the mutation rs1229984 has been widely associated with protection against alcoholism (e.g. Higuchi et al., 2004; Osier et al., 2002; Chen et al., 1997). This mutation resides within a haplotype block in the Maori population, together with at least one other SNP, rs698. Due to the close proximity of these two mutations, it would be expected that the typing of one of these two mutations could be used to predict the result of typing the other mutation. However, as seen in Table 2.4, the two most common haplotypes in Maori have the same rs698 allele (A), while the two most common European haplotypes have the same rs1229984 allele (G). The minor allele frequencies for rs698 and rs1229984 differ in Maori by 15% and in Europeans by 40% (Table 2.1), further demonstrating that the typing of one of these SNPs cannot be used as a proxy for the other SNP.

Another SNP, rs1693482, has been previously determined to be in complete LD ($D' = 0.99$, $R^2 = 0.96$) with rs698 for individuals who participated in the Framingham Heart Study in Massachusetts, USA (Djoussé et al., 2005). This is a potential candidate for another nearby SNP that might be in complete LD with rs1229984, but due to overlapping blocks (see Wall and Pritchard, 2003), linkage is not necessarily inherited from nearby SNPs. In order to establish conclusively how far the primary haplotype block of rs1229984 extends, a more detailed analysis of the surrounding region would be required.

The D' values give an indication of *linkage disequilibrium* (LD), the degree to which an allele at one location is able to predict the allele at another nearby location. The D' value can also be used, together with LOD scores, as an indicator of how well haplotypes within the region between two SNPs are preserved. High D' values indicate that ancestral haplotypes have been carried down through many generations without recombination, keeping the haplotype signature from the ancestral population intact. A D' value of 1 suggests that no recombination has occurred in the observed population.

However, there is always a possibility that there *has* been a recombination event, but subsequent mutation has reverted the particular variant under test back to what it was before the recombination happened. For this reason, it is important to consider nearby variants when making inferences about the degree to which recombination has happened within a region.

It is interesting that the two haplotype blocks identified in Maori within the *ADH* gene region have different frequency profiles both when compared to each other, and when compared to the same haplotype block region in the European population. The *ADH4* haplotype (Table 2.3) is typed as GG for over 90% of Maori individuals, and the next most common haplotype in both Maori and European populations is a mutation at *both* locations, from GG to AA. For the *ADH1B-ADH1C* haplotype, the second most common haplotype in Maori (GG) is again two mutations different from the most common haplotype (AA), whereas the most common haplotype in European (GA) and the second most common haplotype (GG) differ by only one mutation. It is possible that there may be a functional advantage – with regards to alcohol metabolism – in the inheritance of specific variants within these regions as a unit, and this advantage results in the putative haplotype block patterns observed in these regions.

2.5.2 Block Sizes Consistent With Other Populations

Previous studies of haplotype block sizes in the human genome have indicated that a block distance of less than 10-30kb would be consistent with that observed in other populations (see Introduction, Section 1.3.2). The two fully-linked regions that have been identified lie within this range: rs1042364 and 1126671 have a marker separation distance of around 3kb, and rs1229984 and rs698, have a separation of 11kb. However, it was expected that the Maori population might have larger block sizes than the European population because the Maori population is more recently established (around 600-800 years ago), and it is therefore necessary for more

detailed studies of genetic structure surrounding the *ADH* region in Maori to determine the extent of blocks observed here.

2.6 Extensions and Future Work

The regions of high LD that Edenberg et al. (2006) found in Europeans appear to correspond with the regions of high LD that have been observed here in the Maori population. The present results indicate that Maori and European populations differ greatly across these regions at both an allele and haplotype frequency level, which suggests that these differences will still be apparent when looking at detailed linkage patterns within the region. No such study of these patterns has been carried out, but it would be an obvious next-step to enhance the understanding of Maori genetic variation near the ADH genes.

2.7 Conclusion

This study has contributed to understanding of the population genetic structure of the *ADH* genes in Maori, and has demonstrated that this structure differs in Maori and European populations. It is anticipated that this improved understanding will aid future researchers in clarifying the link between alcohol dependence and the *ADH* genes.

Of the nine SNPs typed in the current study, seven were also included in the set of 110 SNPs typed in the study by Edenberg et al. (2006), and the two SNP sets span a similar region on chromosome 4. Of these SNPs, rs1042364 (located in the $3'$ UTR of the *ADH4* gene) was identified by Edenberg et al. (2006) as being significantly associated with alcohol dependence.

In this study marked differences were observed in the allele frequency between Maori and European groups at six SNPs spanning the *ADH* gene

region. Very different haplotype signatures have been identified at the alcohol-metabolising genes in Maori compared with Europeans. A region of apparent high LD including the well-known *ADH1B* variant was identified in Maori, which is perhaps indicative of a large haplotype block of Polynesian origin.

These findings probably reflect the unique genetic history of the Maori population and provide important information for designing association studies of the *ADH* genomic region in alcohol-related traits in Polynesians.

Chapter 3

Sequence Variation at the Monoamine Oxidase A Gene Region in the Maori Population

3.1 Overview

In 2002, Gilad et al. reported evidence for positive selection within the human monoamine oxidase A (*MAOA*) gene region on chromosome X in 7 populations: Pygmy, Aboriginal Taiwanese, Chinese, Japanese, Mexican, and Russian. They did not genotype individuals from Maori or Polynesian populations, but repeated founder effects due to the migration history of Polynesia suggest that selective signals may also be evident in the Maori population. This chapter investigates *MAOA* gene variation in Maori in order to determine if a signal of positive selection may be present within this gene region.

The *MAOA* gene has been identified as a candidate for influencing susceptibility to alcoholism and other impulse control disorders (see Ibanez et al., 2000; Morell, 1993). Variation in *MAOA* genotypes have been im-

plicated in both alcoholism (Hsu et al., 1996) and smoking behaviour (Jin et al., 2006) in Asian populations. A study of genetic variation within the *MAOA* gene region should help to describe the range of genetic variation in the Maori population that may influence alcoholism susceptibility. Also, the characterisation of *MAOA* sequence variation in Maori should aid in the design of future genetic association studies for alcohol and drug dependence.

Data from this chapter, in part, have been abstracted in poster form at the International Congress of Human Genetics, poster #1329 (Lea et al., 2006), and published in The New Zealand Medical Journal (Lea and Chambers, 2007a).

3.2 Background

3.2.1 Biochemistry of Monoamine Oxidase A

Monoamine oxidases (MAO, EC 1.4.3.4) are flavoenzymes that are bound to the outer mitochondrial membrane. This protein class has two known members, each with a different substrate specificity. Monoamine oxidase B (*MAOB*) has low activity towards serotonin but high activity towards dopamine in human platelets (Glover et al., 1977), while monoamine oxidase A (*MAOA*) preferentially oxidises serotonin in human liver (Grimsby et al., 1996). In SH-SY5Y cultured neuroblastoma cells, the predominant monoamine oxidase is *MAOA*, and it appears to influence apoptotic pathways (Fitzgerald et al., 2007). The two proteins are identical at 70% of their amino acid positions (see Shih et al., 1999), and the transposition of substrate binding domains between *MAOA* and *MAOB* (residues 161-375 for *MAOA*, residues 152-366 for *MAOB*) causes each protein to acquire substrate binding affinities similar to the other form of monoamine oxidase

(Grimsby et al., 1996). For reviews of the literature on *MAOA* biochemistry, see Shih et al. (1999) and Nagatsu (2004).

3.2.1.1 X-chromosome Inactivation

The *MAOA* gene resides on the X chromosome, so care must be taken in the interpretation of association study results in females. Inactivation of one of the two X chromosomes during early development results in a mosaic of phenotypes across all the cells of the body. Although some X chromosome genes can escape inactivation, *MAOA* has been found to be monoallelic in skin fibroblasts (Nordquist and Oreland, 2006), so one copy is likely to be inactivated.

Females who are homozygous for a particular genetic variant will have that same variant in all cells, regardless of which chromosome is inactivated, so can usually be considered to have a similar genetic profile to hemizygous males (although statistical tests should be carried out first in order to confirm that male and female data can be combined). However, the nature of the X inactivation process can mean that it is difficult to know which variant may be active in the particular cells of interest for a heterozygous female. Due to these mosaic effects, it is probably best to remove X-chromosome data for heterozygotes from association studies, because inclusion of such data could generate false results.

A removal of heterozygous data is not necessary for investigations of genetic patterns and chromosomal recombination, as in the current study. These types of analyses aim to determine the *history* of a variant, rather than its current effect. Mosaic effects from X-chromosome inactivation in females will not be expected to affect the outcome of the study, because this study just counts alleles.

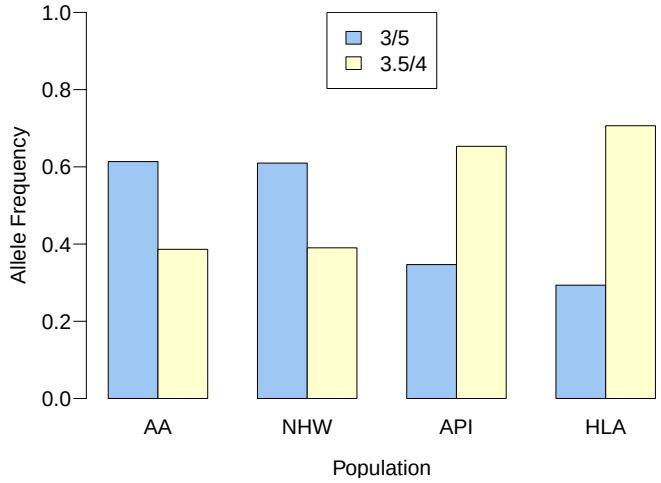

Figure 3.1: Bar plot demonstrating a wide range of *MAOA-uVNTR* frequencies in 4 different population groups (AA – African American; NHW – Non-Hispanic / White; API – Asian / Polynesian; HLA – Hispanic / Latino). Frequencies of the high expression haplotypes (3.5-repeat and 4-repeat) vary from 0.39 in the African American group to 0.71 in the Hispanic / Latino group. Data used to generate this figure are from Sabol et al. (1998).

3.2.2 Sequence Variation in the *MAOA* gene

The two known MAO proteins are encoded by two separate (but closely linked) genes, monoamine oxidase A (*MAOA*) and monoamine oxidase B (*MAOB*). They are found almost tail-to-tail (20kb separation) on the X chromosome (region Xp11.3). In the promoter sequence 1.2kb upstream of the *MAOA* gene, there is a variable-number tandem repeat polymorphism (VNTR) of a 30bp repeat sequence element (*MAOA-uVNTR*, see Figure 3.2) that is known to influence the expression levels of the gene (Sabol et al., 1998), and occurs at different frequencies in different populations (see Figure 3.1). The 3 and 5-repeat variants result in low expression, while 3.5 and 4-repeat variants result in high expression of the gene. This promoter polymorphism has been found to be in LD with genetic markers within the *MAOA* gene (Ibanez et al., 2000).

Gilad et al. (2002) carried out a study of nucleotide diversity at the *MAOA* gene region in humans and discovered extensive variation across seven different population groups: Ashkenazi, Pygmy, Aboriginal Taiwanese, Chinese, Japanese, Mexican, and Russian. A direct sequencing approach was used in order to ensure a high resolution analysis of total nucleotide variation across the region. A total of five segments of the gene were selected with total combined length of 18.8kb, i.e. about 20% of the entire 90kb gene region. Exonic portions of the gene were preferred when choosing regions to sequence, but intronic sequences were used where no exonic sequence existed within a particular region. Overlapping 1kb regions were sequenced in males to provide full haplotype information for each of the five segments.[†] A total of 41 polymorphic sites were observed: 33 Single Nucleotide Polymorphisms (SNPs), 7 deletions, and the *MAOA-uVNTR*. The polymorphic status of mutations was not consistent across all populations. Only 12 of 41 sites were polymorphic in all genotyped populations (i.e. 29 mutations were non-polymorphic in at least one population).

3.2.3 The Case for Selection at the *MAOA* gene

Based on their observations of linkage disequilibrium (LD) patterns across the *MAOA* gene region, and reduced diversity within populations, Gilad et al. (2002) suggest that there may have been positive selection in this region acting on *MAOA*-related phenotypes. The most obvious evidence for selection that was found within this region was consistently higher LD than that expected under a neutral recombination model throughout the region. Gilad et al. (2002) supported their evidence for positive selection at this gene locus with an observation of low within-population diversity combined with high between-population diversity.

[†]Males only have one X chromosome, so mutations can be trivially combined into a haplotype

3.2.3.1 Statistical Tests for Neutrality

Neutral theory predicts that DNA changes at an approximately constant rate over time for sites that are selectively equivalent (i.e. the survival chance for different variants is the same), and that this applies to the majority of the genome (see Kimura, 1991). The constant rate of change also means that between-group variation will increase in proportion to within-group variation, and variation within groups should be proportional to the rate of evolution. Because regions of the genome with functional importance can acquire mutations that are deleterious (and selected against), Kimura reasons that those functional regions will evolve slower than non-functional regions with no selective pressure. This view is supported by Aguadé et al. (1989) and Charlesworth et al. (1993), who found that selection against deleterious mutants resulted in a reduction of variability in regions near the selected locus, and this effect is more pronounced when recombination rates are low.

The index of nucleotide diversity (π) measures the average number of differences per site between randomly chosen DNA sequences (Nei and Li, 1979). It is an estimator of $\theta = 4N\mu$, the population mutation parameter. At the time of writing their initial paper, Nei and Li knew that this quantity varied between populations even within the same species, so it is important to establish a baseline diversity before using the statistic to determine deviation from normality. Gilad et al. (2002) calculated π for all populations that were typed, and use it as an estimator for mutation frequency for the calculation of recombination rate. They reported that nucleotide diversity is similar (about 0.05% per base-pair) in all typed populations, and that this value is similar to average values reported previously for X chromosome sequences.

Tajima (1989b) has proposed a statistic, D (more commonly known as Tajima's D), which under neutral mutation conditions fits a beta distribution with a mean of approximately 0, and a variance of approximately 1. It is

assumed that mutations that are tested using this statistic are sampled at random from the population. Values of D that deviate from 0 to a large degree reject a neutral mutation hypothesis, although it is possible that recent bottleneck effects can result in a large negative D value (see Tajima, 1989a). The statistic can also be influenced by hitchhiking.

Hitchhiking is a process that can give clues to whether a genetic variant has been under selection. Neutral genetic variants appear at a higher than expected frequency when posited near a selected variant due to low recombination rates near the selected region. The degree of hitchhiking increases when selection rates are high, and decreases when recombination rates are high. Fay and Wu (2000) discuss a statistical test (the H test) that is used to test for departure from neutrality in the presence of hitchhiking. This statistic compares two estimators of the mutation rate, θ_π (based on average heterozygosity) and θ_H (based on homozygosity). The H test is defined as the difference between these two estimators, with negative values indicating that a hitchhiking event has occurred.

While Tajima's D (Tajima, 1989b) compares low frequency and intermediate frequency variants to determine if hitchhiking has occurred, the H test compares the high frequency and intermediate frequency variants to find signals for hitchhiking. Fay and Wu (2000) state that only demographic models and positive selection can explain an excess of high-frequency variants. Gilad et al. (2002) found H values more negative than those expected under a neutral selection model for four of the seven genotyped populations, and also for the combination of all seven populations, giving a strong indication that variation within the *MAOA* gene region has been influenced by positive selection. The Taiwanese population was the only population that Gilad et al. (2002) found to have significant ($p < 0.05$) deviation from neutrality for Tajima's D, and this population was one of the three that did not show evidence for hitchhiking using the H test.

3.2.4 The Addition of Maori *MAOA* Data

The study carried out by Gilad et al. (2002) did not include Polynesian individuals, so an obvious extension for this project is to investigate genetic selection in this region in a Polynesian population. The current study has extended this work to include a Polynesian population (Maori), with a goal of investigating whether there is also evidence of selection within this region for the Maori population.

3.3 Methods

This study has an intra-population and inter-population design, looking at haplotype block patterns in the neighbourhood of the *MAOA* gene region. Haplotypes were scored in males and females using banked DNA samples from the Maori population described in Chapter 2. The principal aim is to investigate evidence for positive selection within this region in the Maori population.

3.3.1 Variants Typed

All references to locations of genes and mutations in this chapter are based on the NCBI reference assembly for *Homo sapiens*, build 36, March 21, 2006. Gene names are the same as those in the NCBI database as of May 2007.

The genomic sequence for the *MAOA* gene, including 10kb of flanking sequence (both upstream and downstream), was also retrieved from the NCBI database. Flanking sequences for *MAOA-uVNTR* and 13 of the SNPs typed by Gilad et al. (2002) were requested from the authors (see Figure 3.2), and located within the retrieved *MAOA* genomic sequence (see Table 3.1 for mutations and flanking sequences).

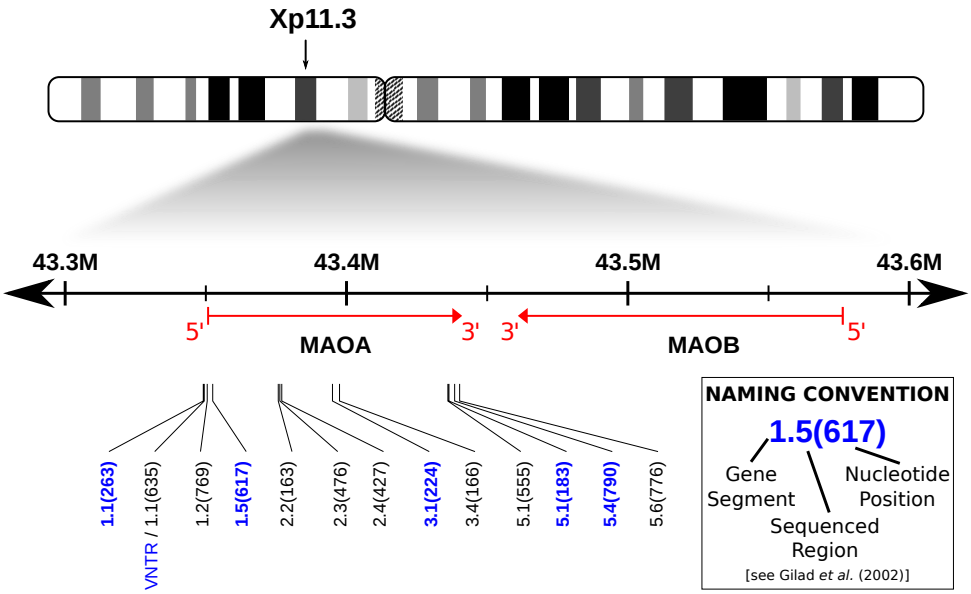

Figure 3.2: A section of chromosome X, showing the relative positions of *MAOA* and *MAOB* genes and mutations typed in this study. Mutations that were polymorphic in the Maori population are coloured blue. The labels for each mutation refer to the segment, region, and nucleotide position from Gilad et al. (2002).

All 13 of these SNP positions were confirmed as present within the *MAOA* gene by using the *ssearch* program (Smith and Waterman, 1981).[†] To determine the precise location of SNPs within the X chromosome (according to the NCBI reference sequence), flanking sequences for the mutations were used as input for the web-based *BLASTn* program.[‡] After carrying out this process, it was noticed that the first SNP, 1.1(635) resides *within* the 30-bp repeat sequence. Hits within chromosome X were found for SNPs except 2.3(476). A search for 2.3(476) alone against mutations and flanking sequences on chromosome X produced over 100 results with greater than 92% homology.

It is likely that no results were returned initially for this SNP for the *BLASTn* search on the entire genome because too many matches were

[†] `http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml`
[‡] `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

Mutation	$5'$ **– Sequence –** $3'$
1.1(263)	AGAGCCCACGCGGCTACACCCAC **R** TCTACTCCCCCACTCTC
1.1(635)	ACCAGTACCGGCACC **R** GCACCAGTACCCGCACCAGTACCGG
VNTR	ACAGCCTGACCGTGGAGAAG ... TCCGAATGGAGCGTCCGTTC
– [repeat]	[(ACCGGC)(ACCGGC)(ACCAGT)(ACCCGC)(ACCAGT)]$_{3-5}$
1.2(769)	CAAAAGGGTTCGCCCCGC **S** CACAGTGCCCGGCTCCCCCCGG
1.5(617)	AGTGATCTACAACCATA **M** TGCTTTTAGGAGGCTTGCCTAGT
2.2(163)	GAGTTGCTGAGAAGCAGGTTTTT **Y** AGCATGGAGATAAAGAA
2.3(476)	AGAATTGCTTGAACCC **R** GGAGGCGGAGCTTGCAGTGAGCCG
2.4(427)	GTGTAGGCTATGCATAG **Y** CTTTTACAGTATGTTAAGATGGG
3.1(224)	TTACATGGATCATT **Y** AACAAAAATAATATATAGCCAGCAAT
3.4(166)	TATCACAGTGTCTGGG **R** GGATGTGGCCCTGCCCCCTACTAC
5.1(183)	GACAACTATTTCTAGAATTTGCA **Y** TGAACTCTGCTTTTCCT
5.1(555)	GTATACTTTGCTCTT **M** CCATTTTCTTGATTAGGGAAGACAT
5.4(790)	TCAGGTTCTTGTACCCAGAT **R** TCTTTCTCGGTCACCTTCCC
5.6(776)	ACACCAGGGTCCAGCA **M** CTTAGGTTTGAATTTATGATAAGG

Table 3.1: The mutations that were typed for the *MAOA* gene region study. Start and end primer sequences for the VNTR are displayed in this table, as well as the repeated sequence. The references here indicate the segment name and sequence position within each region from the Gilad et al. (2002) study. Mutation codes are as follows: M – A/C; R – A/G; S – C/G; Y – C/T.

found for this search. Some of these matches may be duplicates, as flanking sequences for adjacent mutations in NCBI can overlap. However when restricting the search to only the X chromosome, at least 20 different DNA sequences were observed (with only one inside the *MAOA* gene region), suggesting that at least 20 different partial matches for this sequence are on chromosome X alone.

3.3.2 Study Population

The 13 SNPs, plus the 30bp-repeat VNTR, were typed by in 47 unrelated Maori individuals with a self-reported ancestry of four Maori grandparents (in fact the same 47 individuals that were typed for the alcohol dehydrogenase (*ADH*) study in Chapter 2). Genotyping was done via a service contract

with the Australian Genome Research Facility (AGRF)[†], which used the Sequenom MassArray Genotyping system (Buetow et al., 2001).

Genotype data were analysed using *Haploview* (Barrett et al., 2005) to establish whether particular combinations of SNPs could be treated as a unit (i.e. a haplotype block). Female haplotype counts were initially calculated using Haploview and verified manually by a maximum-parsimony method: if it were possible that a previously observed haplotype could generate the observed unphased genotype data, then that haplotype was recorded as present in that individual (the same method used by Clark, 1990). The significance of haplotype and VNTR frequency differences between Maori and other populations was determined using the `chisq.test` function of R (R Development Core Team, 2008), treating the total Gilad et al. (2002) data as an expected probability, and simulating the χ^2 distribution when genotype counts for either Maori or all Gilad populations were below 5.

Neutrality tests carried out by Gilad et al. (2002) were repeated for the set of 11 SNPs genotyped in Maori, together with data from Maori males, using the program *VariScan* version 2.0.2 (Hutter et al., 2006).[‡]

3.4 Results

3.4.1 Polymorphism in the Maori Population

Of the 13 SNPs typed (all polymorphic in at least one of the seven populations that Gilad et al. (2002) had genotyped), two were unable to be genotyped, 1.1(635) and 2.3(476)[§], and five (of the remaining 11) were found to also be polymorphic in the Maori population (1.1(263), 1.5(617), 3.1(224),

[†] `http://www.agrf.org.au`

[‡] `http://www.ub.es/softevol/variscan/`

[§] This was most likely due to the repeat locus and abundance issues described previously in Section 3.3.1

Locus	Typed	Variants		Polymorphic	
		Common	Rare	All[#]	Maori
1.1(263)	Y	A	G	Y	Y
1.1(635)	N	G	A	N	-
uVNTR	Y	3[*]	4	Y	Y
1.2(769)	Y	C	G	N	N
1.5(617)	Y	C	T	Y	Y
2.2(163)	Y	T	C	N	N
2.3(476)	N	A	G	Y	-
2.4(427)	Y	C	T	N	N
3.1(224)	Y	C	T	Y	Y
3.4(166)	Y	A	G	N	N
5.1(183)	Y	C	T	Y	Y
5.1(555)	Y	C	A	N	N
5.4(790)	Y	G	A	Y	Y
5.6(776)	Y	A	C	N	N

Table 3.2: Genotyping overview, indicating the polymorphic status of mutations typed in the Maori population. Two mutations, 1.1(635) and 2.3(476), were unable to be genotyped by AGRF.
The polymorphic status represents whether a mutation was polymorphic in *all* populations typed by Gilad et al. (2002), and whether a mutation was polymorphic in the Maori population.
***** The common variant globally for *MAOA-uVNTR* is a sequence repeated 4 times, whereas the common variant in Maori is a sequence repeated 3 times (globally rare). For all other mutations, the common variant in Maori was consistent with the common variant in other populations.

5.1(183), and 5.4(790); see Table 3.2 and Figure 3.3). The *MAOA-uVNTR* was also polymorphic in the Maori population. The remaining six SNPs were all non-polymorphic in every Maori individual tested, with all individuals having the most common variant (with respect to other populations) at each site. Of the 11 SNPs that were genotyped, mutations that were polymorphic in the Maori population were also polymorphic in all populations typed by Gilad et al. (2002), and only polymorphic in Maori in this globally-polymorphic case.

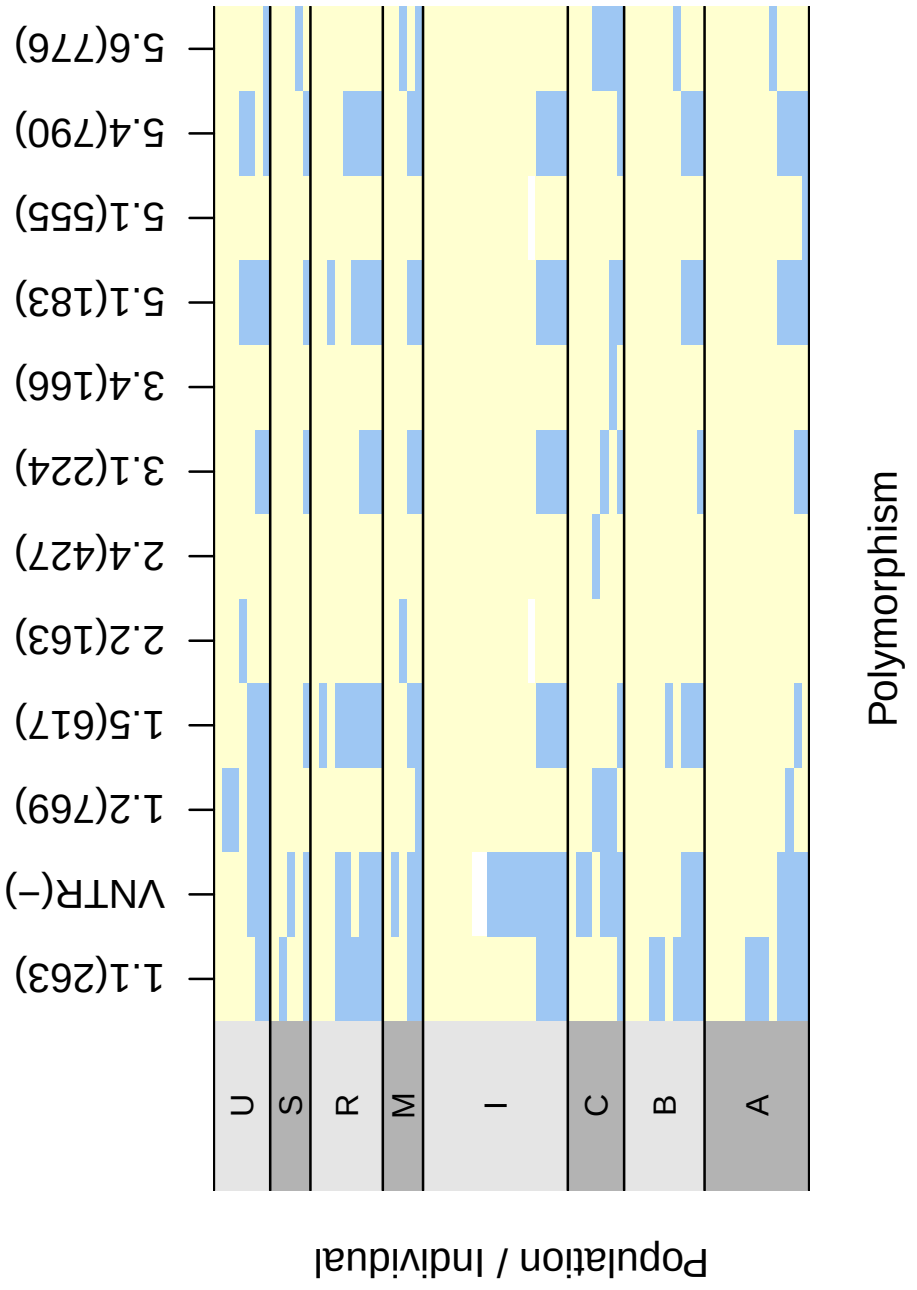

Figure 3.3: Matrix of SNP and VNTR genotypes (56 males) from Gilad et al. (2002), only including SNPs genotyped in Maori, together with genotype data from 18 Maori males. The first column indicates the population, subsequent columns indicate genotypes. Common variants are shown as light yellow, rare variants are light blue. Failed genotypes appear white in this figure. Population letter codes (from bottom to top) are as follows: A – Ashkenazi; B – Bedouin; C – Pygmy; I – Maori; M – Mexican; R – Taiwanese; S – Asian; U – Russian.

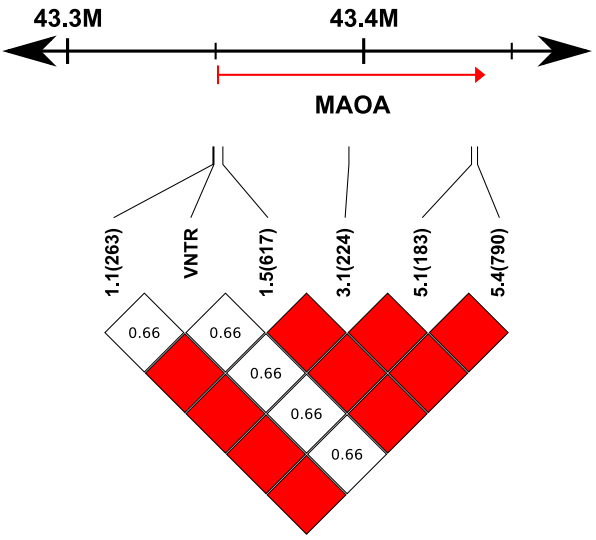

Figure 3.4: Haploview diagram showing full LD in the Maori population (18 males, 29 females) across the *MAOA* gene. Red diamonds indicate full LD between two SNPs. Numbers inside squares represent D', D' is 1 where not specified. The VNTR was treated as a dimorphism and included in this Haploview analysis.

3.4.2 Full LD Among All SNPs

Genotype data from Maori females were analysed using *Haploview* (see Figure 3.4), which demonstrates complete LD between all SNPs ($D' = 1$), but incomplete LD between all SNPs and the VNTR ($D' = 0.66$). Only three SNP haplotypes were observed within the *MAOA* region in Maori females (AGCCG, GATTA, AGTTA), out of a possible 32 that would be expected for five unlinked dimorphic SNPs.

3.4.3 Haplotype Counts for the *MAOA* Gene

Given that LD was found to be complete for all 5 polymorphic SNPs across the entire *MAOA* gene region, these SNPs were combined into 5-SNP haplotype blocks, but treated separately from the *MAOA-uVNTR* (as LD in both Maori males and Maori females was not complete between SNP and VNTR

variants, see Figure 3.4). Table 3.3 summarises the count data recorded for these polymorphisms in all genotyped individuals, including data from the Gilad et al. (2002) study for the five SNPs that were polymorphic in the Maori population. Counts for heterozygous 3/4 females appear separately in the table (7 females had a heterozygous VNTR genotype that could not be assigned with certainty to a particular chromosome).

Only two *MAOA* 5-SNP haplotypes were observed in the 18 Maori males that were genotyped: 14 AGCCG haplotypes and 4 GATTA haplotypes. These two haplotypes are the most common globally, and differ at all five sites. Because males have only one X chromosome, these haplotypes can be inferred directly from the genotypes. Among the 29 females that were genotyped across the *MAOA* region, three different haplotypes were observed: 45 AGCCG, 12 GATTA, and 1 AGTTA haplotype. These haplotype counts do not differ significantly from haplotype counts for Maori males ($\chi^2 = 0.33, p > 0.8$). There were 10 females who were heterozygous across this gene region with both the AGCCG and GATTA haplotypes, and another female was found to be heterozygous with an AGTTA haplotype combined with the AGCCG haplotype.

The 3-repeat VNTR variant was observed in 10 Maori males and in 9 Maori females (a total of 11 3-repeat VNTR variants were counted in females, see Table 3.3). The 4-repeat VNTR variant was observed in 6 males and 20 females (33 4-repeat VNTR variants counted in females). The proportion of haplotypes with the 3-repeat VNTR is much higher in males (0.62) than in females ($0.09, \chi^2 = 12, p < 0.003$).

In all cases where a 3/4 heterozygote VNTR genotype was observed, both the rare and common SNP haplotypes were also observed (AGCCG and GATTA respectively). No male was observed with the GATTA haplotype combined with the 4-repeat VNTR, but one female was observed with two copies of the GATTA haplotype and was also homozygous for the 4-repeat VNTR.

Maori Male VNTR Proportions

5 SNP Haplotype	3	4	NG	n	pHT
AGCCG	0.50	0.50	2	14	0.78
GATTA	1.00	0.00	0	4	0.22
Total	0.62 (10)	0.38 (6)	2	18	1.00

Non-Maori Male VNTR Proportions

5 SNP Haplotype	3	4	NG	n	pHT
AGCCG	0.17	0.83	0	23	0.41
GATTA	1.00	0.00	0	10	0.18
GGCCG	0.00	1.00	0	7	0.12
GACTA	0.67	0.33	0	3	0.05
GGCTA	1.00	0.00	0	2	0.04
AGCTG	0.50	0.50	0	2	0.04
AACCG	0.00	1.00	0	2	0.04
AGCTA	0.00	1.00	0	1	0.02
GACCG	1.00	0.00	0	1	0.02
GACCA	1.00	0.00	0	1	0.02
AACTA	1.00	0.00	0	1	0.02
AGTCG	1.00	0.00	0	1	0.02
GGTTA	1.00	0.00	0	1	0.02
GATTG	1.00	0.00	0	1	0.02
Total	0.45 (25)	0.55 (31)	0	56	1.00

Maori Female VNTR Proportions

5 SNP Haplotype	3	3/4	4	NG	n	pHT
AGCCG	0.14	0.00	0.86	3	17	0.59
AGCCG/GATTA	0.00	1.00	0.00	3	10	0.34
AGCCG/AGTTA	-	-	-	1	1	0.03
GATTA	0.00	0.00	1.00	0	1	0.03
Total	0.09 (2)	0.32 (7)	0.59 (13)	7	29	1.00

Table 3.3: Proportions of VNTR variants for SNP haplotypes observed in the *MAOA* gene region for 18 Maori males, 56 non-Maori males (from Gilad et al. (2002)), and 29 Maori females. Column heading abbreviations have been used to conserve space: NG – number of individuals who could not be genotyped for any VNTR variant; n – total number of individuals with this SNP haplotype; pHT – proportion of this SNP haplotype in the population.

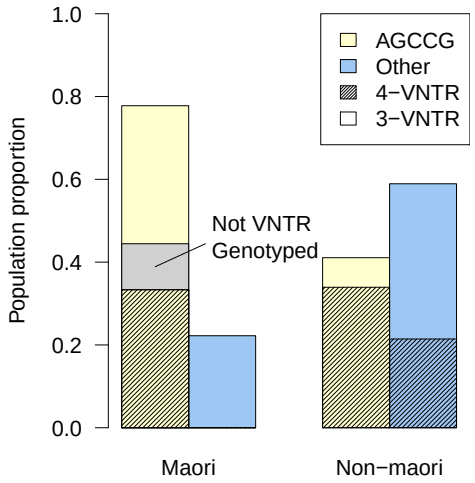

Figure 3.5: Comparison of haplotype frequencies between Maori males and all other populations in Gilad et al. (2002). Shaded lines indicate the frequency of the common (4-repeat) VNTR allele. The grey area on the Maori AGCCG bar indicates individuals who were unable to be typed at the VNTR locus.

3.4.4 *MAOA* Region Haplotype Frequency Comparisons between Maori and non-Maori Populations

Haplotypes were initially dichotomised into common (AGCCG) and not common (any other haplotype) in an attempt to better compare Maori and other human populations (see Figure 3.5). The proportion of the most common SNP haplotype (AGCCG) is greater in Maori males (0.78) than in non-Maori males (0.41, $\chi^2 = 10.02, p = 0.0015$). Of those individuals who have the higher frequency AGCCG haplotype, a smaller proportion of Maori (0.50) have the 4-repeat VNTR allele than non-Maori populations (0.83, $\chi^2 = 10.36, p < 0.008$).

Figure 3.6 shows a mutation network diagram for 5-SNP *MAOA* haplotypes that were found. The figure can be broken into two parts comprising mutations likely to be related to each of the two most common haplotypes, AGCCG and GATTA, separated by dotted lines. While GGCCG was the third most common haplotype observed by Gilad et al. (2002), only two

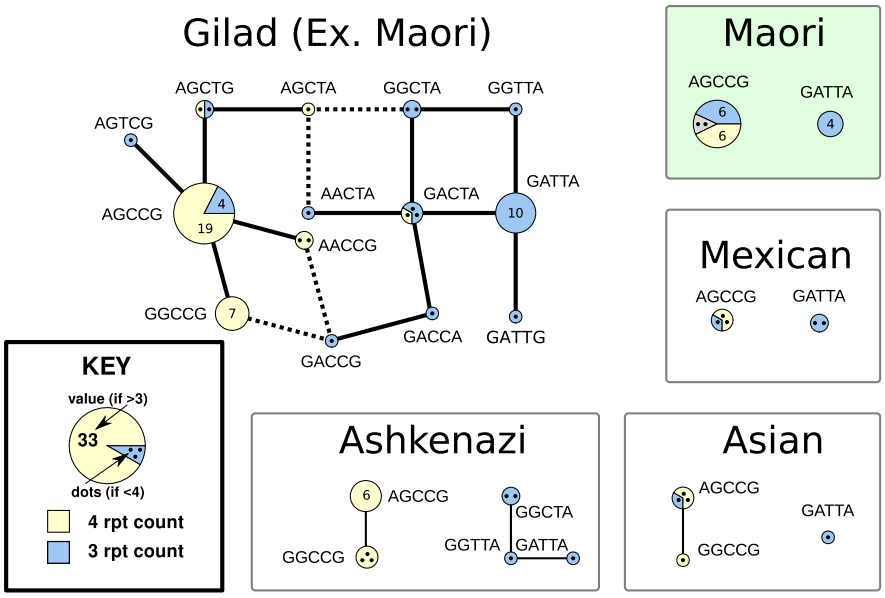

Figure 3.6: Haplotype mutation network diagram for the *MAOA* gene region, showing observed 5-SNP *MAOA* haplotypes in male individuals for all Gilad et al. (2002) populations combined, as well as Maori, Mexican, Asian, and Ashkenazi populations. The SNP haplotypes are indicated by labels, and VNTR variants are indicated by coloured segments; the common 4-repeat VNTR variant is indicated in blue, the rare 3-repeat variant is shown as yellow, and not-genotyped is shown as green. The area of the circles is proportional to the number of haplotypes found (also indicated by numbers, or by dots for fewer than four samples). Lines indicate all single mutations that would change from one observed haplotype to another. Dotted lines require both a SNP and a VNTR mutation to be consistent with observed haplotypes.

haplotypes were found that differed from it at a single locus, AGCCG (the most common haplotype) and GACCG. The common 4-repeat VNTR polymorphism was not found at all in conjunction with the lower frequency GATTA SNP haplotype in any population. Only one GATTA-related haplotype, GACTA, was found in conjunction with the common 4-repeat VNTR allele. However, the 3-repeat allele was found in conjunction with three AGCCG-related haplotypes (AGCCG, AGTCG, and AGCTG).

The frequency difference for the common haplotype between Maori and non-Maori populations is similar when combining counts for single muta-

tion variants of the most common haplotype (GGCCG, AACCG, AGTCG, AGCTG). When including these variants, the proportion of non-Maori populations with common variants is 0.625 ($\chi^2 = 1.7926, p = 0.18$). The AGTTA haplotype that was found in a single female chromosome (see Table 3.3) was not present in any of the haplotypes reported by Gilad et al. (2002).

3.4.5 Re-analysis of Neutrality Tests

Neutrality tests carried out by Gilad et al. (2002) were repeated in the Maori male population, restricting analysis to the 11 SNPs genotyped in Maori (see Table 3.4). Sequence data for the entire gene region were unavailable for Maori, so statistics that require full sequence data were excluded, since they could not be calculated. The sensitivity of statistics to sequence changes was determined by embedding short stretches of non-polymorphic pseudo-sequence between polymorphic sites. Statistics were excluded when the calculated statistic with pseudo-sequence added differed from that when pseudo-sequence was absent. Fu and Li's D statistic (Fu and Li, 1993) has also been excluded because although it was present in their population variability parameter table, the statistic was not discussed by Gilad et al. (2002). The r^2 statistic has been included due to advice received that D' and r^2 should be reported together (See introduction, Section 1.3.1.1).

Most statistics calculated for Maori males lie within the ranges observed for the male populations genotyped by Gilad et al. (2002). However, there are two clear outliers: the number of distinct haplotypes (K), and r^2. Assuming a normal distribution for these statistics in the non-Maori populations, the number of distinct haplotypes in Maori is $2.96SD$ from the mean ($p = 0.0015$), and the r^2 statistic is $5.1SD$ from the mean ($p \ll 0.001$).

Population	N	S	K	D'	r^2	Tajima's D
Ashkenazi	13	8	7	1	0.187	-0.329
Bedouin	10	6	6	0.958	0.262	0.544
Pygmy	7	9	6	0.893	0.224	0.444
Taiwan	9	5	7	0.805	0.315	1.520
Asian	5	6	4	1	0.528	-0.668
Mexican	5	8	4	0.851	0.488	1.028
Russian	7	8	6	0.854	0.255	0.722
Total Gilad	56	11	23	0.762	0.105	0.630
Maori	18	5	2	1	1	0.820

Table 3.4: Neutrality tests of the *MAOA* gene region in the Gilad populations and Maori population for 11 SNPs genotyped in Maori. N is the number of individuals genotyped, S is the number of polymorphic sites, and K is the number of observed distinct haplotypes. Some statistics used by Gilad et al. (2002) were sensitive to sequence data (θ, π, H test), and have been excluded from this table due to the absence of *MAOA* sequence data for Maori. The D' and r^2 statistics in this table are mean linkage values across the entire gene region.

3.5 Discussion

This study has identified three distinct haplotypes in Maori males and Maori females across a 90kb region encompassing the *MAOA* gene (see Section 3.4.3). Haplotypes for males were determined directly from genotype data for the X chromosome; the low number of observed haplotypes in Maori allowed haplotypes to also be determined for females, through a simple process of elimination.

The results from the analysis of the *MAOA* region (18 males, 29 females) in Maori show that there is less genetic variation across the *MAOA* gene region in the Maori population than in other populations (see Figure 3.6. The individuals genotyped in this study were not known to be closely related to each other, but the Maori population in general has reduced genetic diversity when compared to other populations (Hall, 2004; Shepherd et al., 2004) – a characteristic that is predicted when considering the migration history of Maori (see Whittle, 2010).

This study has found that a relatively high proportion of Maori individuals have the uncommon 3-repeat VNTR, when compared with the populations genotyped by Gilad et al. (2002). There are two possible ways in which this variation could have been introduced, either through recombination, or through mutation. If a recombination event is assumed, the most common repeat polymorphism would be expected to be inherited in tandem with the most common *MAOA* haplotype. Such a tandem inheritance is inconsistent with the data presented here. It is more likely that there has been a recent mutation within *MAOA-uVNTR* (i.e. a change from a 3-repeat uVNTR variant to a 4-repeat variant) in the ancestral Maori population.

3.5.1 Statistical Tests for Neutrality

Of the statistical tests that were compared between populations typed by Gilad et al. (2002) and the Maori male population, two are clear outliers in Maori, r^2 and K (see Table 3.4. Other statistics lie within the range of what has been observed in non-Maori populations, but this should not be considered evidence against selection, as selection within this region has already been demonstrated in non-Maori populations.

The value of Tajima's D does not deviate much from 0 in the Maori population (0.82), and is of a smaller magnitude than Tajima's D for the Taiwanese and Mexican populations (see Table 3.4). Therefore the null hypothesis of neutral mutation is not rejected when considering this statistic for the Maori population. This is not particularly surprising, given that the calculated value of Tajima's D only exceeded Gilad et al. (2002)'s threshold of significance ($p < 0.05$) in one population that they typed.

Kimura (1991) suggests that the most prevalent form of natural selection is stabilising selection, and it is interesting to observe that the two haplotypes observed in Maori males are also the two most common haplotypes found in non-Maori males. Tajima (1989a) reports that under an infinite

allele neutral mutation model, founder effects followed by a fast recovery of population size can have a strong influence on the average number of pairwise differences between chromosomes sampled from the population (i.e. π), but little effect on the number of segregating sites (i.e. S). In contrast, current population size has a strong influence on S, but not on π. The ancestors of the Maori population probably experienced repeated founder effects during migration through Polynesia (see Introduction, Section 1.5.2), so it is therefore not surprising that the number of unique haplotypes in Maori is low (resulting in lower π), while the number of segregating sites is within the range of what is observed in non-Maori populations.

It is unfortunate that the H test (Fay and Wu, 2000) is sensitive to sequence data (see Section 3.4.5), because this statistic is prominent in the evidence presented by Gilad et al. (2002) for selection within the *MAOA* gene region. However, given that hitchhiking has occurred within the *MAOA* gene in a number of different populations as demonstrated by H test results from Gilad et al. (2002), it is reasonable to evaluate if the available data suggest that hitchhiking has taken place within this gene region. Only two SNP haplotypes were found in Maori males, AGCCG and GATTA. These two haplotypes differ at all five SNP loci, so it does not make sense to consider that the GATTA haplotype is present in an increased frequency due to selection for the AGCCG haplotype. In this sense, there does not appear to be evidence for deviation from neutrality through a hitchhiking event, as might be found by an increase in intermediate-frequency genetic variants. There has, in fact, been a removal of variation, even though the presence of haplotypes with low frequencies in other populations demonstrates that these low frequency variants are not likely to be deleterious.

3.5.1.1 Association between SNP Haplotypes and Number of VNTR Repeats

However, the frequency of the *MAOA-uVNTR* variant in Maori differs from that observed in non-Maori populations (see Section 3.4.4): there is a relatively low frequency of the 4-repeat VNTR allele in Maori, corresponding to a high frequency of the 3-repeat VNTR allele. If the 3-repeat VNTR allele is considered to be selectively equivalent to the 4-repeat allele, this may indicate that the more common AGCCG is under positive selective pressure. This indicator alone is not enough to be strong evidence for positive selection, but is enough to warrant further investigation at a sequence level of the *MAOA* gene in Maori.

Seemingly contrary to this indicator of selection, the frequency of the 5-SNP *MAOA* haplotypes found in the Maori population do not differ significantly from those in other populations if single-mutation variants of the common SNP haplotype are included in counts of the common haplotype (see Section 3.4.4 and Figure 3.6). It is reasonable to consider these extra haplotypes (GGCCG, AACCG, AGTCG, AGCTG) together with the common haplotype (AGCCG), because it would be expected that the haplotype with the largest frequency in a well-established population is older and therefore likely to accrue additional mutations over time. One of these minor variants, GGCCG, seems to be an outlier in that it is observed at a fairly high frequency in three populations typed by Gilad et al. (2002), namely Ashkenazi, Bedouin, and Asian. In fact, GGCCG is the third most frequent haplotype beyond the two core haplotypes, differing from AGCCG at the first SNP in the haplotype, 1.1(263). Hence another variant has been observed at higher than expected frequency on the background of the most common SNP haplotype, AGCCG.

These two polymorphisms, *MAOA-uVNTR* and 1.1(263), are separated by around 400 base pairs (see Figure 3.2), and are both in the promoter region of the *MAOA* gene (about 1.5kb from the $5'$ end of the *MAOA*

gene itself). The remainder of the SNPs that are polymorphic in Maori reside wholly within the *MAOA* gene. As mentioned previously, only demographic models and positive selection can explain an excess of high-frequency variants (Fay and Wu, 2000). The interpretation of these results appears to indicate that a specific variant of the *MAOA* gene, namely that described by an XGCCG SNP haplotype, is under positive selection both in Maori (as evidenced by an increased frequency of the 3-repeat VNTR allele) and non-Maori (as evidenced by an increased frequency of the GGCCG haplotype). The location of hitchhiking variants suggests that this selection is on the gene variant, rather than the expression of the variant.

3.5.2 Large *MAOA* Haplotype Block Found in the Maori Population

The *Haploview* analysis suggests that the 30bp VNTR is not linked to either of the two major SNP haplotypes in Maori (see Figure 3.4). However, the 4-repeat VNTR was never observed with the GATTA SNP haplotype in males, and only observed with the GATTA haplotype in females that had both GATTA and AGCCG SNP haplotypes (see Table 3.3). It is consistent with the observed male data and female homozygote data to assume that the 3-repeat VNTR resides on the same haplotype as the GATTA SNP haplotype in heterozygous females. When this assumption is considered, the GATTA SNP haplotype always predicts a 3-repeat *MAOA-uVNTR* variant, and LD therefore extends across the entire *MAOA* gene region (*including* the promoter region VNTR).

The *Haploview* analysis indicates that all SNPs within the 90kb *MAOA* gene region were tightly linked in Maori males and Maori females (see Figure 3.4). This complete linkage is a feature of the Maori population that was not observed in any of the populations that Gilad et al. (2002) typed (see Table 3.4). The most similar population found (with respect to linkage

across the entire region) was a combined Chinese and Japanese population ($D' = 1$, $r^2 = 0.528$, $n = 5$).

Gabriel et al. (2002) carried out a study that identified haplotype blocks in 51 autosomal regions covering 13Mb of sequence dispersed throughout the human genome, with an average size of 250kb per region. Their study identified blocks in African-American genomes of up to 94kb in length, (mean length of 9kb), and up to 173kb in length (mean length of 18kb) in European and Asian genomes. Considering the SNPs within the *MAOA* region, the block length of 90kb identified here (Figure 3.4) lies within these ranges, but is at the higher end.

Recombination in the X chromosome occurs less frequently than in the autosomes, so a 90kb haplotype block may not necessarily represent a "large" unit on this chromosome. Although Gilad et al. (2002) have mentioned that the neighbourhood of the *MAOA* region experiences high recombination (up to $4.58cMMb^{-1}$), they also observed no decay of LD across the 90kb region identified here. The results presented here are similar with respect to LD decay (or lack of it), but differ in that a greater proportion (i.e. 100%) of typed SNPs show complete LD in the Maori population.

It is established from the observation of LD across the *MAOA* region (Figure 3.4) that the haplotype block that *MAOA* resides on is *at least* 90kb, but in reality it is probably longer and genotyping of the region surrounding the *MAOA* gene would be required to establish the full extent of this block. In fact, an unpublished study has found that this block spans a much greater distance in a Maori population, possibly almost 1Mb (Rod Lea, Personal communication, 2008).

Across the entire 90kb region, only two SNP haplotypes were identified in Maori males (out of 32 possible haplotypes), where 14 different haplotypes were observed in non-Maori males (see Figure 3.6). The two *MAOA* haplotypes found in Maori are also the most common haplotypes found in non-Maori populations, but differ in frequency between Maori

and non-Maori populations.

Only one possible recombinant haplotype was observed (AGTTA, see Table 3.3), so the entire SNP variation across *MAOA* can be described by genotyping just two SNPs, i.e. 1.1(263) and 3.1(224). The VNTR polymorphism was not found to be a good predictor of SNP haplotypes in Maori (i.e. it is not tightly linked to particular haplotypes), as both the 3-repeat VNTR and 4-repeat VNTR variants were found together with the most common SNP haplotype (AGCCG). Including the VNTR would require 3 genotyping assays for the entire 90kb *MAOA* region, which would be a suitable proxy for sequencing the entire gene in Maori individuals.

3.6 Concluding Remarks

It is clear that the Maori population shows minimal variation within the *MAOA* gene region, most likely because the population is very new (650-750 years before present) in terms of the global history of human migration. Lack of recombination within this region – only one instance of recombination was found in 47 individuals – indicates that some positive selection has occurred within the Maori population. A reanalysis of male haplotype data from Gilad et al. (2002), combined with an analysis of Maori haplotypes indicates that the gene variant described by an XGCCG haplotype is under positive selection.

Two clear limitations of the study carried out here are the number of individuals genotyped (18 males, 29 females), and the gene coverage (i.e. 90kb). The Maori population described here was only genotyped across a small section of the *MAOA* gene region, and it would be helpful to analyse genotype data across a larger region to confirm that selection is only on variants of the *MAOA* gene itself. The number of genotyped Maori individuals was chosen to be comparable in size to the study carried out by Gilad et al. (2002), and as an individual population, exceeds the number

of individuals typed for all other populations at this region (see Table 3.4). Given the observation of substantial variation in other populations, variation beyond the two core haplotypes should have been seen in Maori if it existed in substantial proportions of the Maori population. However, in light of the reduced cost of genotyping now available through SNPchip technology (see Introduction, section 1.4.2.1), both these issues would be easily overcome in a future study of this gene region in Maori.

A review of literature on the metabolic role of *MAOA* has revealed a surprising scarcity of primary research, despite considerable effort to look for recent articles; it seems that few recent studies have demonstrated that primary substrates of *MAOA* in humans actually include serotonin *per se*. Such a study is needed, particularly in light of the results from the present study that suggest a particular gene variant (and hence probably expressed protein) has been selected for in human populations.

Many recent review articles[†] that talk about *MAOA* preferentially acting on serotonin do not reference this statement, treating it as an accepted fact, rather than a hypothesis (e.g. Gokturk et al., 2008; Jacob et al., 2005). Some articles do cite studies in mammals other than humans (e.g. Guo et al., 2008), and follow a (usually implicit) transitive argument: because *MAOA* preferentially acts on serotonin in other animals, it should act similarly in humans. This argument may be incorrect, given that the human and rat *MAOA* have different structures, different polymerisation states, and different catalytic profiles (Son et al., 2008). Lewis et al. (2007) advises caution when extrapolating studies of monoamine oxidases from animal models to humans. Other articles cite papers that only refer to interactions in the introduction section, with the research outcome of the cited paper different than what is expected from the citation (e.g. Bach et al., 1988, cited

[†]i.e. articles that contain no original research

in addition to other references in Sabol et al., 1998).[†]

Papers that do cite original research for the interaction between *MAOA* and serotonin in humans are difficult to find, possibly because research on substrates for the monoamine oxidases was mostly carried out in the late 1970s and early 1980s. The scarcity of current primary research suggests that there is a need for new studies to confirm preferential action in humans, even though it may have been observed in other closely related mammals.

3.7 Controversy

During the 2006 International Congress of Human Genetics (ICHG, see Appendix C), preliminary research on the *MAOA* gene in Maori carried out by the author and collaborators was presented in poster form (Lea et al., 2006). The reaction to this research made the author more reluctant to publish further details of the investigations of this gene region subsequent to presentation at ICHG, but most of these details have been presented in this chapter. The media response to the author's *MAOA* gene research (see Appendix B) has emphasised the need to be extremely cautious when reporting results, particularly because any misinterpretation of prior research can be passed on to other researchers and the general public. Because of the continuing controversy about genetic research on this region, it was decided to cease further investigation on the *MAOA* gene for this thesis.

The arrival of SNPchip data and subsequent data-crunching (see Chapter 4) provided another reason to delay further investigations of the Monoamine Oxidase A gene until a later date. The next chapter explores Maori genetic structure at a genomic scale, to determine if large genotype frequency differences observed between Maori and European populations in

[†]Bach et al. (1988) compared human liver cDNA sequence of *MAOA* and *MAOB*, but do not discuss how sequence difference alters substrate specificities. They already knew that the two proteins were distinct molecules, and their key findings are the cDNA sequence differences themselves, not what these differences mean in terms of enzyme substrates.

the *ADH* and *MAOA* gene regions are representative of genotype differences across the entire genome.

Chapter 4

Quantification of Genomic Ancestry in a Maori Population

4.1 Overview

This chapter describes the genomic ancestry of a Maori tribe, Ngati Rakaipaaka of Nuhaka. This profiling has only been possible within the last few years, due to a combination of new technologies, the formation of the Rakaipaaka Health and Ancestry Study, and informed participation from almost all members of the tribe.

A set of genetic markers for distinguishing between Maori and European genomic ancestry was discovered in two small groups of non-admixed Maori and European individuals, and then validated in two independent (and larger) groups of non-admixed Maori and European individuals. The selected markers were then typed in almost the entire adult Rakaipaaka population (including admixed and non-admixed Maori and European individuals), providing a good estimate of the distribution of Maori genomic ancestry and European admixture within Rakaipaaka.

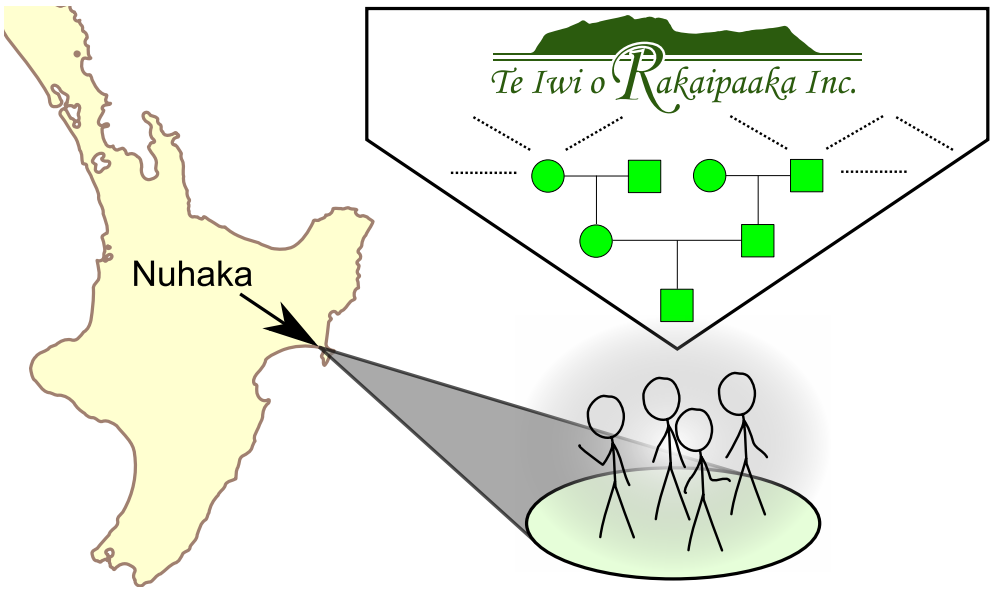

Figure 4.1: Nuhaka – the home of Te Iwi o Rakaipaaka

4.2 Background

4.2.1 Rakaipaaka and Nuhaka

Ngati Rakaipaaka are descendants of the Maori chief Rakaipaaka, and have a rohe (home area) of Nuhaka, New Zealand (see Figure 4.1. Rakaipaaka was expelled from a pa (Maori village) on the banks of the Waipaoa River (north of Gisborne), and fled to a mountain in Nuhaka, where he and his family set up a pa and governed the neighbouring district. This mountain, now called Momoukai (literally "waste food"), was the site where a Maori musket army had been repelled around 200 years later by Ngati Rakaipaaka, partly due to the attackers running out of food – Moumoukai had good access to a fresh water spring, plenty of food and storage, and possibly also an underground route to the coast for fishing (Walker, 2008).

The genesis of Ngati Rakaipaaka, formerly considered to be people of Ngati Kahungunu, began in the mid 1980s with the then Labour gov-

ernment's promotion of iwi (tribe) development. In 1996, descendants of Rakaipaaka decided to carve their own identity by establishing themselves as a new tribe and organisation, Te Iwi o Rakaipaaka Incorporated (TIORI), created as a vehicle to promote cultural and tribal development (Johnny Whaanga, TIORI Day, 2008).[†] They now have a regular yearly event (Rakaipaaka Day) to reaffirm their identity and pride as descendants of Rakaipaaka.

Nuhaka is a small town, with a general store, fire brigade, garage, and local school. There are six Marae (Maori meeting houses) in Nuhaka: Kahungunu, Manutai, Tamakahu, Te Rehu, Taane, and Kotahitanga, each comprising different whanau (extended family units) and genealogical backgrounds (Eva Paea, TIORI Day, 2008). Census data from 2006 indicates that Nuhaka has a population of around 300 people, with a mean income of around $15,000.[‡] Similar to most other iwi groups, a fair proportion of Rakaipaaka do not live in the traditional rohe, but have connections back to Nuhaka. There are about two thousand people registered with TIORI, and it is estimated that about eight thousand Rakaipaaka members live in New Zealand, most in Auckland, Hamilton, Wellington and Napier / Hastings (Johnny Whaanga, TIORI Day, 2008).

4.2.1.1 Rakaipaaka Health and Ancestry Study

Different whanau (extended families) in Nuhaka have high rates of some common diseases (including heart disease, cancer and diabetes), and questions amongst families created an interest in why these high rates existed. This led to an interest in health development, particularly the genetic and environmental contributions to health and wellbeing within the community. Rakaipaaka Discussions with Dr. Rod Lea and the Institute of Environmental Science and Research (ESR) about a health and ancestry project indicated

[†]Personal communication: TIORI Day, 11 December 2008, Wellington, New Zealand

[‡]`http://www.stats.govt.nz/Census/2006CensusHomePage/QuickStats/AboutAPlace/SnapShot.aspx?id=3545303`

a need for the community to better understand their health status and use that knowledge to provide future benefits (Johnny Whaanga, TIORI Day, 2008).

These discussions led to the creation of the Rakaipaaka Health and Ancestry Study (RHAS), launched at Rakaipaaka day in 2005. The study was set up in order to answer questions about why particular ailments were common among Rakaipaaka descendants, and to determine ways in which the wellbeing of individuals can be improved to protect against current and future health issues within the community. The study is Iwi-governed, focusing on identifying social, environmental, and genetic determinants of health. About 300 adult participants have enrolled in RHAS, filling out a questionnaire about heath-related traits, and donating blood for biochemistry and DNA analysis.

Part of the outcome of the initial discussions was an acknowledgement that both lifestyle and ancestral history contribute towards the health of a population, and the ancestral component of this suggested a need to have a better understanding of the genetic background of Rakaipaaka (Whaanga, 2008). The beginning of RHAS coincided with the start of this PhD research project, and my interest in genetic structure has blended well with a study of genomic ancestry in the Rakaipaaka population.

An understanding of population genetic structure is important for disease studies because many diseases have an underlying genetic basis (see Introduction, Section 1.5.5), and genetic structure at disease-associated loci can vary in different populations (see Chapters 2 and 3). In some cases where there is a substantial difference in disease frequency in different populations, it is possible to derive false genetic disease associations when population ancestry is not taken into account (see Pritchard and Donnelly, 2001).

4.2.2 Genetic History

The ancestral history of Maori represents a unique opportunity for genomic research, due to the low genetic diversity of Maori coupled with relatively recent (< 500 years) admixture with Northern Europeans.

Studies carried out at Victoria University and elsewhere (see Chapter 1, Section 1.5) have recreated the genetic history of Polynesians through sequencing of regions of mitochondrial DNA (e.g. Whyte et al., 2005). Other studies on Polynesian populations have been carried out on DNA from Y-chromosomes (e.g. Kayser et al., 2006b), a structure which has similar properties to mtDNA in its single line of derived ancestry (see Introduction, Section 1.2.3).

4.2.2.1 Uni-parental Ancestry

Recent Maori-European admixture is difficult to determine using DNA data that are derived from a single line of ancestry (i.e. maternal and paternal lineages). Low mutation rates and a lack of recombination reduce the variation of this DNA as it is passed down from generation to generation, so genetic differences between two people of similar origins may be difficult to determine. Also, the single line of ancestry excludes a large proportion of the genealogical history of a person – even three generations back, mitochondrial and Y-chromosomal DNA can only capture (at most) genetic information from one eighth of the ancestors of an individual (see Introduction, section 1.2.3). This can introduce substantial errors when the individual's ancestors are from many different geographical locations (see Koenig et al., 2008, Chapter 10, p. 207-8). Studies of recent admixture require a *genomic* approach, capturing recent genetic variation introduced by recombination, particularly in autosomes.

4.2.2.2 Genome-wide Analyses

It has only recently been possible to carry out cost-effective *genome-wide* studies on DNA. The information on flanking DNA sequence for Single Nucleotide Polymorphisms (SNPs) can now be retrieved as a result of the human genome project (Sachidanandam et al., 2001), together with a large-scale genotyping effort by the HapMap project (International HapMap Consortium, 2007). The genetic information provided from these efforts has led to the development of a genome-wide microarray genotyping assay, the SNPchip (Gunderson et al., 2005). This technology enables genotyping of thousands of different genetic loci in one assay, greatly reducing the cost and effort of typing large numbers of genetic variants in a group of individuals. Such genome-wide analysis is expected to reveal a great deal of information about variance in the genome as a whole, without requiring full genome sequences for all individuals involved.

4.2.3 Genome-Wide Association Studies

Genome-wide Association Studies (GWAS) typically involve determining the degree of association between genetic markers and a heritable trait. Most often, these studies look for associations relating to susceptibility for particular diseases (e.g. Wellcome Trust Case Control Consortium, 2007; Mathew, 2008), but some have also looked at traits that are not directly associated with disease (e.g. blood lipid phenotypes in Kathiresan et al., 2007). However, replication consistency can be an issue in GWAS (see Kathiresan et al., 2007), suggesting a large proportion of false positive associations.

This study will test for genetic association with Maori genomic ancestry, a trait which has no environmental component (i.e. a heritability of 100%). Contributions towards this trait come from everywhere in the genome,

greatly reducing (and one hopes eliminating) the chance of false positive associations.

Genomic ancestry is also a useful trait for this study because of reported good agreement between self-reported genealogical ancestry and estimated genomic ancestry from genetic data (Walsh et al., 2003). Further benefits come from looking at *Maori* genomic ancestry, because the Maori gene pool is less diverse and more differentiated when compared to other outbred populations (European, Asian, African, etc.). The Maori population originated from a series of long voyages between isolated islands, with admixture happening only recently in the population between Maori and European individuals (see Chapter 1, section 1.5). This allows easy sampling of an observed phenotype with fairly high accuracy by asking participants about their recent genealogical history.

4.3 Objectives

Discussions during the implementation of the Health and Ancestry Study evoked an interest in how knowledge of genetic features might better target treatments and interventions that could impact on the health of the community. This knowledge can be enhanced by a comparison of genetic differences between individuals of European descent and individuals of Maori descent, as well as an analysis of admixture within the Rakaipaaka community. The study presented here is an attempt to investigate autosomal genetic variation in the Rakaipaaka population, using the SNPchip platform for genotyping within this population. This has been done with the following goals in mind:

1. Establish a framework for the ethical conduct of genetic studies in partnership with Te Iwi o Rakaipaaka

2. Carry out a genome-wide scan to identify markers that are informative for ancestry

3. Use these markers to estimate ancestral structure within the community

4. Establish the utility of these markers for explaining differences in disease risk

4.4 Methods

This study used a whole-genome approach to derive a suitable SNP set for the estimation of genomic ancestral fractions for most of the individuals in the Rakaipaaka community (see Figure 4.2).

A total of 30 Maori individuals with 100% reported Maori ancestry were initially compared with 90 European individuals for this study (see Figure 4.3). After removing markers that had a high rank variance in simulated subsamples from the two groups, 59 markers were selected as candidate ancestry-informative SNPs. This set was then reduced to a set of 23 SNPs with minimal loss of information. Of these SNPs, 14 were chosen to be validated using 95 more Rakaipaaka individuals (RHAS) from the Ancestry Study, and 271 European individuals (PDC) from a Parkinson's Disease study (see Fung et al., 2006).[†]

4.4.1 Genotyping

A total of 30 Individuals from Rakaipaaka with full Maori genealogical ancestry were genotyped at 317,503 SNPs, using the Illumina 317k SNPchip.[‡]

[†] `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000089.v3.p2`

[‡] BDCHP-1x10-HUMANHAP300v1-1-11219278-C

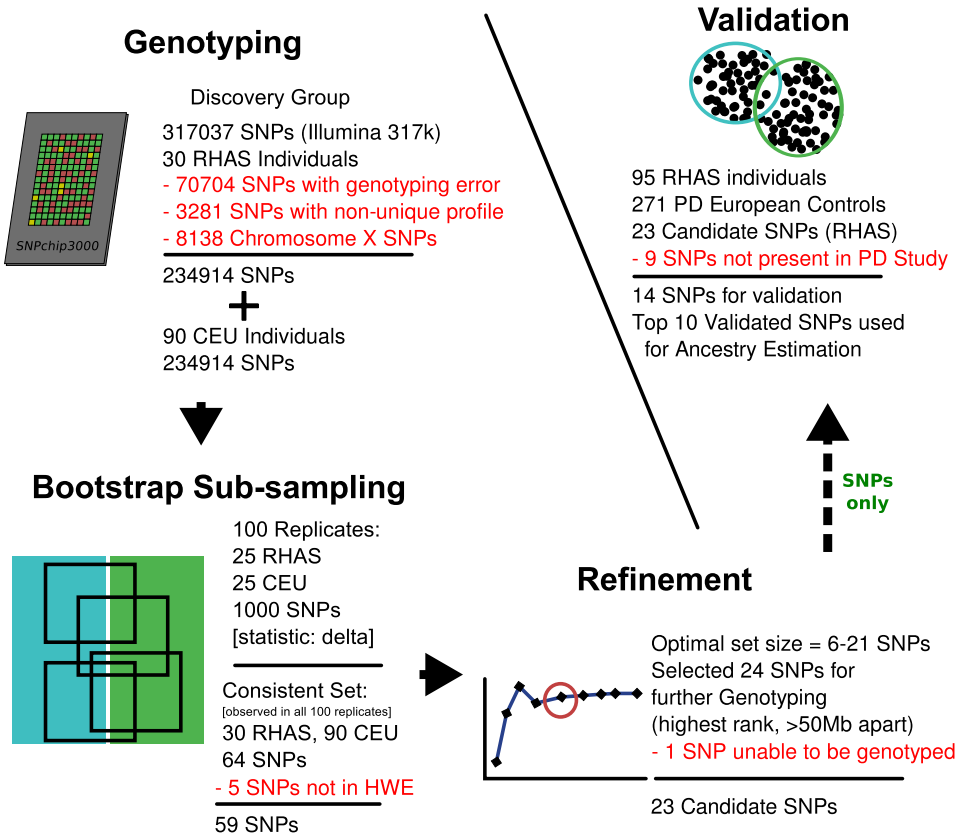

Figure 4.2: A visual summary of steps carried out during the investigation of Genomic Ancestry in the Rakaipaaka population. Candidate markers were discovered by genotyping 30 Rakaipaaka individuals (RHAS) and comparing population allele frequencies to 90 HapMap European individuals (CEU). Markers were further refined by determining the difference of mean group Q values, as calculated by *structure*, over increasing numbers of markers. Validation of markers was carried out with an additional 95 Rakaipaaka individuals, together with 271 European Individuals from a Parkinson's Disease study (PDC).

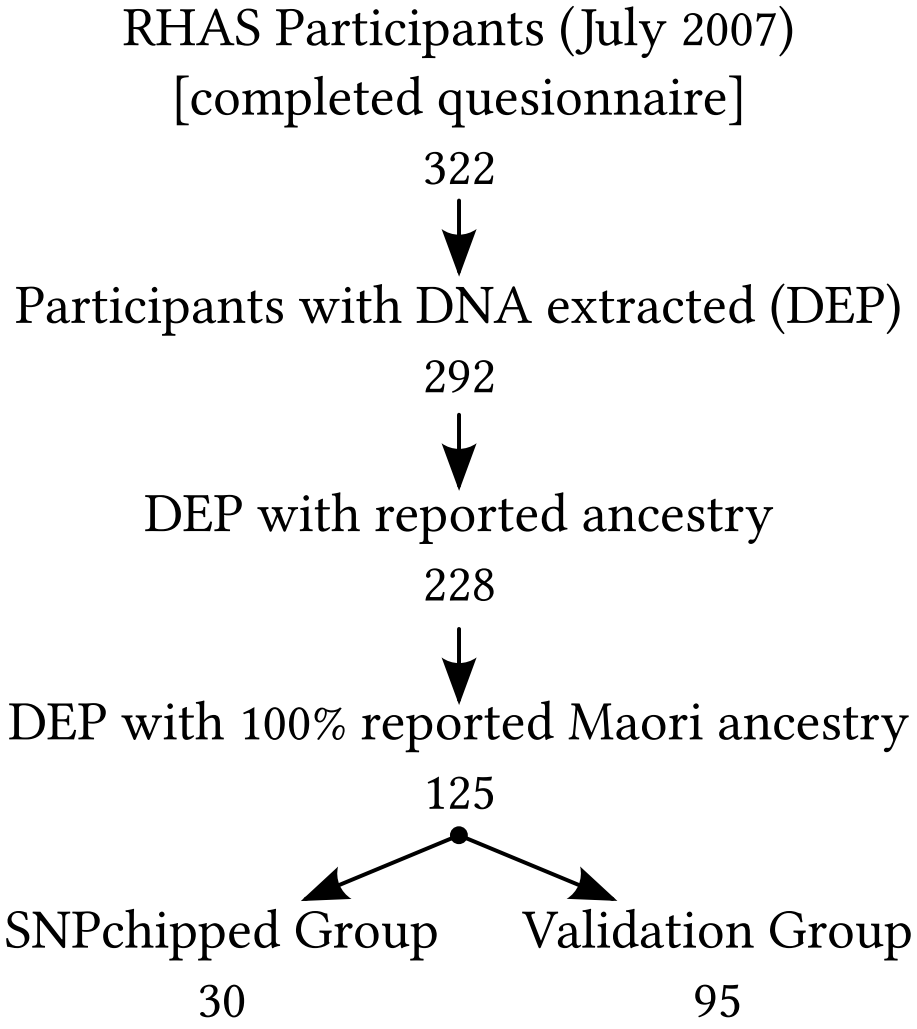

Figure 4.3: A breakdown of the participants from the genomic study of the Rakaipaaka population. The study was initially conducted with a SNPchipped group of 30 RHAS individuals, validated using the 95 remaining individuals with 100% reported Maori ancestry, and then used to generate an estimated genomic ancestry profile for all 292 participants who had DNA extracted.

These data were then combined with matching data from the HapMap project (`http://www.hapmap.org`) for the CEU population (90 individuals), and filtered to only include SNPs that had different profiles across all individuals in both populations (i.e. no two SNPs had the same reported genotype in *all* individuals). The data were filtered again to remove any SNPs that did not have genotypes for all 120 individuals, as well as removing SNPs on the X chromosome, resulting in full genotype data at 234,914 filtered SNPs (approximately 74% of the Illumina 317k SNP set).

4.4.2 Population Sub-sampling

A population sub-sampling procedure was carried out, where 100 random subgroups of 25 individuals from each of the two populations were chosen. SNPs were ranked based on the allele frequency difference between these two subgroups, and only those SNPs that were consistently in the list of 1000 top-ranked SNPs for all 100 groupings were chosen for the next stage, resulting in a consensus list of 64 SNPs. Of these SNPs, five were found to deviate from Hardy-Weinberg Equilibrium, so were removed from subsequent analysis.

4.4.3 Using *structure* to Determine SNP Set Effectiveness

The computer program *structure* (Pritchard et al., 2000) was used to determine how effective different groups of SNPs were at classifying Maori vs European ancestry. This program uses iterative methods to determine group membership ancestry coefficients (Q values) for each genotyped individual, representing how likely it is that an individual belongs to a particular group. These Q values are usually interpreted as an estimate of the proportion of genomic ancestry from that group.

This program attempts to fit the individuals to a model with a specific number of populations. The algorithm used by *structure* is an iterative

algorithm that generates the next potential solution from previous attempts at solutions. The probability distribution, or model, that this algorithm simulates is the probability of each individual being assigned to a specific cluster. Eventually, the process settles into an equilibrium from which the overall probability distribution of different states – in this case, the allocation of individuals to population clusters – can be estimated. The population clustering model is based on allele frequencies, assuming that within pseudo-populations there is no Linkage Disequilibrium (LD) and markers are in Hardy-Weinberg Equilibrium (HWE). Modifications have been made to the original program to include the possibility of weakly linked markers that may result from admixture (Falush et al., 2003).

Each *structure* run used a bootstrap of 10,000 iterations, followed by an additional 90,000 iterations. Lambda was set to 0.85, and separate alpha values were calculated for each population (initially set to 0.3). A preliminary set of *structure* runs were carried out on the set of 59 SNPs to validate a two-population hypothesis, comparing estimated probability for varying numbers of populations from 1 to 10.

In order to determine the minimum number of SNPs required to provide adequate information about Maori ancestry (i.e. a minimal SNP set), *structure* runs were carried out with increasing numbers of SNPs, using the 59-SNP set with SNPs ranked based on the allele frequency difference between the 30 RHAS and 90 CEU individuals (i.e. delta). The effectiveness of particular SNP sets was quantified by a difference of means test, calculating differences between mean Q value for each population and its associated Standard Error for 1 to 59 SNPs.

4.4.4 Validation of 10-SNP Marker Set

A further 262 Rakaipaaka individuals were then typed by AGRF for 24 of the top markers from the 59-SNP set. Including the original 30 SNPchipped

RHAS individuals, this represents about 89% of Rakaipaaka living in the Nuhaka area (of 327 reported in the 2006 census). There was a genotyping failure for one SNP, and 5 additional markers were removed due to being posited within 5MB of another marker (the markers with the lowest delta were removed, with ties broken by removing the marker with lowest F_{ST}, then by removing the marker further from the start of the chromosome according to the marker location in the NCBI database), leaving 18 markers for use in the estimation of admixture in the Maori population. These 18 SNPs were used for validation in two independent replication groups, 95 RHAS-Maori individuals with full Maori genealogical ancestry, and 271 European controls from the Parkinson's disease study mentioned in Section 4.4. Due to a different initial genotyping platform being used for the Parkinson's disease study,[†] only 14 of these SNPs were typed in both groups and therefore represent all that were available to be used for validation.

As it had been established that fewer than 14 SNPs would be sufficient to determine ancestry coefficients, the 10 SNPs with the highest frequency difference were selected as a final signature set of ancestry-informative SNPs. To estimate the accuracy of this signature set, the Q value outputs for the final 10-SNP set were compared with Q value outputs from the larger 59-SNP set. Self-reported ancestral fractions were also compared to Q values for the 10-SNP set as another estimate of accuracy.

4.5 Results

4.5.1 Distribution of Delta Throughout the Genome

Figure 4.4 shows the distribution of delta values across the genome. Large delta values are observed throughout the entire genome, with no regions

[†]Illumina Infinium 1

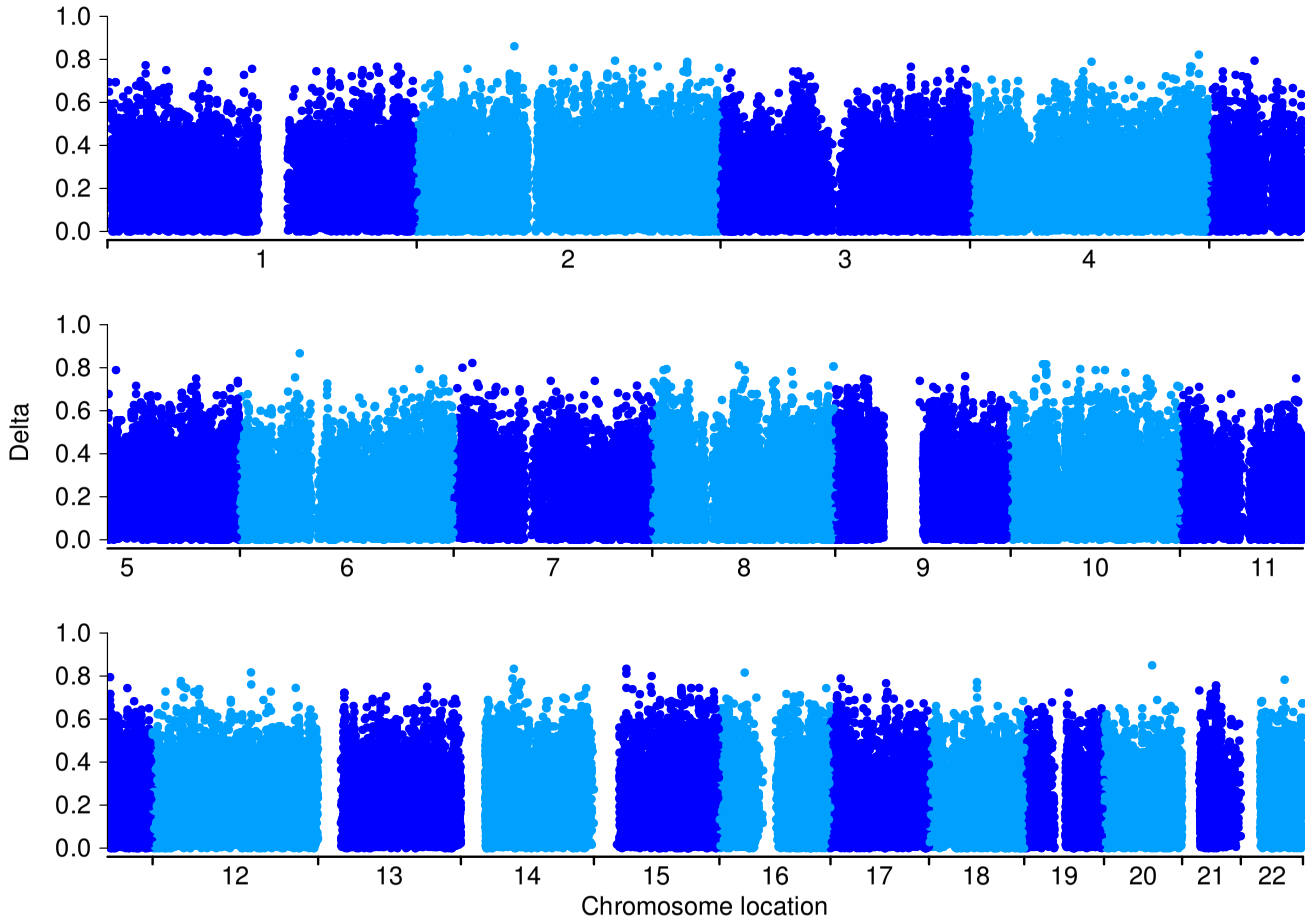

Figure 4.4: Chromosome delta plot indicating allele frequency difference (delta) between SNPchipped RHAS individuals and CEU, as well as genomic location of 234914 SNPs. Alternating colours are used here as an aid to distinguish different chromosomes. Gaps appear in all chromosomes at the centromere, where genotype data is not available.

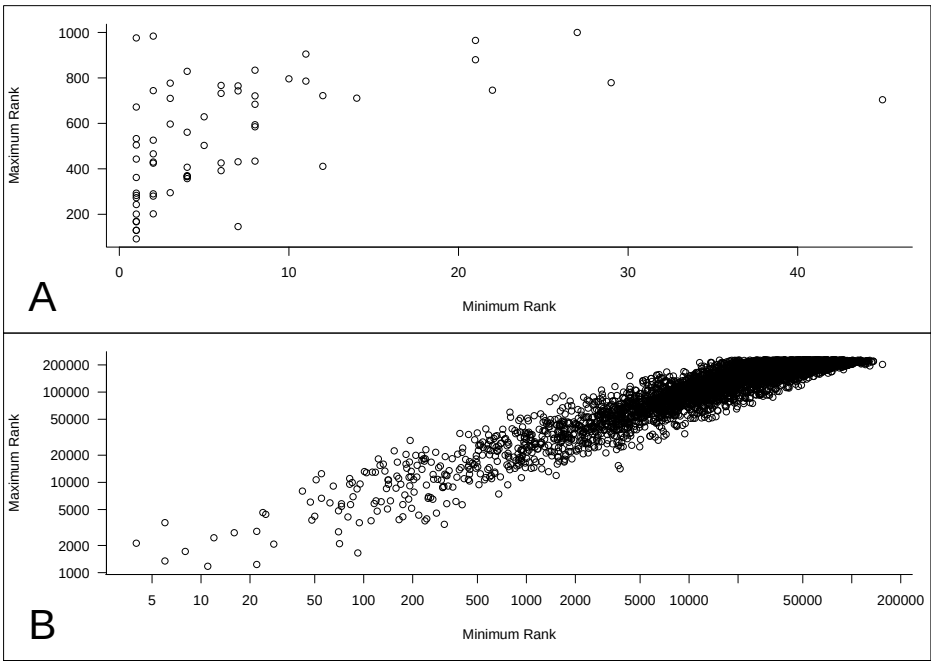

Figure 4.5: Scatter plot indicating maximum rank over all bootstrap sub-samples vs. minimum rank in any bootstrap sub-sample for the bootstrap-consistent set of 64 SNPs (A), and a random sample of 5000 of the remaining SNPs (B). A total of 105803 SNPs (not included when generating these graphs) were unranked in at least one bootstrap sub-sample, as no genetic difference was observed between RHAS and CEU populations with that SNP. The difference between minimum and maximum rank gives an indication of the reliability of a particular marker for association testing in a general population. Of those markers in this bootstrap-consistent set of 64 SNPs, 59% were ranked in the top 600 markers in all bootstraps. Of the remaining 234850 SNPs, 59% (138388) had a maximum rank of 200000 or more (including 105803 unranked SNPs).

that can be easily identified as showing particularly large differences between the two populations when compared with the general trend across the genome.

4.5.2 Population Sub-sampling

Figure 4.5 compares minimum and maximum rank of SNPs, both those 64 SNPs that were ranked in the top 1000 SNPs in all bootstrap sub-samples,

and the remaining 234850 SNPs. Each bootstrap sub-sample used a different group of 25 individuals from each population. Most of the consistent set of 64 SNPs had a maximum rank below 600, whereas most of the remaining 234850 SNPs had a maximum rank above 200000.

Five of these SNPs were excluded due to HWE probabilities of less than 5%, resulting in a HWE-filtered set of 59 SNPs (see Figure 4.6).

4.5.3 Estimation of Maori Ancestral Fraction

Figure 4.7 shows the two-population output for the HWE-filtered set of 59 SNPs. There is a greater spread of Q values among RHAS individuals ($\overline{x} = 0.92, \sigma = 0.095$) than among CEU individuals ($\overline{x} = 0.0055, \sigma = 0.014$). All RHAS individuals were within 3 SD of the mean Q values, while two outliers ($> 3SD$) were observed in the CEU population (with Q values of 0.075 and 0.11). With all individuals included, there is a difference of 0.60 between the most extreme Q values of the two populations (i.e the Q value ranges for the two populations do not overlap).

Increasing numbers of SNPs were run through *structure*, using the candidate set of 59 markers (Figure 4.8). In each run, the top n SNPs were tested together, with n ranging from 1 to 59 SNPs. Effectiveness of particular SNP sets was quantified by a difference of means test, calculating differences between mean Q for each population and associated Standard Error for each run of SNPs. The maximum mean difference of 0.95 $(SE = 0.0095)$ was observed with 6 SNPs, after which the difference drops for increasing numbers of SNPs, then reaches a local peak of 0.93 $(SE = 0.018)$ with 21 SNPs.

Figure 4.9 shows the unsorted two-population output for the set of 14 community-validated SNPs. Spread is still greater among RHAS individuals ($\overline{x} = 0.81, \sigma = 0.18$) than among Parkinson's Control (PCE) individuals ($\overline{x} = 0.017, \sigma = 0.035$), but greater than that observed in the initial test of 59

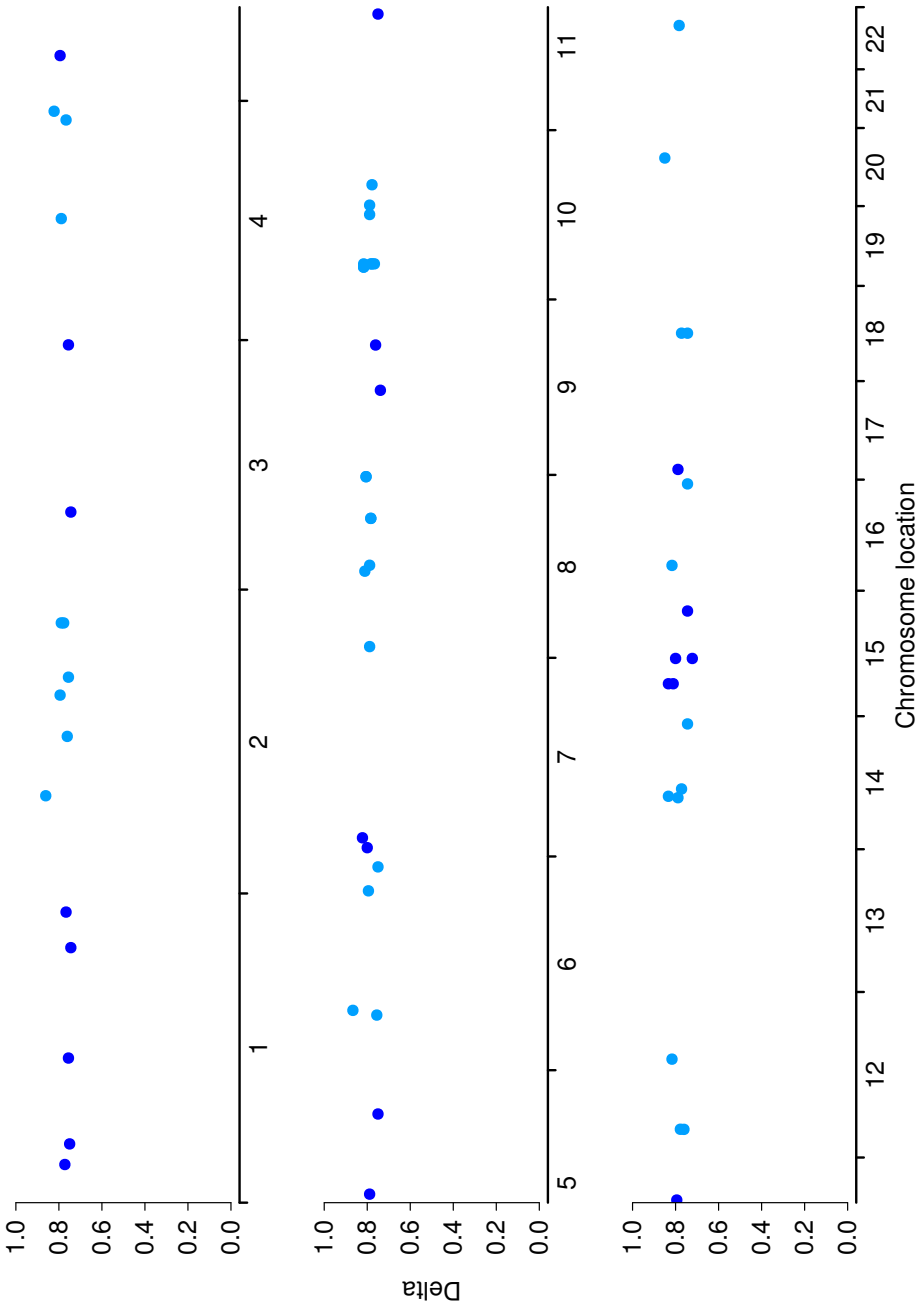

Figure 4.6: Chromosome delta plot indicating allele frequency difference (delta) between SNPchipped RHAS individuals and CEU, as well as genomic location of the consistent set of 64 SNPs. No SNPs consistently ranked in the top 1000 were found on chromosome 13, 19, or 21.

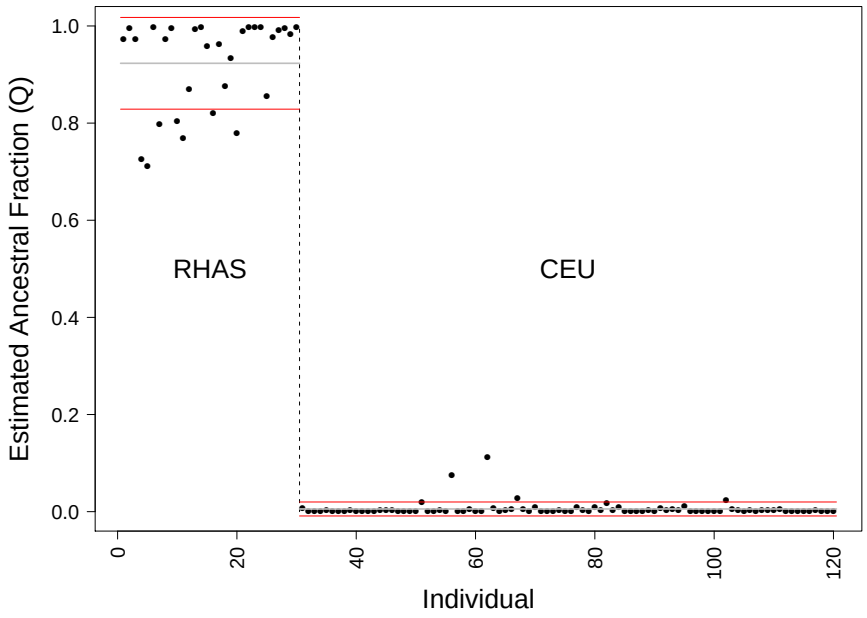

Figure 4.7: Structure output (K=2) for 59 SNPs, showing Q values for the SNPchipped RHAS individuals and the CEU population. Mean Q value (grey line) and the standard deviation of Q values (red line) are also indicated for each population.

SNPs. All RHAS individuals were within 3 SD of the mean Q values (with one individual 2.98 SD from the mean), while six outliers were observed in the CEU population (with Q values of 0.12, 0.15, 0.17, 0.20, 0.25, and 0.37). If these outliers are included, there is an overlap of 0.12 between the most extreme Q values of the two populations, otherwise the difference is 0.15 between the most extreme values.

The spread of Q values from the Rakaipaaka community genotyping study (see Figure 4.10) is still greater among RHAS individuals (mean $= 0.685$, SD $= 0.280$) than among European individuals (mean $= 0.028$, SD $= 0.040$, not shown in figure), and also greater than that observed in the initial test of 59 SNPs. All RHAS individuals were within 3 SD of the mean Q values, and eight outliers were observed in the European populations (with Q values of 0.16, 0.16, 0.18, 0.24, 0.26, 0.26, 0.30, and 0.37). Choosing the

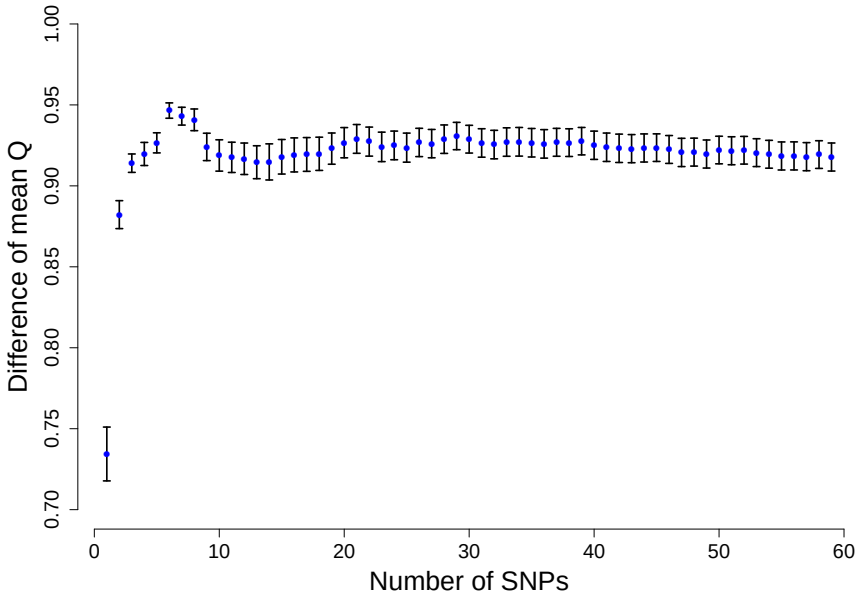

Figure 4.8: Difference of means tests for the top 1..59 SNPs comparing mean Q values between 30 SNPchipped RHAS individuals and 90 CEU individuals. The difference of mean Q values between the two populations rises steeply for the first six SNPs, then gradually drops for increasing numbers of SNPs.

maximum European Q value as a cutoff point, 86% of RHAS individuals have a Q value greater than this value. Also, 92% of RHAS individuals have a Q value greater than 3SD from the mean Q value of the European population (0.15).

45 Maori individuals have an estimated ancestral fraction greater than 0.97, which cannot be distinguished from a value of 1.0 using our test.

4.5.4 Final List of SNPs

A total of 10 SNPs (Table 4.1) have been found that have good discrimination power for comparing Rakaipaaka and European ancestry, covering nine chromosomes (two on chromosome 12). Of these 10 SNPs, six reside within genes (or hypothetical genes). The minimum delta observed

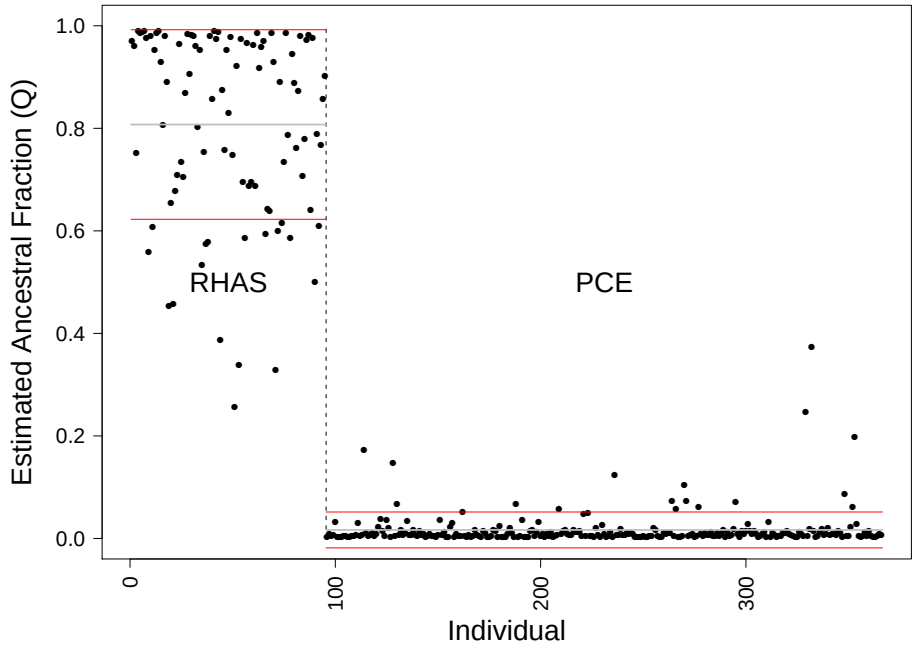

Figure 4.9: Structure output (K=2) for 14 SNPs, showing Q values for the 95 individuals from the Rakaipaaka validation group (RHAS) and the Parkinson's Control (PCE) population. Mean Q value (grey line) and the standard deviation of Q values (red line) are also indicated for each population.

Marker	Mutation	c/s	Location (Mb)	Delta	Nearest Gene
rs1160638	C/T	2	158.594	0.67	*UPP2**
rs10485317	A/G	6	47.841	0.73	*OPN5*
rs6950662	G/T	7	14.917	0.72	*DGKB*
rs6558383	C/T	8	144.845	0.72	*ZNF707**
rs7911256	C/T	10	25.897	0.67	*GPR158**
rs11224580	C/T	11	100.444	0.74	*PGR**
rs10842036	A/G	12	22.596	0.68	*KIAA0528*
rs1592672	G/T	12	78.653	0.76	*LOC100133105**
rs12440301	A/G	15	46.177	0.73	*SLC24A5*
rs10502789	A/G	18	38.324	0.75	*LOC284260**

Table 4.1: Location information for the 10 community-validated SNPs with highest delta values, together with their nearest gene (including hypothetical genes with 'LOC' prefix). A '*' indicates that the SNP resides within the gene. Delta values represent a comparison of 123 RHAS individuals who reported full Maori ancestry with 361 European individuals.

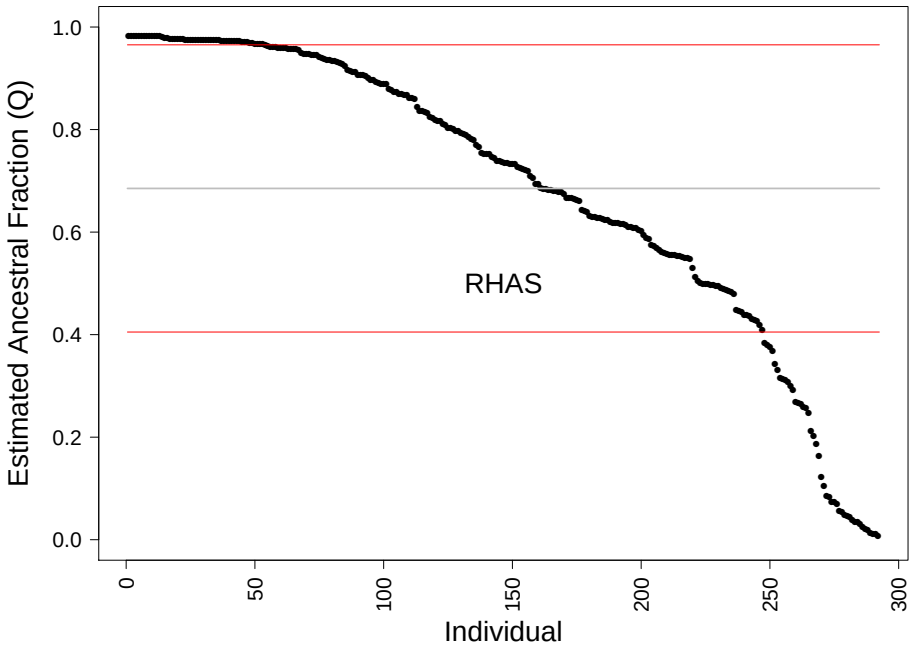

Figure 4.10: Structure output (K=2) for a the final set of 10 SNPs, showing Q values for all 292 participants from Rakaipaaka who consented to DNA extraction and analysis. Mean Q value (grey line) and the standard deviation of Q values (red line) are also indicated for this population.

[

(comparing 123 RHAS individuals who reported full Maori ancestry with the total set of 361 European individuals) was 0.67, while the maximum observed delta was 0.76.

4.5.5 Accuracy of reported Q values

Using the initial 30 RHAS and 90 CEU individuals, Q values were compared between sets of increasing numbers of SNPs, from 1 to 59 SNPs (Figure 4.11). Accuracy was calculated using the absolute difference between the 59-SNP set and the set under test, subtracting the result from 1.

The range in accuracy decreases as the number of SNPs in the set in-

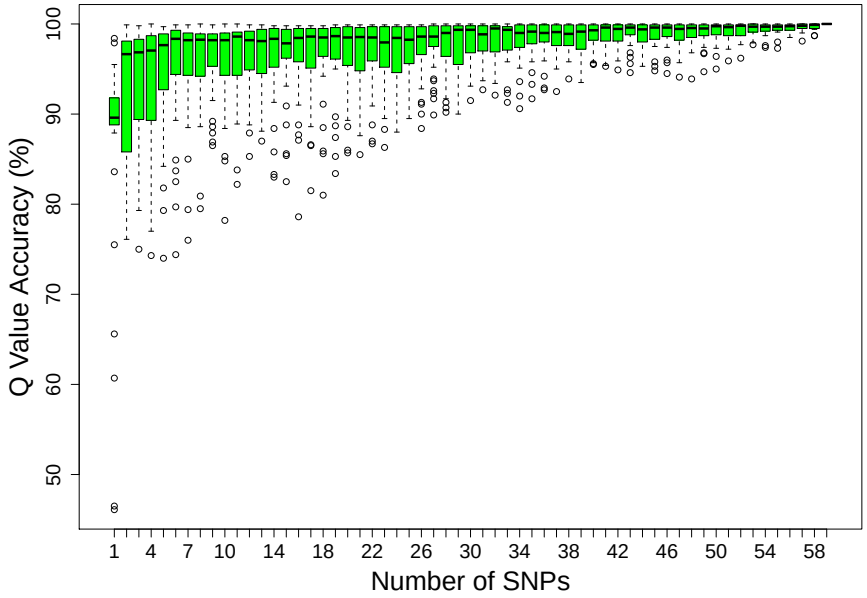

Figure 4.11: Boxplots demonstrating SNP set accuracy (Y axis) for increasing numbers of SNPs (from 1 to 59, X axis) assuming the 59-SNP set produces perfect Q values. The Q value for each individual was compared with the Q value for the 59-SNP set, and the error rate was assumed to be the same as the Q value difference. Data were generated using *structure*, using the 30 SNPchipped RHAS individuals and the 90 CEU individuals. Distributions shown in this figure are based on Q value differences for the 30 RHAS individuals only.

creases. The lower bound of accuracy is greater than 75% for all SNP sets of size 8 and greater, and median accuracy is greater than 95% for all SNP sets of size 2 and greater. The maximum Q value difference for the 10 SNPs set was 0.218 (accuracy of 78%), and median difference for the 10 SNPs set was 0.0431 (accuracy of 96%).

Figure 4.12 compares Q values derived from *structure* with reported ancestry from 228 RHAS individuals. The Pearson's correlation of Q vs reported ancestry is $r = 0.72$ ($p < 10^{-15}$). For the 123 individuals with a reported ancestry value of 1 (i.e. four Maori grandparents), no Q values were observed below 0.3. For the 10 individuals who reported no Maori ancestry, no Q values were observed above 0.09.

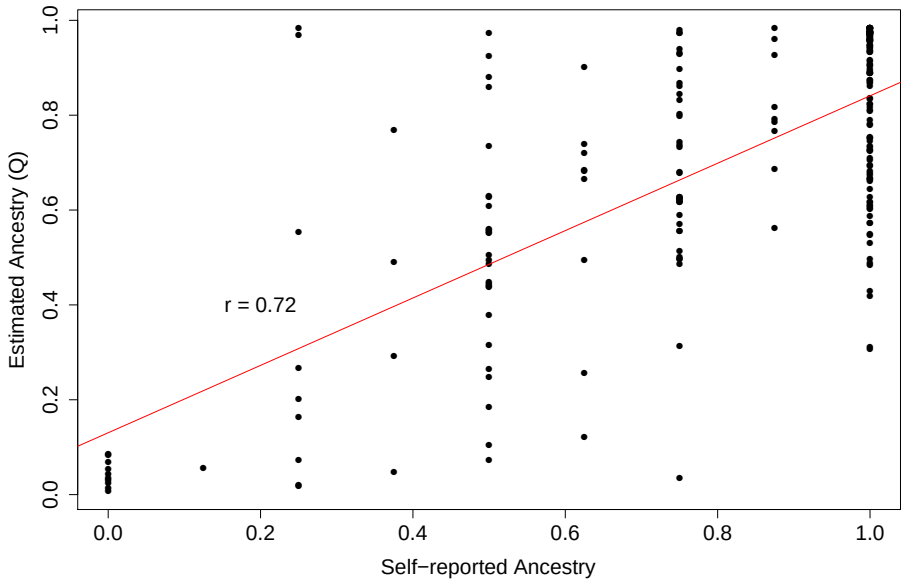

Figure 4.12: Scatter plot comparing structure-derived Q values with self-reported ancestry (228 RHAS individuals).

4.6 Discussion

A genome-wide scan of SNPs associated with Maori ancestry has been carried out, identifying those most consistently associated among subgroups, and these SNPs have been validated in independent populations. This study is the first genome-wide SNP analysis of a subset of the Maori population, and has demonstrated that only 10 SNPs (and possibly as few as 6) are required for a reliable estimation of European admixture in Maori.

The approach used here is similar to a Genome-wide Association Study (GWAS) in that two populations are compared in order to search for genetic markers across the whole genome that differ in frequency between the two populations. However, here a population sub-sampling process is applied to the set of markers in order to remove marker-population associations in

a discovery group that may not be applicable to a more general population.

The discovery of this set of 10 SNPs has also demonstrated the utility of the population sub-sampling method developed over the course of the investigation. For more details about the application of this method, see Chapter 5.

This set of SNPs is able to be used for the quantification of Maori *genomic* ancestry in an individual. Note that this is different from the typical genealogical interpretation of ancestry (see Frudakis, 2008). Genomic ancestry refers to what is passed down through the DNA, while *genealogical* ancestry refers to a person's direct ancestors, whether or not that genealogical history is detectable in the DNA. A demonstration of this can be found in Chapter 1, figure 1.5, where a chromosome has a genealogical ancestry of all eight ancestors, but a genomic ancestry that is composed of only six of those eight ancestors in quite unequal proportions. In this context, references to Maori ancestry relate to DNA inherited from the ancestral Maori population (i.e. a hypothetical population from which all Maori are descended).

4.6.1 Large Genomic Differences in Allele Frequencies between Maori and European Populations

The SNPs identified here describe a *genomic signature* – a genome-wide set of SNPs that can be used to estimate genomic ancestry. Although there are two polymorphisms (rs10842036, rs1592672) on this list that both reside on chromosome 12, their frequencies are likely to be independent, as the distance between these two SNPs is greater than 50MB (see Introduction, Section 1.2.4 and Section 1.3.1). Apart from those two SNPs, all other SNPs are on different chromosomes, so can be considered independent due to segregation at meiosis. Unlike gene based approaches for trait signatures (e.g. Huang et al., 2007), the process outlined here has found differences

throughout the genome, regardless of the presence of genes near mutations of interest. Such an approach considers the possibility that DNA sequence outside of genes may have a functional effect on a particular trait.

The genome-wide delta chromosome plot (Figure 4.4) indicates that the two populations are clearly distinct at a genomic level. The distribution of delta is spread consistently across the entire genome. This is what would be expected when looking at two groups separated by substantially different genomic ancestry.

4.6.2 Population Profile of Rakaipaaka Indicates Substantial Maori Ancestry

A total of 292 Adults from Rakaipaaka were profiled for the genotyping study (see Figure 4.10), with mean estimated Maori ancestral fraction within Rakaipaaka of 0.685 (SD = 0.280). The European individuals that they were compared against were the 90 HapMap CEU individuals and 271 PCE individuals from the discovery/validation phases. The mean estimated ancestral fraction for European individuals was 0.028 (SD = 0.040). The difference of the mean estimated ancestral fraction between the RHAS and CEU groups was 71.3%, suggesting 28.7% European admixture within Rakaipaaka.

Just over 15% of Rakaipaaka participants had an estimated European ancestral fraction that was smaller than the error of the test. For a theoretical perfect genomic ancestry test and uniform recombination, an individual would drop below this level with one European ancestor 5-6 generations back. While this is possible given that Europeans arrived in New Zealand earlier than 5-6 generations before present, it is also likely that these individuals have no European ancestry at all. This appears to support a hypothesis that there are still Maori who are not of European blood lines in New Zealand, despite some comments to the contrary (e.g. Brash, 2004).

4.6.3 Population Sub-sampling is an Effective Tool for the Identification of Ancestry-Informative Markers

A multi-marker approach for quantifying genetic variation has been presented, using an ideal model population (Rakaipaaka) for this task. The benefit of using multiple markers has been shown by Marchini et al. (2005), who found that a multi-marker approach will generate more informative results, even after considering the multiple-testing cost. The approach here has used a sub-sampling method, which may help in the removal of false positive signal that is common in GWAS (see Wellcome Trust Case Control Consortium, 2007; Healy, 2006).

The sub-sampling approach is advantageous because it reduces bias towards a particular grouping of individuals, and also buffers against a SNP being falsely selected due to genotyping error in an individual from the tested populations – this effect is more of an issue in smaller populations, especially when the observed minor allele frequency is low. This sampling approach should also reduce the influence of admixture on the selection of SNPs, as regions of admixture are unlikely to be consistent in unrelated individuals.

The high delta values observed in the genome-wide plot (Figure 4.4) are a strong indication that the tested populations are not substantially admixed, i.e. The Maori individuals who were genotyped reported no European ancestry in the past three generations, and these data support that.

The difference in the maximum rank between SNPs included in the bootstrap-consistent set of 64 SNPs and the remainder of the SNPs (Figure 4.5) suggests that the less-informative SNPs for given subgroups are being filtered out. These filtered SNPs are likely to be associated with particular groupings of individuals, rather than the general Rakaipaaka population (i.e. a false positive signal), but might be included in a typical

GWAS study where such a sub-sampling process is not carried out. The large number of SNPs captured in all 100 bootstrap sub-samples supports a hypothesis that these SNPs are not related to particular groupings, and reflect the general populations as a whole. In the presence of substantial admixture, fewer SNPs would be expected to be captured in all sub-samples, and the maximum rank for SNPs included in a bootstrap-consistent set would be higher (depending on the degree of admixture in the sample).

Rosenberg et al. (2003) have derived a formula that approximates the number of markers required for calculating ancestral fraction, $n = 1/(8 * \Delta^2 * V)$ (where V is the accepted variance, and Δ is the mean delta for the markers). Using this formula with an accepted standard deviation of 0.2 and delta of 0.7, the estimated number of markers for informative classification is 7, which is similar to the number of markers that have been used in this investigation.

The results shown in Figure 4.8 suggest that sampling 20-30 SNPs from the informative set of 59 SNPs would be more than sufficient to capture information about Maori genetic ancestry. While choosing the 6 top SNPs may also be effective, the chance of genotyping errors and the risk of false positive signal for one or more SNPs means that choosing the absolute minimum (for non-validated SNPs) is not the best choice.

Given the observation in Figure 4.8, it appears that as few as 6 SNPs are required to capture a large portion of the genetic differences between the two groups. Treating standard error as an indication of the normal range of values, this graph also suggests that 15 SNPs should be just as effective as 59 SNPs at quantifying genetic differences between the RHAS and CEU populations. This is evidence that only a small number of SNPs (15 or fewer, based on these results) are required in order to quantify the majority of the genetic variation in Maori ancestry.

4.6.4 Accuracy

The data available can not be used to determine the accuracy of the Q value output with respect to its ability to quantify true Maori genetic ancestry. In order to determine accuracy correctly, full ancestral information from a number of individuals with varying degrees of admixture would be required. Alternatively (or supplementary to this), full genome-wide SNPchip data from a similar group of individuals with varying admixture, independent of the initial discovery group, could serve a similar purpose.

However, estimates of accuracy have been determined, using the 59-SNP set (see Figure 4.11) and the self-reported ancestral fraction (see Figure 4.12) as reference values.

The boxplots in Figure 4.11 provide one estimate of accuracy and assumes that Q values generated using the 59-SNP set represent the true Q value for a fully informative set of markers. Difference in Q value (rather than difference from a Q value of 1) was used as a statistic to test accuracy because it removes systematic error associated with *structure* Q values (i.e. the estimated proportion of shared ancestry between the two populations). The median accuracy for a 10-SNP set was 96% , indicating that a 10-SNP set produces very accurate results for a large proportion of individuals. These results suggest that the choice of a 10-SNP signature set for the estimation of ancestry is reasonable in this case.

There is uncertainty about the true accuracy of the test when using this method of comparing Q values. Error is likely to be underestimated due to the assumption that Q values from the 59-SNP set represent the true genetic ancestry. Differences based on Q values from a larger set of SNPs are expected to be greater, but including other markers would require selecting them from the set of markers rejected by the sub-sampling process, increasing the risk of false positive associations and overfitting. Also, an underestimate of error is likely due to the marker set being designed to

maximise differentiation of the two population groups that are used to test accuracy.

Comparisons with self-reported ancestry (Figure 4.12) indicate a much higher error that exceeds 50% in some instances, but indicate that the Q values *underestimate* Maori ancestral fraction. However, the comparison with self-report is likely to overestimate error, as the self-report will also include error associated with phenotypic variation. For those individuals who reported no Maori ancestry, the maximum difference of Q from 0 was 0.09, which is consistent with the observed error in Q values in Figure 4.11.

Even when considering substantial error in observed Q values for the Parkinson's control individuals (PCE) in Figure 4.9, the 10-SNP set would still be effective as a qualitative test to determine if a person has substantial Maori genomic ancestry. A suggested cutoff Q value for this would be 0.3, based on the self-report comparison data.

Another method for estimating accuracy is by observing the spread of ancestry estimates for which the true ancestry value is known. In this case, CEU European individuals clearly do not have *any* Maori ancestry, so the true ancestral fraction for all CEU European individuals is really 0%. The mean estimated European ancestral fraction was 2.8%, so an assumption of a 97% accuracy in this Q value statistic can be made. This accuracy estimate is similar to that derived from the comparison of the 59-SNP signatures with 10-SNP signatures (median 98%, IQR 95-99%).

4.6.5 Genomic Ancestry is not a Good Representation of True Genealogical Ancestry

A genomic test for ancestry can never be a true representation of the genealogical (or biological) ancestry of an individual. Even if ancestral coefficients match the constructed genetic history, the random nature of recom-

bination produces an uneven spread of genetic sequences from different ancestral chromosomes.

It would be expected that given the random nature of recombination, a much more accurate representation of genetic ancestry could be obtained by using more markers. For example, a selection of SNPs could be made with 1 marker per 50cM, or around 60 markers across the entire genome. The rationale behind this would be that varied admixture in individual chromosomes would balance out over the whole genome. However, this does not appear to be the case, as can be demonstrated by the results in the 10-SNP to 59-SNP comparison (Figure 4.11), where a 10-SNP signature appears to be consistent with a larger 59-SNP signature (estimated 98% accuracy) for most individuals.

A couple of possible explanations for this come to mind. First, the random nature of recombination may exacerbate, rather than alleviate, admixture proportions – i.e. the true genetic structure is not a good representation of genetic history. Second, the selection process for the signature set of 10 SNPs has discovered markers that provide a very good approximation of admixture. These markers were chosen to maximise allele frequency differences between populations, and minimise variation in those frequency differences. This selection for minimal variation should ensure that the genetic history of blocks that markers reside in is consistent in different individuals, and therefore consistent throughout the population.

4.6.6 Conclusions

An estimate of the admixture profile of the Rakaipaaka population has been produced using the SNPchip platform together with publicly available data. Also, a 10-SNP signature set of genetic markers has been produced that has good accuracy in quantifying Maori genomic ancestry.

4.6.6.1 Applications

One direct use of the set of Ancestry-informative SNPs described here is as a genetic ancestry measure to control for admixture in disease studies. It may also be possible to use this value as an included factor for disease prediction, moving away from prediction based on group membership (i.e. ethnicity) and towards more personalised, gene-based prediction profiles.

Large allele frequency differences were used in combination to derive a set of genetic markers that could be used to estimate Maori ancestral fraction in this chapter. A population sub-sampling (bootstrapping) method was used in the process of generating this marker set, but this bootstrap sub-sampling method is also applicable to other traits that have a genetic basis. This method of sub-sampling is investigated in more detail in the next chapter (Chapter 5), where bootstrapping is used to identify a set of markers that could be used to estimate genetic risk for Type 1 Diabetes.

Chapter 5

Internal Validation of Genetic Associations for Type 1 Diabetes

5.1 Overview

Personalised medical treatment based on genome profiles is a major goal of genetic research in the 21^{st} century (see Avery et al., 2009; Province and Borecki, 2008). However, complex genotype-environment interactions for common diseases make it difficult to determine which specific genetic features should be used to construct such profiles. Hence the prediction of genetic risk is a major challenge of the 21^{st} century.

The introduction of large-scale Single Nucleotide Polymorphism (SNP) genotyping systems has enabled genetic variants to be typed *en-masse*, shifting the main effort required in a genetic risk study from genotyping to data analysis (or bioinformatics). The investigation of genetic markers for Type 1 Diabetes (T1D) in this chapter is a demonstration of how a population sub-sampling method developed in the previous chapter may assist in the identification of risk markers for a complex disease. This can be considered an alternative application of the method used in Chapter 4,

where the sub-sampling method was used for estimating Maori genomic ancestry.

This chapter is laid out in a slightly different fashion from a typical investigative study, in order to emphasise the theory behind the marker selection method. Method theory and results are combined into one section (Section 5.3), and the discussion of results for this particular study of T1D risk appear in the final section of the chapter (Section 5.4).

5.2 Background

5.2.1 Type 1 Diabetes

Type 1 Diabetes mellitus (T1D) is a disorder characterised by an absence of insulin-producing beta cells in the pancreas. This disorder shares with the more common Type 2 Diabetes mellitus (T2D) a characteristic symptom of high blood glucose levels. In some cases, this glucose also passes through to the urine, creating a sticky/sweet substance that attracts ants (see Ekoé et al., 2002, pp. 7,11). In T2D, this high blood glucose is caused by cells not responding to insulin (insulin resistance), while in T1D the excess is caused by a reduction in insulin production (insulin dependence).

The incidence of T1D varies throughout the world, with rates of incidence as low as 0.0006% per year in China, 0.02% in the UK, up to nearly 0.05% per year in Finland. About 50-60% of cases of T1D manifest in childhood (younger than 18 years), and the disease is believed to be caused by an abnormal immune response after exposure to environmental triggers such as viruses, toxins or food (see Daneman, 2006).

5.2.1.1 Symptoms, Diagnosis and Management of T1D

Typical symptoms of T1D include excess urine output (polyuria), thirst and increased fluid intake (polydypsia),blurred vision, and weight loss. When left untreated, this form of diabetes can lead to a build-up of ketone bodies and a reduction of blood pH (ketoacidosis), reducing mental faculties and causing a loss of consciousness (see Ekoé et al., 2002, p. 7).

Diabetes can be diagnosed by a single *random*[†] blood glucose test, as long as symptoms are present and blood glucose levels are found to be in excess (typically $> 11.1 mmoll^{-1}$) of those normally observed. In situations where symptoms are less obvious and/or glucose levels are at the high end of the normal range, a glucose tolerance test (GTT) is used for diagnosis. In this test, fasting patients have their blood glucose level tested, patients then consume a measured dose of oral glucose, and blood glucose levels are measured 2 hours later. A fasting glucose level in excess of $6.1 mmoll^{-1}$, or post-load level in excess of $11.1 mmoll^{-1}$ is considered diagnostic for both forms of Diabetes Mellitus. Type 1 Diabetes (as distinct from T2D) encompasses a range of diseases that involve autoimmunity. It can be diagnosed by the presence of antibodies to glutamic acid decarboxylase, islet cells, insulin, or ICA512 (see Ekoé et al., 2002, p. 19).

As the symptoms of T1D are caused by high blood glucose levels (hyperglycaemia) due to a lack of insulin, these symptoms can be relieved by the introduction of insulin into the blood. This is typically carried out by supplying measured doses of insulin via intramuscular injections or by the use of insulin pumps (see Daneman, 2006). Individuals with T1D need a constant supply of insulin for survival, together with occasional insulin bursts to control variable blood glucose levels throughout the day (e.g. after meals). Individuals with T2D only require insulin for survival in rare cases (see Ekoé et al., 2002, p. 16). Slow-release insulin and consumption of

[†]i.e. taken at any time of the day, as opposed to a *fasting* glucose test taken at least 8 hours after the last meal

foods with a low glycaemic index can help to reduce the extremes of T1D symptoms.

Improperly managed treatment can cause further medical complications in a diabetic patient. Too much insulin, excessive physical activity, or not enough dietary sugar can result in low blood glucose levels (hypoglycaemia), which produce short-term autonomic and neurological problems such as trembling, dizziness, blurred vision, and difficulty concentrating. Hypoglycaemia is treated either by ingestion of sugar, or by intravenous glucose in severe cases (see Daneman, 2006).

5.2.1.2 Complications of T1D

The initial symptoms of T1D are not usually severe, and the disease may progress for a few years before a diagnosis is made and treatment is given. However, long-term complications can appear when the disease is not managed appropriately (see Ekoé et al., 2002, p. 8). Retinal damage progresses in about 20-25% of individuals with T1D, with later stages causing retinal detachment and consequent loss of sight. Renal failure is also a problem in diabetic individuals, which is indicated by high urinary protein levels. When individuals have these high levels, progression to end-stage renal disease occurs in about 50% of cases. Neural defects are also a potential complication of T1D, most commonly damage to peripheral nerves, leading to ulceration, poor healing and gangrene unless good care is taken of the body extremities (see Daneman, 2006).

5.2.1.3 Genetic Contribution to T1D Risk

Type 1 Diabetes has a heritability of around 88% (Hyttinen et al., 2003), indicating that a substantial proportion of variance in disease susceptibility can be attributed to genetic factors. About 50% of the genetic contribution to T1D can be accounted for by variation in the *HLA* region on chromosome

6, and 15% is accounted for by variation in two other genes, *IDDM2* and *IDDM12* (see Daneman, 2006). Incidence rates in migrant populations quickly converge to those of the background population, suggesting that although the genetic contribution to the disease is high, environmental factors probably play a significant role in triggering the onset of disease (see Daneman, 2006).

5.2.2 Wellcome Trust Case Control Consortium Study

The Wellcome Trust Case Control Consortium (WTCCC)[†] was established in 2005 to identify novel genetic variants associated with seven common diseases, including Type 1 Diabetes (Wellcome Trust Case Control Consortium, 2007). 2000 individuals with T1D, and 1500 individuals from the National Blood Service (NBS)[‡] were genotyped for the WTCCC using an Affymetrix GeneChip 500k Mapping Array Set (See Introduction, Section 1.4.3.1).

The Wellcome Trust Case Control Consortium (2007) reported associations near five gene regions that had been previously associated with T1D: The major histocompatibility complex (*MHC*) on chromosome 6, *CTLA4* and *IFIH1* on chromosome 2, *PTPN22* on chromosome 1, and *IL2RA* on chromosome 10. The insulin gene (*INS*) on chromosome 11 was also associated with T1D; the only SNP tagging *INS* failed quality control filters, but also indicated strong association with T1D when examined. A number of other regions showed evidence of association with T1D in the Wellcome Trust Case Control Consortium (2007) study: 4q27 (chromosome 4); 10p15 (chromosome 10); 12p13, 12q13 and 12q24 (chromosome 12) 16p13 (chromosome 16); and 18p11 (chromosome 18). Most of these regions include genes involved in the immune system. However, only two genes are in 16p13, and both have unknown functions (*KIAA0350* and dexamethasone-induced

[†] `http://www.wtccc.org.uk`

[‡] The study also typed 2000 individuals for each of the six other diseases: a total of 14,000 cases genotyped for seven diseases.

transcript). The strongest association signal for T1D was detected within the *HLA* region of chromosome 6, a region in which multiple SNPs had strong associations with T1D, but only one of those SNPs (rs9272346) was reported in the results table of the strongest associations (see Wellcome Trust Case Control Consortium, 2007, table 3).

5.2.3 Replication Issues in GWAS

The Genome-wide Association Study (GWAS) is currently seen as the best way to tackle the search for genetic contributions to complex human diseases. The outcome of these studies is to determine the degree of association between single genetic markers and a heritable trait. Commonly, an analysis is carried out on a large number of genetic variants in a large number of people, allowing the detection of small genetic effects that are associated with a trait.

A study style that is built around correlation and association rather than a hunt for causal variants requires extreme care to ensure that observed associations are valid *and* causal. Studies need to have good within-study validation to reduce the likelihood of false-positive results being obtained and treated as true associations, and need to be supported by good independent validation. The distinction between association and causation is important – GWAS are used as hypothesis-generating tools to narrow down, through association, the search for potential causative loci. After the associations have been validated, it is expected that they will be followed up with studies attempting to determine the true causative status of that association. Such causative studies are difficult, and progress towards understanding the aetiology of common disease has been slow (see Dermitzakis and Clark, 2009).

5.2.4 Sampling Errors in GWAS

Natural variation of genotypes within populations means that any particular sample from the population may not represent the true genotype frequencies within that population (see Introduction, Section 1.4.6). This may lead to the observation of marker-disease associations when no such association exists. This is particularly important when considering populations with mixed ancestry, where markers that are informative for distinguishing population ancestry may become accidentally associated with a particular disease (see Pritchard and Donnelly, 2001).

Bootstrapping by repeated re-sampling of a representative draw made from a group can estimate population variation in genotype frequencies by observing variation within the sub-samples. The marker selection method used in this chapter utilises a re-sampling technique similar to that used in Chapter 4 in order to reduce the influence of allele frequency variation in producing false-positive results for particular samplings of the population.

5.3 Method and Results

5.3.1 Method Summary

The Wellcome Trust Case Control Consortium (WTCCC) have genotyped 2000 individuals diagnosed with T1D, and 1500 individuals from the National Blood service (NBS) using the Affymetrix 500k chip (500568 SNPs). These genotypes were obtained from WTCCC for subsequent computer-based research exploring the utility of the author's new bootstrap sub-sampling method for genome-wide association studies. Genotype data were filtered at a SNP level to remove those SNPs that were present on the X chromosome; individuals flagged by WTCCC as having potentially invalid genotype data were removed from the dataset.

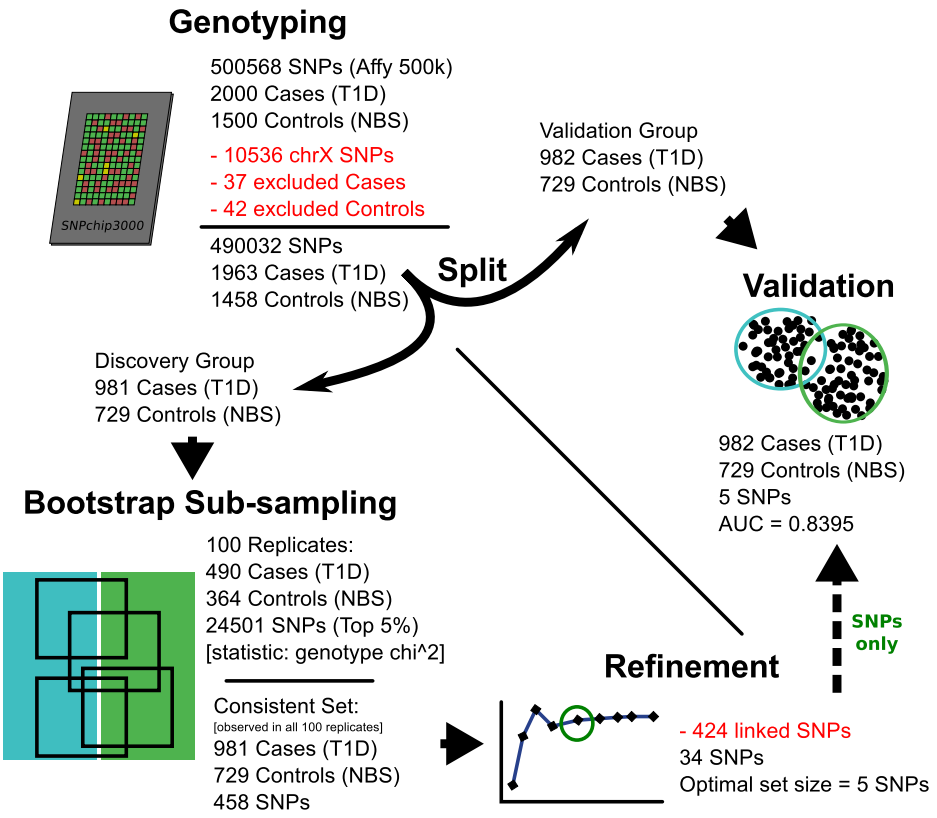

Figure 5.1: An overview of the marker set construction procedure, using an initial validation/discovery split, bootstrap sub-sampling, set refinement, and internal validation.

The study group was randomly split into two equal-sized groups: a *discovery* group (981 T1D cases, 729 NBS controls), and a *validation* group (982 T1D cases, 729 NBS controls). Subsequent filtering, analysis, and selection of SNPs was carried out on the discovery group, while the validation group was *only* used to test the effectiveness of the selected SNP set in a situation distinct from that used to generate this set of SNPs (see Figure 5.1).

A bootstrap sub-sampling method was used to reduce the initial panel of 500k SNPs down to a set that consistently produced associations on all bootstrap sub-samples. Sub-samples of the Type 1 Diabetes (T1D) cases and

National Blood Service (NBS) controls were used to estimate the in-group variance of association statistics throughout the genome. Markers that were informative and had low variance were selected as candidate markers for a minimal informative set of 45 markers.

The final refinement step tested sets of SNPs in combination, rather than single SNPs alone, in the hope that those sets would be able to capture a wider range of genetic variation than any single marker (or combination of data for single markers alone) could provide. Once a suitable SNP set was found, that set was tested in the validation group to confirm the utility of the set for distinguishing T1D cases from NBS controls.

5.3.2 Genotyping and Filtering of Individuals and SNPs

Genotype data from 2000 T1D cases and 1500 NBS controls were provided by WTCCC. This genotyping had been carried out on an Affymetrix 500k SNPchip (500568 SNPs). The purpose of the initial genotyping procedure is to obtain a large sample of candidate markers ($> 100,000$) from which to pick the most informative.

Due to sex differences in expression for X-chromosome SNPs, all 10536 X chromosome SNPs were excluded. The WTCCC dataset also included a list of 37 T1D cases and 42 NBS controls to exclude for a number of reasons (e.g. high proportion of missing genotype data, duplicate individuals, non-European ancestry), so these individuals were also removed from the present study (X chromosome filtered set of 490032 SNPs, 1963 cases, 1458 controls).

5.3.2.1 Separation of Discovery and Validation Groups

Individuals were randomly assigned into one of two groups, a discovery group with 981 T1D cases and 729 NBS controls, and a validation group

with 982 T1D cases and 729 NBS controls. To ensure a robust analysis, the validation group was only used for the final validation of a generated SNP set, and not used for any part of the SNP discovery procedures.

An alternative approach to genotyping a group of individuals from a target group is to use large public (and free) datasets from the closest matching groups to generate the initial set. This approach would be used if large-scale genotyping from the target group were unavailable, or would be prohibitively expensive.

While genotyping similar groups is much cheaper, it also has a high chance of misses for relevant markers, as the marker mutation profile of populations can differ quite significantly between populations. As a demonstration of this, allele frequency differences between HapMap CEU and HapMap CHB populations were calculated for SNPs in a 10Mb region centred on the ADH region. In this case, the HapMap CHB population was used as a proxy for the Maori population. A total of 37 SNPs were selected, all with allele frequency differences greater than 0.5, for then genotyping in 45 Maori (the target group). Of these 37 SNPs, 19 had allele frequency differences between HapMap CEU and Maori populations of less than 0.2, and only 2 SNPs had allele frequency differences greater than 0.5.

5.3.2.2 Marker Association Values Across the Entire Genome

Association scores were calculated across the entire autosomal genome (see Figure 5.2). A genotype χ^2 test was used, comparing observed genotype counts for each group to expected genotype counts for both groups combined. High association scores ($\chi^2 > 100$) were found on chromosomes 3, 6, 12, 16, and 22, but the scores were only consistently high for a region of about 10Mb in the middle of the short arm of chromosome 6 (30-40Mb from the $5'$ end of the forward strand).

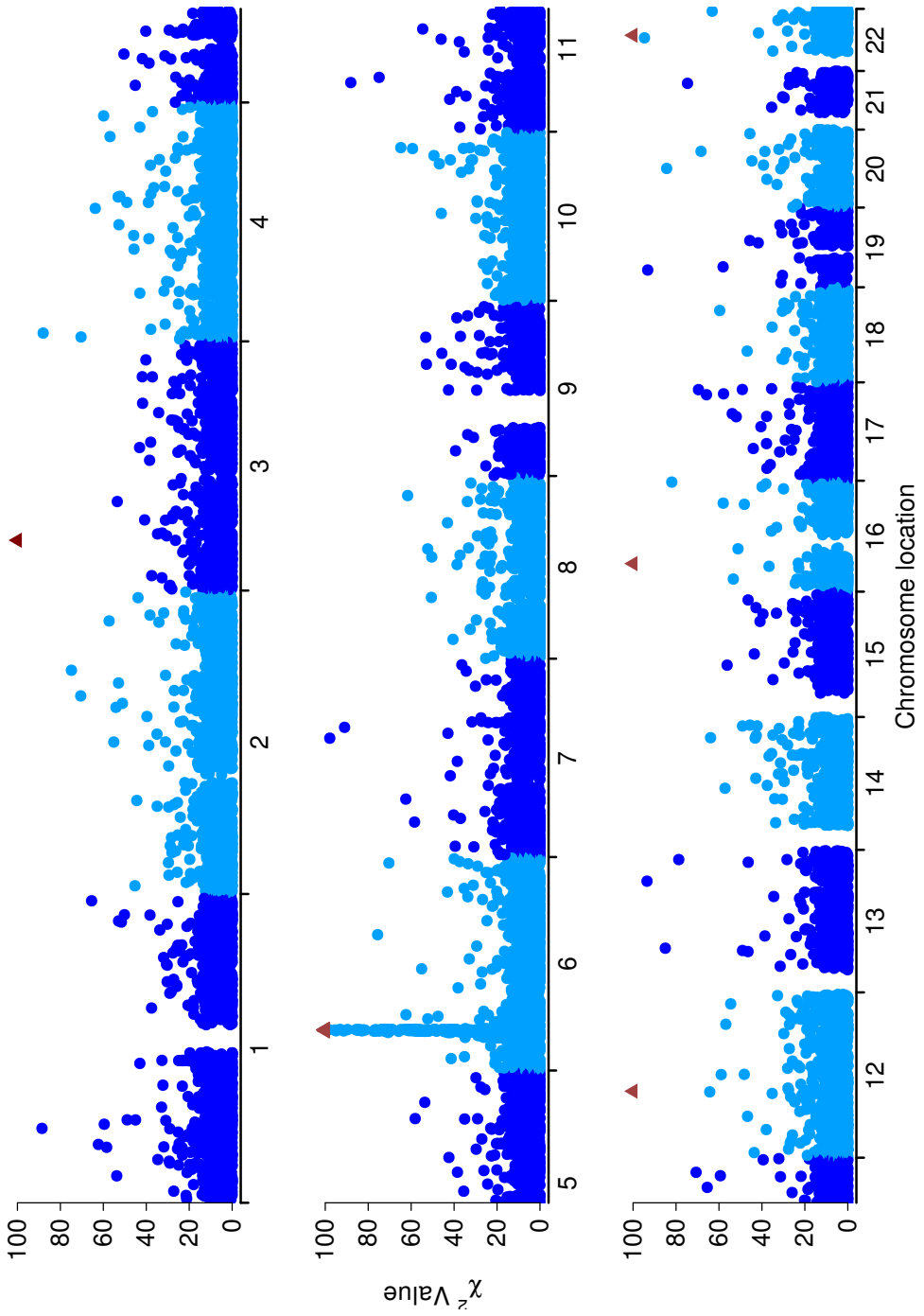

Figure 5.2: Scatter plot indicating unfiltered marker association values across the autosomal genome in the discovery group. Alternating colours are used here as an aid to distinguish different chromosomes. Values greater than the range of this graph ($\chi^2 > 100$) are shown at the top of the graphs as a triangle symbol. A large spike of high association values can be seen near the middle of the short arm of chromosome 6. Gaps appear in all chromosomes at the centropere, where genotype data is not available.

5.3.3 Bootstrap Sub-sampling of the Discovery Group

A bootstrap sub-sampling method was then carried out, generating 100 subsample replicates of the discovery group, each replicate having 490 T1D cases and 364 NBS controls (i.e. retaining the same proportions as the original 981 cases and 729 controls), sampled from the original discovery set without replacement. The SNPs were then ranked by χ^2, and a *bootstrap-consistent set* of 458 SNPs was identified, each ranked in the top 5% of SNPs (24501 SNPs) in every bootstrap sub-sample (see Figure 5.3-A). Most of these 458 SNPs had a maximum rank below 5000, whereas most of the remaining 489574 SNPs had a maximum rank above 350000 (see Figure 5.3-B).

5.3.4 Bootstrap Sub-sampling

The bootstrap sub-sampling method was used to eliminate those markers from the initial X chromosome filtered set of 490032 SNPs that were not effective for genetically distinguishing case and control groups. In each iteration of the bootstrap process, a sub-sample of individuals from each group was carried out, then markers were ranked based on a statistic that evaluates the effectiveness of each marker (see Figure 5.4). Markers that consistently had a high association statistic in each bootstrap sub-sample were selected for the next stage in the process.

5.3.4.1 Comparison of Bootstrap Sub-sampling With a Simple Ranking Method

The bootstrap sub-sampling process attempts to eliminate markers that are specific to the particular sample of individuals under study, rather than the more general population those individuals have been sampled

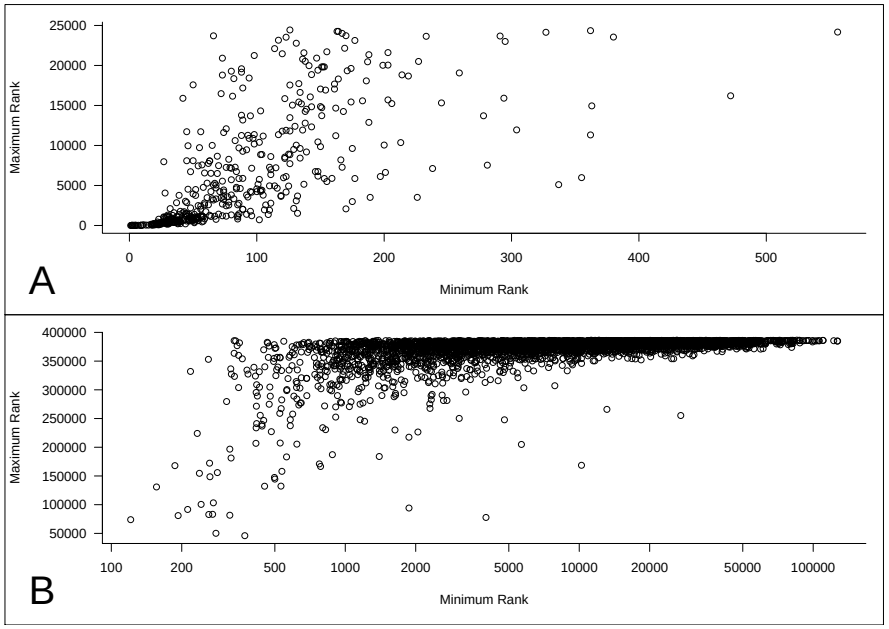

Figure 5.3: Scatter plot indicating maximum rank over all bootstrap sub-samples vs. minimum rank in any bootstrap sub-sample for the bootstrap-consistent set of 458 SNPs (A), and a random sample of 5000 of the remaining SNPs (B). A total of 126519 SNPs (not included when generating these graphs) were unranked in at least one bootstrap sub-sample, as no genetic difference was observed between case and control groups with that SNP. The difference between minimum and maximum rank gives an indication of the reliability of a particular marker for association testing in a general population. Of those markers in the bootstrap-consistent set of 458 SNPs, 57% were ranked in the top 5000 markers in all bootstraps. Of the remaining 489574 SNPs, 95% (4464977) had a maximum rank of 350000 or more (including 126519 unranked SNPs).

from. The effect of using a simple ranking procedure that has no sub-sampling would be to identify the markers that are most differentiated in that particular sample of individuals. However, natural variation in genotype frequency introduces noise into association analyses, so markers that are differentiated in a particular sample may not be differentiated in the population the sample was derived from.

The problem of discovering associated features that are not present in the more general case is known as overfitting (see Russell and Norvig, 2003, Chapter 14, pp. 661-663). In the conventional GWAS context, overfitting

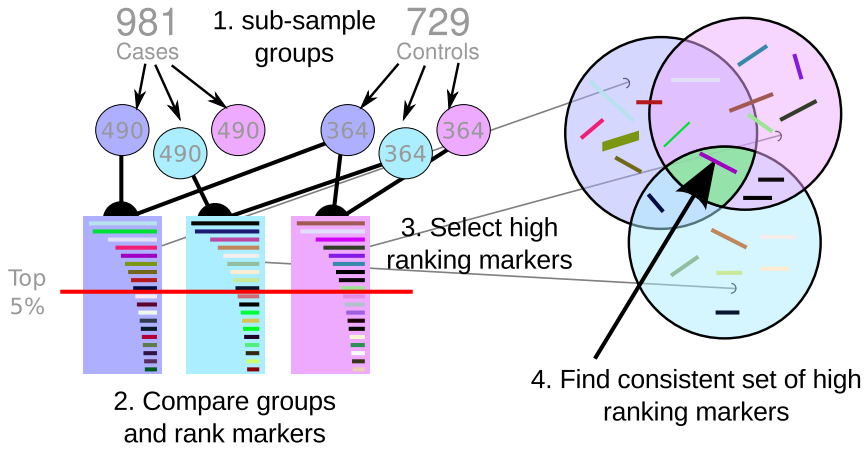

Figure 5.4: Visual representation of the key points of the bootstrap process. The groups are sub-sampled a number of times, and marker ranking statistics are calculated for each sub-sample (bootstrap). Markers are then ranked, identifying the markers with the highest association statistic for each sub-sample. Markers that were consistently ranked in the top 5% in all sub-samples were passed onto the next stage of the selection process.

produces false positive associations, where an association with a particular genotype does not extend to the general population. Using a method that includes bootstrap sub-sampling should reduce the degree of overfitting by removing markers that are only relevant for distinguishing between the specific groups involved in marker discovery.

5.3.4.2 Choosing a Marker Ranking Statistic

A ranking statistic is necessary for the bootstrap process to determine which markers are more likely to be associated with the phenotype of interest. The purpose of this statistic is to rank the effectiveness of markers in distinguishing groups, rather than give a precise indication of their utility. This means that the actual statistic used is not important, as long as it is able to rank an informative marker higher than a less informative marker. Statistics useful for evaluating genetic association are discussed earlier in

this thesis (see Chapter 1, Section 1.4.4). In this case, a genotype-based χ^2 statistic was chosen for evaluating marker effectiveness. This statistic considers situations where a heterozygous genotype may have a strong association that is not present in either homozygous genotype, as well as identifying strong associations for homozygous genotypes.

5.3.4.3 Ranking Markers Using the Observed Distribution of Ranking Statistics

A non-parametric ranking method selected markers based on rank order across all bootstraps. Markers are assigned a rank within each bootstrap: the marker with the most informative statistic is assigned rank 1, the second-most informative is assigned rank 2, and so on. The minimum, maximum, and mean marker rank are determined for each marker in all bootstrap sub-samples (see Table 5.1).

Markers that are not ranked in the top 5% of markers in *any* sub-sample are excluded from further analysis. When using this process on the T1D discovery group, a *bootstrap-consistent set* of 458 SNPs were found in the top 24501[†] SNPs in *all* 100 sub-samples. Of these bootstrap-consistent SNPs, 182 (40%) are located between 30Mb and 33Mb from the beginning of chromosome 6, near the *HLA* region. The remaining 276 SNPs are distributed fairly evenly throughout the genome (see Figure 5.5). From these observations of chromosomal location, T1D appears to have a very strong association signal near the *HLA* region on chromosome 6, and limited signal elsewhere in the genome.

5.3.5 Linkage Refinement

Linked SNPs were removed from the bootstrap-consistent SNP set in order to reduce the redundancy of associative signal produced by the generated

[†] $24501 = \lfloor 0.05 * 490032 \rfloor$

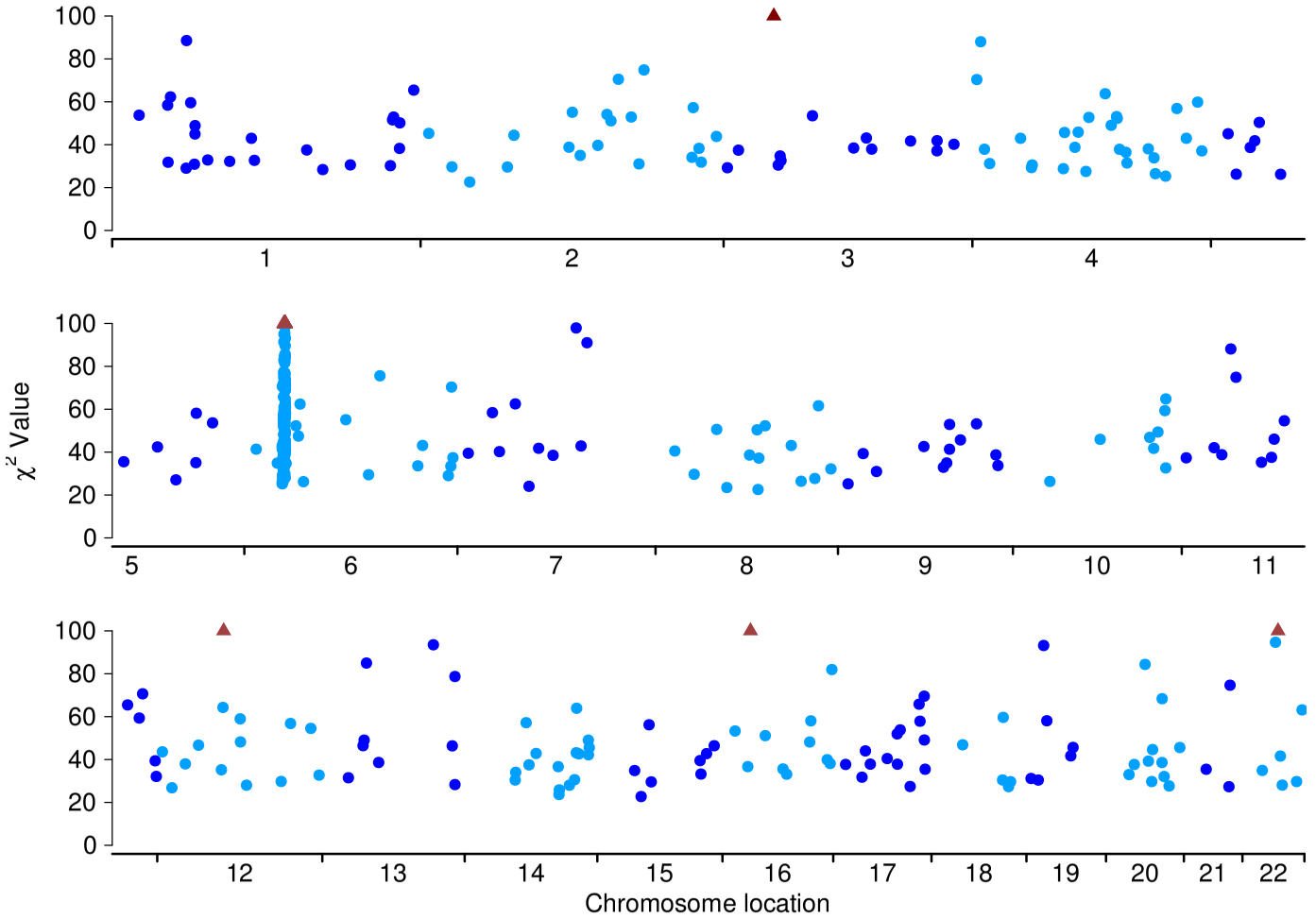

Figure 5.5: Scatter plot indicating marker association values for the consistent set of 458 SNPs across the autosomal genome in the discovery group. Values greater than the range of this graph ($\chi^2 > 100$) are shown at the top of the graphs as a triangle symbol. The large spike of high association values still remains near the middle of the short arm of chromosome 6.

Marker	Min Rank	Max Rank	Mean Rank
rs2027852	9	27	17.8
rs3135342	42	1598	196.2
rs16917773	48	2196	261.5
rs10144861	59	2817	491.3
rs10842028	44	3028	543.4
rs10742084	61	5290	679.9
rs7158350	84	6736	730.4
rs16854531	126	13477	735.1
rs1429445	54	7451	975.9
rs17023486	472	16210	2903.6
rs12117563	1508	377782	77651.8
rs11189528	1563	385065	162459.9
rs17038075	47938	453037	172338.7
rs11081211	13893	376453	185579.0
rs11249611	1731	376943	190683.0
rs16831752	119403	461199	222291.4
rs1006931	22398	369366	232357.9
rs9317562	15117	383671	235382.3
rs11205709	477789	477983	477890.9
rs1027341	487082	487165	487114.9

Table 5.1: A sample of markers from the T1D study, showing minimum, maximum, and mean rank in 100 bootstrap sub-samples. In order to demonstrate differences between included (low rank in *all* bootstrap sub-samples) and excluded markers, the first ten markers were sampled randomly from the group of 458 markers ranked in the top 5% of markers in all subsamples, and the remaining ten markers were sampled randomly from the remaining 489574 markers.

SNP set. Markers were ordered based on mean rank order and any SNPs that were linked ($r^2 > 0.1$) with a higher-ranked SNP were removed from the set, leaving an *unlinked set* of 34 SNPs.

Markers within a signature marker set should be unlinked, so it is a good idea to calculate a linkage-associated statistic such as D' or r^2 during the discovery phase of the analysis, and remove the least informative marker among linked high-association pairs. This step is carried out after the bootstrap sub-sampling process in order to reduce the number of

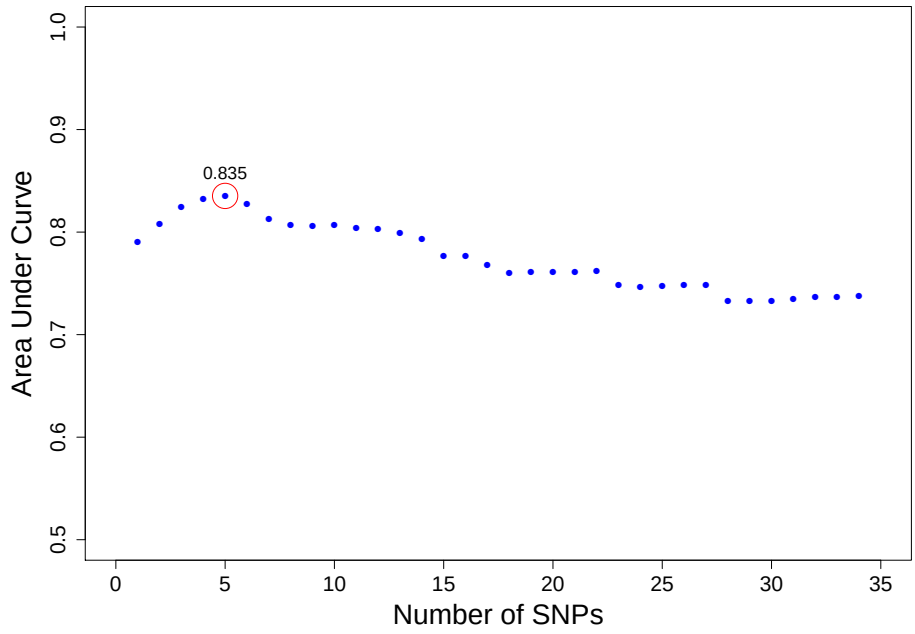

Figure 5.6: A marker refinement plot, showing the effectiveness score (AUC) for increasing numbers of SNPs in the discovery group. The highest AUC value (0.835 for 5 SNPs) is circled in red.

pairwise calculations required for linkage analysis – pairwise calculations for 500 markers would require 124,750 linkage comparisons,[†] while pairwise calculations on 500,000 markers would require around 1.25×10^{11} comparisons.

5.3.6 Set Size Refinement

The optimal marker set size was identified using an Area Under the Curve (AUC) test on the Q-values generated by *structure* (10,000 bootstraps, and 100,000 total runs), finding marker sets with large differences in mean Q value between the two groups (see Figure 5.6). Increasing numbers of markers were selected from the unlinked SNP set based on mean rank

[†] $124,750 = (500^2 - 500)/2$

order identified during the previous (bootstrap sub-sampling) stage. The effectiveness of a given set of markers was evaluated using the *structure* program (see Introduction, Section 1.4.7.1), followed by an AUC calculation for each set of markers based on Q values reported by the program.

The *structure* program outputs values that represent how genetically similar an individual is to a particular group (Q values), attempting to cluster pooled individuals into two "populations".[†] The Q values produced by *structure* are continuous in the range between 0 and 1 inclusive, and are treated as an estimate of the probability that an individual has a particular trait.

Analysis of Q values was used to determine false positive and true positive rates for given Q-value cutoffs (see Figure 5.8). The true positive rate was calculated as the proportion of T1D cases with Q below the cutoff value, and false positive rate was calculated in the same way for NBS controls. The area under the curve of this graph can be used as an indication of the effectiveness of a quantitative test. An AUC of 1 indicates a perfect test (no misclassification), while an AUC of 0.5 indicates a test that cannot distinguish between groups.

The greatest difference between cases and controls was observed when the top 5 SNPs were selected, producing an AUC of 0.8449. This *signature set* of 5 SNPs was considered to be the most appropriate T1D-informative set.

5.3.7 Validation of Final 5 SNP Set

The signature set of 5 SNPs (see Table 5.2) was finally tested on the validation group (982 T1D cases, 729 NBS controls) using *structure*, followed by an AUC analysis of the Q values. There is a small overlap between some

[†]The *structure* program is designed for *population* analysis, but is used here for *group* analysis.

Marker	Chromosome	Location (Mb)	χ^2	Mean Rank
rs9273363	6	32734250	485	1
rs3957146	6	32789508	317	2.2
rs3135377	6	32493377	264	4.3
rs7431934	3	40268801	199	13.7
rs1046089	6	31710946	108	37.9

Table 5.2: Location information for the top 5 SNPs discovered in a bootstrap sub-sampled GWAS for T1D associations, after removing linked SNPs, and choosing the set with the highest AUC value. Mean rank reported in this table is based on the marker rank for 100 bootstrap sub-samples. Out of the five markers, four are within a 2Mb region of chromosome 6.

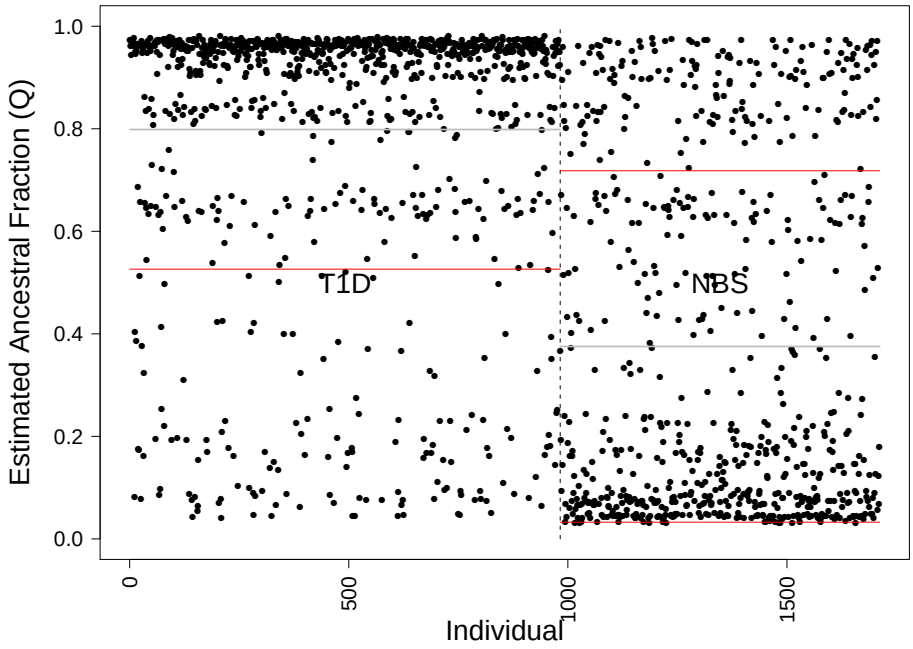

Figure 5.7: Structure output (K=2) for the top 5 SNPs discovered in a bootstrap sub-sampled GWAS for T1D associations, showing Q values for individuals from T1D and NBS groups (using validated group). Mean Q value (grey line) and the standard deviation of Q values (red line) are also indicated for each group.

T1D cases and some NBS controls (Figure 5.7), but most T1D cases cluster together, and are separate from the cluster of NBS controls.

The AUC value associated with this test of the signature set of 5 SNPs in

SNP	Chr Region	Gene Locus
rs9270986	6q21	*HLA*
rs6679677	1p13	*PHTF1-PTPN22*
rs17696736	12q24	*C12orf30*
rs2292239	12q13	*ERBB3*
rs12708716	16p13	*KIAA0350*
rs2542151	18p11	*PTPN2*
rs3741208	11p15	*INS*
rs17388568	4q27	*Tenr-IL2-IL21*
rs7722135	5q14	*Q8WY63*
rs9653442	2q11	*AFF3-LOC150577*
rs6546909	2p13	*DQX1*
rs2666236	10p11	*NRP1*

Table 5.3: A list of SNPs found by other researchers to be associated with T1D risk. The first SNP (rs9270986) yielded the most extreme statistic in the WTCCC analysis (Wellcome Trust Case Control Consortium, 2007). Marker names and locations for the remaining 11 SNPs are from Table 1 of Todd et al. (2007).

the validation group was 0.8395. Setting the false positive rate to 5% (cutoff Q value 0.129) produced a true positive rate of 43%, while setting the true positive rate to 85% (cutoff Q value 0.5583) produced a false positive rate of 38%. The position on the curve nearest to a true positive rate of 100% and a false positive rate of 0% was when the cutoff Q value was set at 0.506, with a true positive rate of 78%, and a false positive rate of 29%.

5.3.8 Comparison with SNP set from Literature

Todd et al. (2007) carried out an analysis of 11 SNPs that were found to be associated with Type 1 Diabetes in genome-wide association studies. This group of SNPs, in combination with the most informative SNP from the WTCCC study (Wellcome Trust Case Control Consortium, 2007), was selected to be compared with the signature set of 5 SNPs in the present study (see Table 5.3). The *structure* program was used in combination with an AUC analysis to evaluate the effectiveness of this group of 12 SNPs

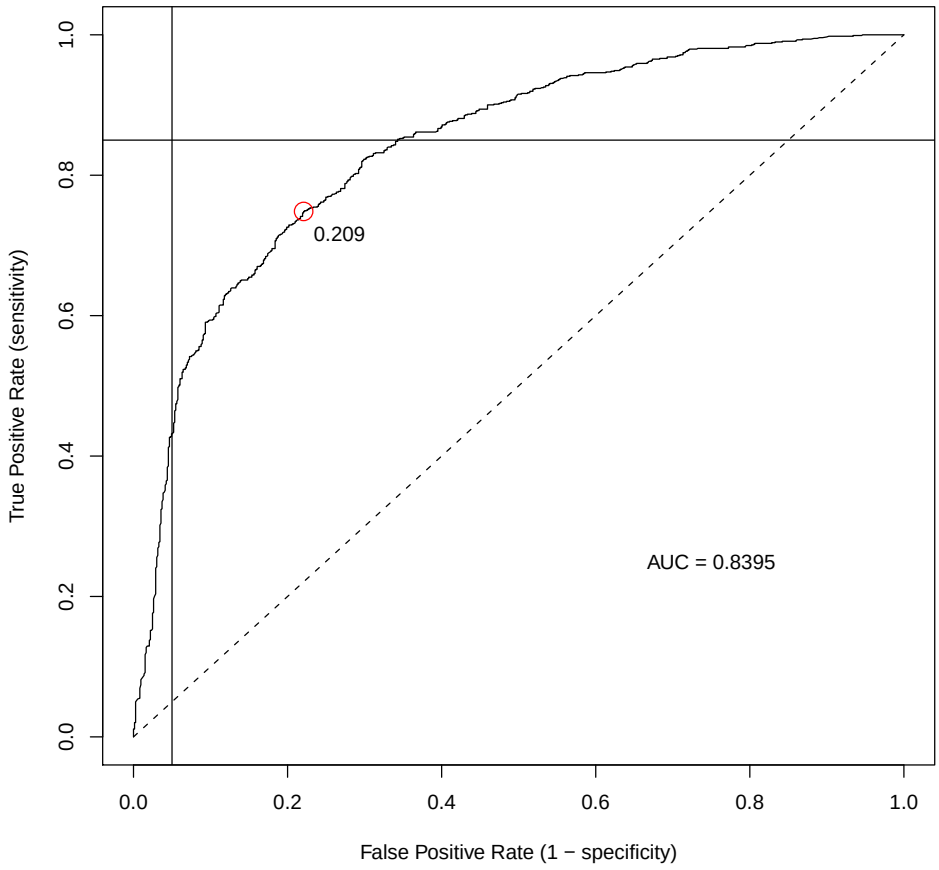

Figure 5.8: Receiver-operator characteristic graph of true positive rate vs. false positive rate based on the structure plot of validated set of 5 SNPs (see Figure 5.7). The area under the curve (AUC) of this graph indicates that when comparing randomly selected individuals from each group, the probability that a T1D case individual will have a higher Q value than a control individual is around 84%.

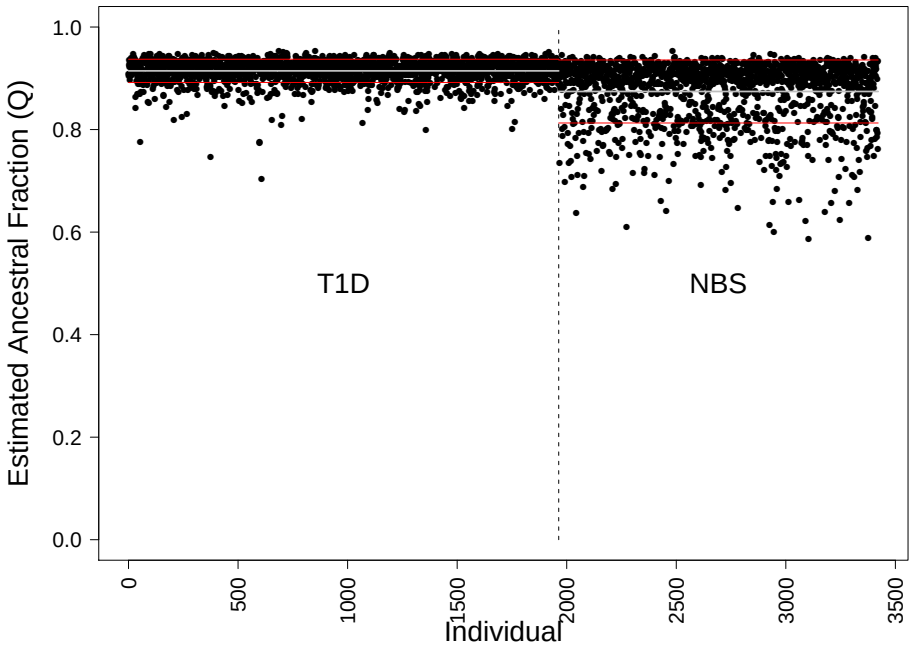

Figure 5.9: Structure output (K=2) for 12 SNPs found by other researchers to be associated with T1D risk (see Table 5.2). Mean Q value (grey line) and the standard deviation of Q values (red line) are also indicated for each group.

for 1963 WTCCC T1D cases and 1458 NBS controls (see Figure 5.9 and Figure 5.10).

This 12-SNP comparison set had an AUC of 0.73 when tested with 1963 T1D cases and 1458 NBS controls. Setting the false positive rate to 5% (cutoff Q value 0.933) produced a true positive rate of 18%, while setting the true positive rate to 85% (cutoff Q value 0.895) produced a false positive rate of 53%. The position on the curve nearest to a true positive rate of 100% and a false positive rate of 0% was when the cutoff Q value was set at 0.910, with a true positive rate of 65%, and a false positive rate of 35%. These results indicate that the signature SNP set discovered in the present study is considerably more informative than a set of T1D-associated SNPs found in other genome-wide association studies.

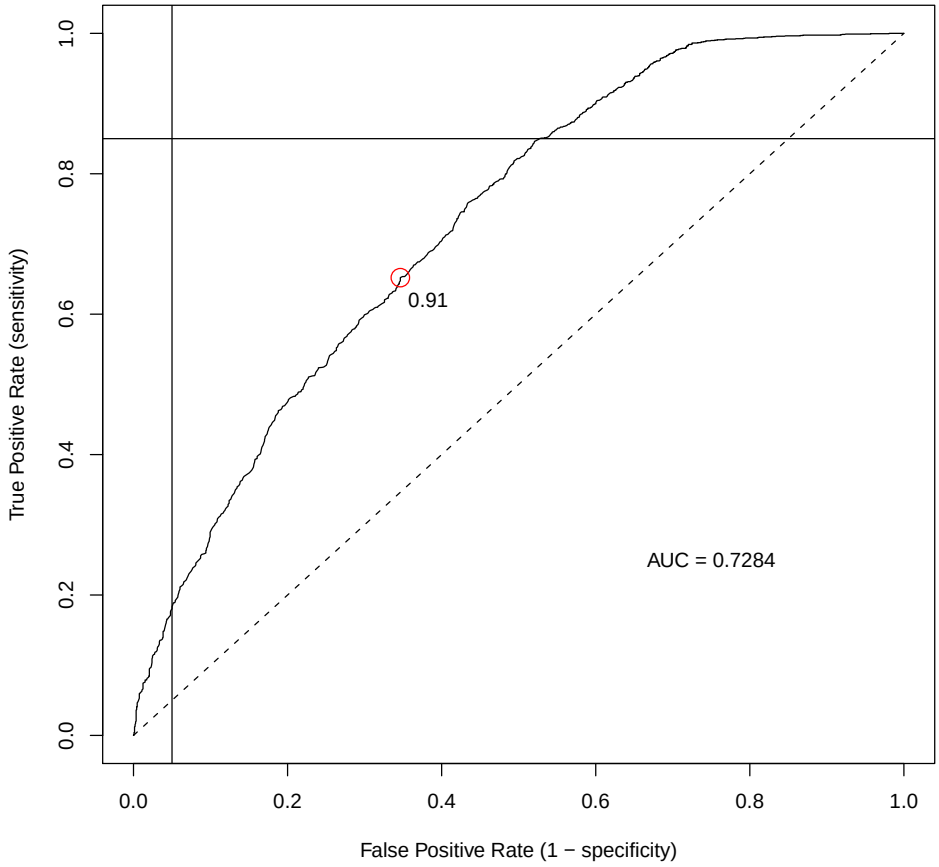

Figure 5.10: Receiver-operator characteristic graph of true positive rate vs. false positive rate based on structure plot of SNPs found by other researchers (see Table 5.2 and Figure 5.9).

5.4 Discussion

This study has identified a group of 5 SNPs that classify individuals with T1D with good reliability (AUC = 0.84, see Figure 5.8). The heritability of Type 1 Diabetes is around 88% (Hyttinen et al., 2003), so the maximum possible sensitivity (true positive rate) of a genetic test for T1D should be 88%, with the remaining 12% of variation being due to non-genetic factors.

One of the assumptions made in GWAS is that the individuals selected as candidates for the phenotypic groups (cases and controls) are ideal members of those groups – affectation status tends to be a binary or integer value that does not allow for intermediate values. Due to the difficulty in qualitatively describing traits, as well as mutation and admixture effects (particularly for population-derived groups), this assumption may be invalidated.

The marker construction method used a bootstrapping procedure as an internal validation to remove markers that had substantial variation in χ^2 values within the tested groups. In an ideal case, a bootstrapping procedure would not be necessary as the genetic makeup of the total population will reflect the makeup of any given subgroup of that population. In such a case, the ranking after each bootstrap will be the same as the overall ranking. However, the comparison of minimum and maximum rankings for SNPs across all bootstrap sub-samples has demonstrated that this is clearly not the case (see Section 5.3.4).

5.4.1 Type 1 Diabetes Study Results

It is known that genetic variation within the *HLA* region on chromosome 6 plays an important role in T1D, accounting for about 50% of the genetic susceptibility for T1D (see Daneman, 2006). This role is supported by the preliminary results in the present study, which show consistently strong

predictive power using genetic markers, all but one from this region alone (see Table 5.2).

5.4.1.1 Accuracy of the Signature SNP Set

The interpretation of accuracy of a genetic test is difficult, particularly when considering what would be expected if the test were used in an untested population. A statistic that can be useful in this case is the positive predictive value (how likely a test is positive, given a positive result).

In order to determine the positive predictive value of a test, it is necessary to establish the prevalence of the trait in the population of individuals who are to be tested. A country which is considered to have a very high incidence of T1D, Finland, has an overall cumulative incidence of around 0.5-0.6% at the age of 35 years (Hyttinen et al., 2003). Also, there has been a general trend of a 2-3% increase in the incidence rate of childhood T1D in South West England over the past 20-30 years, with the incidence in 2003 at around 0.16% per year (Zhao et al., 2003). Even at the higher incidence rate in Finland, fewer than 0.6% of individuals in a typical non-enriched control population would be expected to have T1D.

The NBS controls for the WTCCC study had not been enriched to remove individuals that have T1D. Given an expected prevalence of T1D of 0.6%, it would be expected that around 4 individuals from the validation NBS control group (or 9 from the discovery and validation groups combined) have T1D. Setting the false positive error rate to this value (i.e. 0.6%) is unrealistic for the current data set, as only a small fraction of T1D cases would be identified with that cutoff (just over 5%, see Figure 5.8). However, if a more moderate 5% false positive error rate is accepted (identifying 43% of T1D cases, see Section 5.3.7), then 36 NBS individuals would be identified by this test as at risk for T1D. This is about ten times that expected by cumulative incidence rates for T1D, indicating a positive predictive value

of 10% with the discovered signature set of 5 SNPs. Given that the population prevalence of T1D is so low, the NBS control group should not differ substantially from an enriched control group, and the positive predictive value of this genetic test will remain around 10%.

5.4.1.2 Accuracy in Other Populations

The low positive predictive value of the marker set, together with heritability values of less than 100%, means that it is unlikely that a genetic test using these T1D markers would be useful as a *diagnostic* test for a general population. However, if used in conjunction with other clinical indicators, it may be appropriate to use these genetic markers for a *screening* test, identifying individuals that should be more closely monitored for T1D symptoms. This is because it will still exclude a large proportion of the normal population, while also identifying a high proportion of at-risk individuals. However, the signature SNP set has not been validated in groups of individuals outside the WTCCC study, and caution should be taken in attempting to extrapolate results to non-validated populations.

Taken in the context of disease, it can be very difficult to accurately determine the phenotype of an individual – this is a particular problem when the disease is a continuous (rather than discrete) trait, as often happens with common complex diseases. Phenotype identification is further complicated by non-Mendelian patterns of inheritance. It is possible for there to be numerous paths to the same apparent end disease, and numerous gene-gene interactions that contribute to the same disease. Furthermore, trait variation is often a mixture of genetic and environmental factors (i.e. heritability is less than 100%), so potential gene-environment interactions also need to be taken into account when describing phenotype.

The effectiveness of any given set of markers will be reduced due to the presence of erroneous false positive results (i.e. some of the false positives will later turn out to have T1D). In a situation where the marker set is

constructed to remove as many false positive results as possible, this may result in a refined test that is over-fitted to the initial discovery group of case and control individuals, and is not reliably generalisable to other populations. It is possible that such situations would be apparent when follow-up studies on independent case/control groups for the same trait are carried out, and it is recommended that such validations are carried out before using this signature SNP set.

5.4.2 Overfitting Generates Spurious Associations

For a genetic association study to be successful, individuals must be separable into distinct groups based on a particular phenotype, and some differences between the groups must be attributable to genetic factors. Methods for identifying associated markers in a GWAS relies on a clear distinction between trait and non-trait individuals. In situations where the trait of interest is not easy to classify, an associated marker may not reflect the true distinction between those groups. In addition, a low genetic influence for the expression of a particular trait can mean that even when a trait can be classified completely, the genetic component of that trait (the only component able to be identified by any DNA marker-based method) will not always determine the observed phenotype completely.

Overfitting is the generation of a set of distinctive parameters that relies on irrelevant attributes for the model being observed. The problem exists when vital information about the model is missing, and the discovery algorithm ends up being required to derive a model based on other spurious distinctions between discovery groups (see Russell and Norvig, 2003, Chapter 14, pp. 661-663). Overfitting is applicable to the case of generating minimal marker sets because any such method assumes that a minimal set can be found for the data. When cases and controls are not genetically distinct, and distinct *only* due to the trait under test, any resultant marker set will be invalid. In such a situation, the set of markers generated is

informative only for the specific group of individuals that were used for discovery of that set of markers, and will not be applicable for individuals outside the discovery group. Internal validation within groups, and external validation of results in similar populations, is essential to ensure that overfitting has not occurred.

Bootstrap sub-sampling uses variance among group sub-samples to remove markers that are associated because of *genetic chance* effects rather than the particular phenotype under test. However, it cannot distinguish between genetic differences due to the tested phenotype and genetic differences due to sampling bias. The problem of overfitting is especially relevant for genetic data, where one pattern of genotypes due to a group-associated factor with high heritability may outweigh the disease-causing factor under test. This is similar to the population stratification problem that has been discussed by Pritchard (1999) and Pritchard and Donnelly (2001) who say that due to the influence of *genetic chance* (e.g. genetic drift, founder effects, non-random mating), alleles can appear with high frequency differences between groups within a given population sample even though the differences are not directly associated with the trait of interest. This is particularly important when a population group has a high incidence of a given disease, and the genetic history of the case and/or control subgroups is not known. Pritchard and Donnelly (2001) recommend testing for structured association in case and control groups before carrying out further association tests in order to remove confounding genetic factors that may be present in a case/control study.

5.4.2.1 Genome-wide Trait Contributions

While there may be many gene-gene interactions throughout the genome that all contribute to a particular disease, it is unlikely that *all* genetic variants in the subgroup will influence the trait. In addition, some variants may influence the trait more than others and in some cases may even negate

the effects of another variant. Both of these factors increase the potential for spurious associations and false positive results when carrying out a whole genome scan.

Genotyping carried out in an association study is restricted to a subset of the total genome, because full-genome sequencing is still prohibitively expensive. Also, only a subset of interactions between multiple genetic factors can be studied (if any), because multi-factorial analysis is computationally expensive.†

It is expected that any reduction of SNP set size will result in decreased reliability, as there is an information loss when fewer markers are typed. For a reduction method to be useful, the information lost due to typing fewer markers must be compensated by cost reduction. However, in this investigation, the opposite appears to be true – a small number of markers are useful to distinguish the case and control groups, and appear to provide more information than a full genome set.

5.4.2.2 Interactions from Multiple Genetic Variants

In some cases, a first-pass single association analysis of markers will not be useful for the classification of a trait. This will be the case for traits that have complex interactions that result in non-linear association patterns between marker frequency and trait prevalence. As an example of a complex interaction, two causative variants may interact in a neutralising fashion (i.e. the effects of one variant are cancelled out by another variant). In this sort of case, a simple one-way association test would not work as expected, retaining a lack of observed association even when there is a strong signal (Pickrell et al., 2007). Other non-linear interactions between different markers would also reduce the effectiveness of an association test to determine informative markers.

†It has an exponential complexity with respect to the number of factors studied in tandem.

The ideal situation for investigating complex traits at a genetic level is an analysis of the effectiveness of *every possible* set of marker interactions. Once such an analysis is carried out, the best set of markers will be identified as being the set that is most informative for classifying individuals into groups. However, the computational requirements for such testing combined with the increased danger of overfitting due to small cell sizes, make such an analysis effectively useless when carried out on the total marker set (see Province and Borecki, 2008).

The bootstrapping approach as outlined here does not consider combinations of genetic markers. However, it provides an efficient way to reduce a large set of markers down to a much smaller set. This smaller set can then be used by programs that determine multi-way interactions, which are typically computationally expensive procedures.

5.5 Conclusion

The application of the bootstrap sub-sampling process to marker selection is a useful complement to current GWAS. It can be used to remove potential spurious associations that are specific to the tested groups, and may help to reduce the set of individuals required for initial large-scale genotyping. Bootstrap sub-sampling acts as an internal validation of association signals, which helps to reduce the likelihood of false positive associations in publications. This, in turn, would hopefully make clinicians less likely to use these false positive associations when evaluating disease risk.

The method for identifying a minimal set of SNPs is an association-based method that discovers genome-wide combinations of SNPs for the identification of a particular trait. The method relies on a clear distinction between trait and non-trait individuals, and in situations where the trait of interest is not easy to qualify, the identified SNP set may not reflect the true distinction between those groups. In addition, the non-genetic influence

for the expression of a particular trait can mean that even when a trait can be classified completely, the genetic component of that trait (the only component able to be identified by any DNA marker-based method) will not always determine the observed phenotype.

The signature set of 5 SNPs identified here should be suitable for estimating T1D risk for screening purposes, particularly in combination with other clinical parameters, in at least the UK population. It is essential that this set be externally validated in other populations, but given reasonable validation the set may also be used for a global indicator of T1D risk.

Chapter 6

Conclusions and General Discussion

The process of research carried out in this thesis led to the development of two sets of genetic markers: one set reliably determines Maori-European admixture in New Zealand Maori populations, and another set can be used to screen for individuals who are at high risk for Type 1 Diabetes. Understanding of genetic variation in the Maori population has also progressed, from noticing reduced genetic diversity in Maori compared to other populations, to identifying some of the genetic signatures that are unique to Maori. These genetic patterns probably have clinical significance, and appear to be present throughout the genome.

At this point, it is appropriate to step back, examine the key discoveries of this thesis, and consider how these discoveries can work together to open more avenues for further investigation. The key findings from each chapter are abstracted in the next four sections (Section 6.1 to Section 6.4), and a discussion of linked themes follows (Section 6.5 onwards).

6.1 The *ADH* Gene Region

Alcohol response is a genetically influenced trait, and variation in alcohol consumption is found between New Zealand populations. Maori (especially young men) tend to drink less often but consume much higher volumes of alcohol when compared to Europeans. Chambers et al. (2002b) found that a specific variant of the *ADH1B* gene (ADH1B*47His) is associated with protection against alcohol dependence. This research was extended in Chapter 2 by typing eight additional Single Nucleotide Polymorphisms (SNPs) within the Alcohol Dehydrogenase (*ADH*) gene region in the Maori population. Substantial Linkage Disequilibrium (LD) was found at two areas within the *ADH* gene region: one near the *ADH1B* gene, and another near the *ADH4* gene.

While common and rare haplotype frequencies were found to be similar in both Maori and European populations near the *ADH4* gene, they differ near the *ADH1B*/*ADH1C* genes. This disparity demonstrates the need to consider haplotypes when investigating association at a particular genetic locus, and that SNP associations will not necessarily be consistent for different populations.

6.2 MAOA Gene Structure

The Monoamine Oxidase A (*MAOA*) gene was another candidate gene for influencing alcohol consumption behaviour. (Gilad et al., 2002) reported evidence for positive selection within the human *MAOA* gene region on chromosome X in seven populations. The original study lacked an analysis of any Polynesian populations, but consideration of migration history suggested that a similar study would be appropriate for the Maori population. A comparison of genetic variation between Maori and non-Maori populations in Chapter 3 found a substantial reduction in genetic diversity

at the *MAOA* gene locus, and an increase in the frequency of the most common *MAOA* gene variant in the Maori population. The results support the findings of Gilad et al. (2002), but also demonstrate that a 5-SNP haplotype (XGCCG) can describe the gene variant under selection, and that this variant has also undergone positive selection in the Maori population.

The controversy based around research on the *MAOA* gene (see Appendix B) led to a cessation of gene-based research for this PhD project, but also drove research towards two genome-wide analyses of genetic variation.

6.3 Maori Genomic Ancestry

In Chapter 4, a bootstrap sub-sampling method was developed to generate a set of markers for the investigation of Maori-European admixture within Rakaipaaka, a tribe of Maori from Hawkes Bay. This bootstrapping method was tested with a trait that is 100% heritable, namely New Zealand Maori genomic ancestry.

Genotype data from 30 Maori individuals and 90 European individuals were compared at 300k autosomal polymorphisms, the first genome-wide study carried out in in a Maori population. After bootstrap sub-sampling and evaluating the effectiveness of marker sets of different sizes, a validated set of 10 genetic markers was constructed that estimates individual Maori ancestral fraction with high accuracy (median = 98%, IQR = 95-99). These markers were then used to determine the variation of Maori-European ancestral fraction within Rakaipaaka, and were also used to provide an estimate of the amount of European ancestry within Rakaipaaka (28.7%).

6.4 Validation for T1D Associations

Chapter 5 develops the bootstrap sub-sampling method further by applying the method to a domain outside population genetics, a genome-wide association study for Type 1 Diabetes (T1D).

Previous literature on Type 1 Diabetes identified some DNA variants that are associated with the disease. These variants, even in combination, are not particularly informative for distinguishing between T1D cases and control individuals using genetic information alone ($AUC = 0.7284$). Population sub-sampling helped to filter out noise from genome-wide association data, and increase the chance of finding useful associative signals. Subsequent filtering based on marker linkage and testing of marker sets of different sizes produced a 5-SNP *signature set* of markers for T1D. The combination of markers used in this set, primarily from the *HLA* region on chromosome 6, is considerably more informative than previously known associated variants for predicting T1D phenotype from genetic data ($AUC = 0.8395$). Given this predictive quality, the signature set may be useful alone as a screening test for T1D, and would certainly be useful as a screening test in combination with other clinical cofactors for T1D risk.

6.5 A Combination of Different Approaches

This thesis necessarily describes two different approaches for research. Chapters 2 and 3 investigate candidate genes for influencing alcohol dependence phenotypes, planned out at the beginning of this PhD research project. Both the *ADH* gene region study and the *MAOA* study involved an analysis of additional genotyping carried out on Maori individuals as a follow-up to results in 2002 papers (Chambers et al., 2002b and Gilad et al., 2002 respectively).

The subsequent two chapters of the thesis represent research carried out after SNPchip genotyping of a Maori population (Chapter 4), and publicly available SNPchip data from a UK case-control study (Chapter 5); both these datasets were not available at the beginning of the research project, so the outcomes of these two chapters could not be predicted at the time that the thesis project began. However, it was known at the time of starting research that genome-wide data would be available about a year into study. The RHAS study was based on SNPchip genotyping that was carried out near the end of 2006, while the T1D association study was based on SNPchip data that became publicly available in 2007 (Wellcome Trust Case Control Consortium, 2007).

6.5.1 Recombination and Haplotype Block Genetics

Chapters 2 and 3 investigate the *ADH* genes and *MAOA* gene region respectively, with considerable emphasis on the haplotype block patterns observed in the Maori population. It it important to realise that most genetic variation in human populations is due to recombination (see Introduction, Section 1.2.4), and this is particularly important for recent family history (i.e. 1-5 generations back). An understanding of the physical structure of the genome (where particular genes and genetic sequences lie with respect to each other) can help to explain why particular traits, diseases, or other responses to environmental factors are more likely to be inherited together.

Both the *ADH* and *MAOA* gene studies were carried out on the same group of Maori individuals, although males were of primary interest in the *MAOA* study. Maori individuals were compared with a number of different populations within the *MAOA* gene, but compared only to the European population at the *ADH* gene region. The *MAOA* study concentrates on a single gene, while the *ADH* study considers block structure across a cassette of related genes. In the MAOA study, the common 5-SNP haplotype in Maori matched the common 5-SNP haplotype in other populations. For one

haplotype block near ADH4, the common variant in Maori was identified as the common variant in European, but for another block (near *ADH1B*, *ADH1C*), the common Maori variant matched the *rare* variant in European. Marker densities differed between the two studies, about one marker per 8k for *MAOA* (13 markers in a 90kb region), and one marker per 45k for *ADH* (7 markers in a 400kb region).

Initial genotyping and analysis carried out in Chapter 4 demonstrated large allele frequency differences between Maori and European populations for many SNPs throughout the genome (see Chapter 4, Figure 4.4). This supports the hypothesis that the two populations are different across the entire genome.

The results from Chapter 2 suggest that haplotype block patterns will be different throughout the *ADH* gene region. A more detailed analysis (i.e. higher density marker coverage) of haplotype block structure is recommended, as well as experimental tests that compare alcohol response at a haplotype level, rather than at an allele level.

Chapters 2 and 3 demonstrate that haplotype-level differences exist between Maori and European populations at two gene regions, and that these differences probably exist throughout the genome. These populations differ at a haplotype level within these regions (multiple combinations of genetic variants occurring at the same time), therefore clinical outcomes are less likely to match predictions derived from previous studies in different populations.

6.5.2 Bootstrap Sub-sampling and Internal Validation

The two bootstrap sub-sampling studies (a population study in Chapter 4, and an association study in Chapter 5) differ substantially in many ways, but they achieve similar outcomes; they each produce a small set of SNPs that can be used to describe the phenotype under test. The effectiveness of

the bootstrapping process in generating a validated set of markers, both for Maori ancestry and for Type 1 Diabetes, demonstrates the potential utility of this process for a wide range of different heritable traits.

Some differences reflect the evolution of the bootstrapping method. Delta (or allele frequency difference) was used as an association statistic in Chapter 4, but it was realised that delta failed to capture some forms of genetic differences, so a genotype χ^2 test was used in Chapter 5. The top 1000 markers were chosen in each bootstrap for investigations of Maori genomic ancestry, and the top 5% of markers were chosen for T1D associations. It was not beneficial to repeat discovery of markers in the genomic ancestry study with the revised method, because the set of markers for validation and community genotyping had been chosen (and validation populations had been genotyped) prior to these realisations. However, the genome-wide differences in allele frequencies between Maori and European populations has meant that the methodology differences made little impact on the outcome of the study.

The statistic used to evaluate the effectiveness of marker sets was also different between the two studies. A comparison of the validation plots for each study (Figure 4.9 and Figure 5.7) demonstrates why this was necessary. The Area Under the Curve (AUC) for the Maori ancestry study is effectively 1 (0.9999) because there is only minimal overlap in Q values for the two populations. However, the overlap of case and control groups for the T1D study means that the an AUC analysis can be used to compare the usefulness of different marker sets. While the false positive / true positive curve used in AUC analysis is a more descriptive result for diagnostic tests, the difference of means test is clearly more informative for populations that are distinct at a genetic level.

Other differences were necessary due to the reduced sample size of the RHAS study. The bootstrap sub-samples were limited to 25 individuals for both Maori and European populations, because it was assumed that

choosing half the Maori population for each sub-sample (i.e. 15 individuals) would be too small a group of individuals to produce informative results. However, it is unlikely that the proportion of individuals in each bootstrap sub-sample had an effect on the outcome of the study as allele frequency differences for the 59-SNP consistent set in Chapter 4 were all above 0.7 (see Figure 4.6). To guard against potential genotyping errors and misclassification of ancestry due to small discovery group sizes, more markers were included in the final SNP set for the genomic ancestry study than strictly necessary based on results (i.e. 10 SNPs rather than 6 SNPs).

While the initial marker set sizes are quite different (500k on an Affymetrix SNPchip for the T1D association study, 317k on an Illumina SNPchip for the genomic ancestry study), both platforms capture a similar amount of common variation in human populations (Ele Zeggini, Personal Communication, 2007[†]). The two final, validated marker sets cover quite different regions, and likely reflect the traits that were used to separate the groups in the two studies: a genome-wide set of markers from population groups separated by Maori/European ancestry in Chapter 4, and a marker set largely restricted to the *HLA* region for groups separated by T1D case/control status in Chapter 5.

Given the large number of T1D-informative SNPs found near the HLA region prior to linkage filtering (see Chapter 5, Figure 5.5), there is likely to be more variation that alters genetic risk for T1D within this region beyond what can be captured from SNPs alone. As such, it would be a good idea to attempt full sequencing of some parts of this region, in conjunction with validation of either this 5-SNP signature set or a slightly larger group of markers taken from the 458-SNP consistent set. This validation should be carried out for T1D case/control groups sampled from other populations (e.g. T1D studies from dbGaP) to reduce the chance of population-specific markers being identified as having good global association with T1D.

[†]See Appendix D, Section D.3

Other possible extensions of work presented in this thesis lie in the area of genetic risk calculation. The bootstrap sub-sampling studies carried out in Chapters 4 and 5 demonstrate that bootstrapping association results through population sub-sampling and linkage refinement is an effective tool for identifying useful markers linked to heritable traits. This procedure should extend to other traits with a strong genetic basis. Suitable targets may include drug response, and diseases where early intervention provides benefits.

6.5.3 Tractability and Bootstrap Sub-sampling

An issue that was largely passed over when considering bootstrap sub-sampling results is the number of bootstrap sub-samples that were used to generate a consistent set of SNPs. The number of sub-samples used here (100) is just below the range of $100 \sim 200$ recommended for classification algorithms in computer science applications over 20 years ago (Jain et al., 1987). The number of iterations used for the *structure* program[†] is typically around 100,000 for human populations (Falush et al., 2003), and 10,000 for other animal populations (Falush et al., 2007), which may better represent an accepted value of sub-samples given computing power available today. The main reason for the selection of 100 sub-samples for bootstrap sub-sampling for this research project has been one of tractability – a typical genome-wide analysis (i.e. 500k markers, 100-1000 individuals) with 100 bootstrap sub-samples using code written by the author will take about 30 hours on a present-day desktop computer. The bootstrap results file size scales linearly based on the number of sub-samples (about 1.2GB for the T1D investigation of Chapter 5), and the time taken appears to also scale linearly based on the number of sub-samples.

More bootstrap sub-samples *do* appear to reduce the size of the consistent set of markers – a re-analysis of Maori data with 1000 sub-samples

[†] *structure* does not use population sub-sampling, but it does use repeated simulation.

produced a consistent set around half that of the 100 sub-sample set (data not shown) – but the time taken outweighed any benefit gained by this in the present studies. It is likely that a more efficient bootstrap sub-sampling program will be developed in the future, permitting a greater number of sub-samples to be attempted within a reasonable time period.

6.6 The Potential for Low-cost and Informative Research

Although Chapters 4 and 5 emphasise bootstrap sub-sampling, this sampling method should not be the only tool used to identify mutations associated with traits and diseases with a genetic component. Similarly, the haplotype block approach for analysing gene structure (used in Chapters 2 and 3) should not be the only tool used to investigate gene variation and physiological effects. These tools are emphasised in this thesis as methods that are often overlooked by researchers investigating disease association and gene function.

An approach that uses both of these tools should produce useful results for a number of genes typed in the Maori population. The Maori studies in this project have been small in terms of population size (47 individuals in the studies for the *ADH* and *MAOA* gene regions, 30 individuals in the initial genome-wide study of Chapter 4), yet have led to insightful discoveries. The marker coverage for currently available genome-wide SNPchips will exceed that of both haplotype block analyses in this thesis (1M SNPs covering 3Gb, or around one SNP per 3k), so follow-up studies for at least those regions could be carried out with a similar group size and a high expectation of clinically informative results. Other genes that may be a useful target for a combined study are those that are candidate genes for diseases with different health outcomes for Maori and European populations (see Introduction, Section 1.5.5).

In conclusion, consider a more general situation in which the two approaches are combined, i.e. an internally-validated genome-wide analysis combined with a haplotype approach to investigate association. When investigating mutations and their causal relationship to disease, it may not be appropriate to study single genetic variants, as they could be linked to another variant on the same haplotype block. The investigations in this thesis that looked at haplotype block structure in Maori (Chapters 2 and 3) demonstrate that this is indeed the case, and the methods developed to supplement genome-wide association studies (Chapters 4 and 5) should help to make any future discoveries of association more robust to the genetic complexities that are inherent in human populations.

Bibliography

Abbott, W., Scragg, R. and Marbrook, J. (2001). Differences in disease frequency between Europeans and Polynesians: directions for future research into genetic risk factors. Pacific Health Dialog *8*, 129–156.

Addison, D. J. and Matisoo-Smith, E. (2010). Rethinking Polynesian origins: a West-Polynesia Triple-I Model. Archaeology in Oceania *45*, 1–12.

Aguadé, M., Miyashita, N. and Langley, C. H. (1989). Reduced Variation in the *yellow-achaete-scute* Region in Natural Populations of *Drosophila melanogaster*. Genetics *122*, 607–615.

Alcohol Advisory Council of New Zealand (2005). NZ Statistics: Per capita consumption. http://www.alac.org.nz/NZStatistic.aspx?PostingID=4346.

Alia-Klein, N., Goldstein, R. Z., Kriplani, A., Logan, J., Tomasi, D., Williams, B., Telang, F., Shumay, E., Biegon, A., Craig, I. W., Henn, F., Wang, G.-J., Volkow, N. D. and Fowler, J. S. (2008). Brain Monoamine Oxidase-A Activity Predicts Trait Aggression. The Journal of Neuroscience *28*, 5099–5104.

Anderson, A. (1991). The chronology of colonization in New Zealand. Antiquity *65*, 767–795.

Avery, P., Mousa, S. S. and Mousa, S. A. (2009). Pharmacogenomics in type

II diabetes mellitus management: Steps toward personalized medicine. Pharmacogenomics and personalized medicine *2*, 79–91.

Bach, A. W. J., Lan, N. C., Johnson, D. L., Abell, C. W., Bembenek, M. E., Kwan, S.-W., Seeburg, P. H. and Shih, J. C. (1988). cDNA cloning of human liver monoamine oxidase A and B: Molecular basis of differences in enzymatic properties. Proceedings of the National Academy of Sciences of the United States of America *85*, 4934–4938.

Barnes, H. M., McPherson, M. and Bhatta, K. (2003). Te Ao Waipiro 2000: Māori National Alcohol Survey. Technical report Whariki Research Group.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). Haploview: analysis and visualisation of LD and haplotype maps. Bioinformatics *21*, 263–265.

Bau, C. H. D., Almeida, S., Coster, F. T., Garcia, C. E. D., Elias, E. P., Ponso, A. C., Spode, A. and Hutz, M. H. (2001). DRD4 and DAT1 as modifying genes in alcoholism: interaction with novelty seeking on level of alcohol consumption. Molecular Psychiatry *6*, 7–9.

Berdis, A. J. (2009). Mechanisms of DNA polymerases. Chemical Reviews *109*, 2862–2879.

Bollenbach, T., Vetsigian, K. and Kishony, R. (2007). Evolution and multilevel optimization of the genetic code. Genome Research *17*, 401–404.

Bosron, W. F., Magnes, L. J. and Li, T.-K. (1983). Kinetic and Electrophoretic Properties of Native and Recombined Isoenzymes of Human Liver Alcohol Dehydrogenase. Biochemistry *22*, 1852–1857.

Brash, D. T. (2004). Orewa speech - nationhood. `http://www.national.org.nz/files/OrewaRotaryClub_27Jan.pdf`, accessed 2009-Dec-01.

Buetow, K. H., Edmonson, M., MacDonald, R., Clifford, R., Yip, P., Kelley, J., Little, D. P., Strausberg, R., Koester, H., Cantor, C. R. and Braun, A. (2001). High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proceedings of the National Academy of Sciences USA *98*, 581–584.

Campbell, A. M. and Heyer, L. J. (2002). Discovering Genomics, Proteomics, and Bioinformatics. 1st edition, Benjamin Cummings, San Francisco.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A. and Poulton, R. (2002). Role of Genotype in the Cycle of Violence in Maltreated Children. Science *297*, 851–854.

Chambers, G. K. (2008). Genetics and the Origins of the Polynesians. In Encyclopedia of Life Sciences. John Wiley & Sons, Ltd: Chichester. `http://www.els.net/`.

Chambers, G. K., Day, D. J., Marshall, S. J. and Robinson, G. M. (2002a). Alcohol dependence: Advances in understanding, diagnosis and treatment. New Zealand Science Review *59*, 35–41.

Chambers, G. K., Marshall, S. J., Robinson, G. M., Maguire, S., Newton-Howes, J. and Chong, N. L. (2002b). The Genetics of Alcoholism in Polynesians: Alcohol and Aldehyde Dehydrogenase Genotypes in Young Men. Alcoholism: Clinical and Experimental Research *26*, 949–955.

Charlesworth, B., Morgan, M. T. and Charlesworth, D. (1993). The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics *134*, 1289–1303.

Chen, W. J., Loh, E. W., Hsu, Y.-P. P. and Cheng, A. T. A. (1997). Alcohol Dehydrogendase and Aldehyde Dehydrogenase Genotypes and Alcoholism among Taiwanese Aborigines. Biological Psychiatry *41*, 703–709.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution *7*, 111–122.

Cloninger, C. R. (1987). Neurogenetic Adaptive Mechanisms in Alcoholism. Science *236*, 410–416.

Condit, C. M. (2007). How geneticists can help reporters to get their story right. Nature Reviews Genetics *8*, 815–820.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. and Foldit Players (2010). Predicting protein structures with a multiplayer online game. Nature *466*, 756–760.

Crabb, D. W., Matsumoto, M., Chang, D. and You, M. (2004). Overview of the role of alcohol dehydrogenase and aldehyde dehydrogenase and their variants in the genesis of alcohol-related pathology. Proceedings of the Nutrition Society *63*, 49–63.

Crampton, P. and Parkin, C. (2007). Warrior genes and risk-taking science. The New Zealand Medical Journal *120*, 63–65.

Daneman, D. (2006). Type 1 Diabetes. The Lancet *367*, 847–858.

Dermitzakis, E. T. and Clark, A. G. (2009). Life After GWA Studies. Science *326*, 239–240.

Devor, E. J. (1993). Why There Is No Gene for Alcoholism. Behavior Genetics *23*, 145–151.

Di Piazza, A., Di Piazza, P. and Pearthree, E. (2007). Sailing virtual canoes across Oceania: revisiting island accessibility. Journal of Archaeological Science *34*, 1219–1225.

Dick, D. M. and Foroud, T. (2003). Candidate Genes for Alcohol Dependence: A Review of Genetic Evidence From Human Studies. Alcoholism: Clinical and Experimental Research *27*, 868–879.

Dixon, H. B. F., Cornish-Bowden, A., Liébecq, C., Loening, K. L., Moss, G. P., Reedijk, J., Velick, S. F. and Vliegenthart, J. F. G. (1984). Nomenclature and Symbolism for Amino Acids and Peptides: Recommendations 1983. Biochemical Journal *219*, 345–373.

Djoussé, L., Levy, D., Herbert, A. G., Wilson, P. W., D'Agostino, R. B., Cupples, L. A., Karamohamed, S. and Ellison, R. C. (2005). Influence of Alcohol Dehydrogenase 1C Polymorphism on the Alcohol-Cardiovascular Disease Association from the Framingham Offspring Study. The American Journal of Cardiology *96*, 227–232.

Donohue, M. and Denham, T. (2010). Farming and Language in Island Southeast Asia. Current Anthropology *51*, 223–256.

Du, F.-X., Clutter, A. C. and Lohuis, M. M. (2007). Characterizing Linkage Disequilibrium in Pig Populations. International Journal of Biological Sciences *3*, 166–178.

Eckert, K. A. and Hile, S. E. (2009). Every Microsatellite is Different: Intrinsic DNA Features Dictate Mutagenesis of Common Microsatellites Present in the Human Genome. Molecular Carcinogenesis *48*, 379–388.

Edenberg, H. J., Xuei, X., Chen, H.-J., Tian, H., Wetherill, L. F., Dick, D. M., Almasy, L., Bierut, L., Bucholz, K. K., Goate, A., Hesselbrock, V., Kuperman, S., Nurnberger, J., Porjesz, B., Rice, J., Schuckit, M., Tischfield, J., Begleiter, H. and Foroud, T. (2006). Association of alcohol dehydrogenase genes with alcohol dependence: a comprehensive analysis. Human Molecular Genetics *15*, 1539–1549.

Ekoé, J.-M., Zimmet, P. and Williams, R., eds (2002). The Epidemiology of Diabetes Mellitus An International Perspective. John Wiley & Sons, West Sussex, England. 2002 reprint.

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*, 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*, 574–578.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters *27*, 861–874.

Fay, J. and Wu, C. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

Fitzgerald, J. C., Ufer, C., Girolamo, L. A. D., Kuhn, H. and Billett, E. E. (2007). Monoamine oxidase-A modulates apoptotic cell death induced by staurosporine in human neuroblastoma cells. Journal of Neurochemistry *103*, 2189–2199.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., Altshuler, D. M., Aburatani, H., Jones, K. W., Typer-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W. and Lee, C. (2006). Copy number variation: New insights in genome diversity. Genome Research *16*, 949–961.

Friedlaender, J. S., Fran c. R. F., Reed, F. A., Kidd, K. K., Kidd, J. R., Chambers, G. K., Lea, R. A., Loo, J.-H., Koki, G., Hodgson, J. A., Merriwether, D. A. and Weber, J. L. (2008). The Genetic Structure of Pacific Islanders. PLoS Genetics *4*, e19.

Frudakis, T. N. (2008). Molecular Photofitting: Predicting Ancestry and Phenotype Using DNA chapter 2, p. 35. London, UK: Academic Press.

Fu, Y.-X. and Li, W.-H. (1993). Statistical Tests of Neutrality of Mutations. Genetics , *133*, 693–709.

Fung, H.-C., Scholz, S., Matarin, M., Simón-Sánchez, J., Hernandez, D., Britton, A., Gibbs, J. R., Langefeld, C., Stiegert, M. L., Schymick, J.,

Okun, M. S., Mandel, R. J., Fernandez, H. H., Foote, K. D., Rodrguez, R. L., Peckham, E., Vrieze, F. W. D., Gwinn-Hardy, K., Hardy, J. A. and Singleton, A. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. The Lancet Neurology , *5*, 911–916.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggert, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. Science , *296*, 2225–2229.

Gibbons, A. (2004). Tracking the Evolutionary History of a "Warrior" Gene. Science , *304*, 818.

Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. and Skorecki, K. (2002). Evidence for positive selection and population structure at the human MAO-A gene. PNAS , *99*, 862–867.

Glover, V., Sandler, M., Owen, F. and Riley, G. J. (1977). Dopamine is a monoamine oxidase B substrate in man. Nature , *265*, 80–81.

Gokturk, C., Schultze, S., Nilsson, K. W., von Knorring, L., Oreland, L. and Hallman, J. (2008). Serotonin transporter (5-HTTLPR) and monoamine oxidase (MAOA) promoter polymorphisms in women with severe alcoholism. Archives of Women's Mental Health , *11*, 347–355.

Grimsby, J., Zentner, M. and Shih, J. C. (1996). Identification of a region important for human monoamine oxidase b substrate and inhibitor selectivity. Life Sciences , *58*, 777–787.

Guéguen, N., Jacob, C., Guellec, H. L., Morineau, T. and Lourel, M. (2008). Sound Level of Environmental Music and Drinking Behavior: A Field Experiment With Beer Drinkers. Alcoholism: Clinical and Experimental Research , *32*, 1–4.

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. and Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nature Genetics , *37*, 549–554.

Guo, G., Roettger, M. E. and Cai, T. (2008). The Ingegration of Genetic Propensitites into Social-Control Models of Delinquency and Violence among Male Youths. American Sociological Review , *73*, 543–568.

Hall, D. A. (2004). Genetic Analysis of Polynesian Populations. Honours thesis, School of Biological Sciences. `http://gringer.org/honours/dah_honours_thesis.pdf`.

Hall, D. A., Lea, R. A. and Chambers, G. K. (2007). Haplotype analysis at the alcohol dehydrogenase gene region in New Zealand Maori. Journal of Human Genetics , *52*, 191–194.

Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., Kidd, J. R. and Kidd, K. K. (2007). Evidence of Positive Selection on a Class I ADH Locus. The American Journal of Human Genetics , *80*, 441–456.

Hasin, D. (2003). Classification of Alcohol Use Disorders. Alcohol Research & Health , *27*, 5–17.

Healy, D. G. (2006). Case-control studies in the genomic era: a clinician's guide. The Lancet Neurology , *5*, 701–707.

Hedrick, P. and Kumar, S. (2001). Mutation and linkage disequilibrium in human mtDNA. European Journal of Human Genetics , *19*, 969–972.

Helzer, J. E., Bucholz, K. K., Bierut, L. J., Regier, D. A., Schuckit, M. A. and Guth, S. E. (2006). Should DSM-V Include Dimensional Diagnostic Criteria for Alcohol Use Disorders. Alcoholism: Clinical and Experimental Research , *30*, 303–310.

Higuchi, S., Matsushita, S., Masaki, T., Yokoyama, A., Kimura, M., Suzuki, G. and Mochizuki, H. (2004). Influence of Genetic Variations of Ethanol-Metabolizing Enzymes on Phenotypes of Alcohol-Related Disorders. Annals New York Academy of Sciences , *1025*, 472–480.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. and Cox, D. R. (2005). Whole-Genome Patterns of Common DNA Variation in Three Human Populations. Science , *307*, 1072–1079.

Hopkin, M. (2008). 'ruthlessness gene' discovered. [Published online, `http://www.nature.com/news/2008/080404/full/news.2008.738.html`, doi:10.1038/news.2008.738].

Hsu, Y.-P. P., Loh, E. W., Chen, W. J., Chen, C.-C., Yu, J.-M. and Cheng, A. T. A. (1996). Association of Monoamine Oxidase A Alleles with Alcoholism Among Male Chinese in Taiwan. American Journal of Psychiatry , *153*, 1209–1211.

Huang, H., Shiffman, M. L., Friedman, S., Venkatesh, R., Bzowej, N., Abar, O. T., Rowland, C. M., Catanese, J. J., Leong, D. U., Sninsky, J. J., Layden, T. J., Wright, T. L., White, T. and Cheung, R. C. (2007). A 7 Gene Signature Identifies the Risk of Developing Cirrhosis in Patients with Chronic Hepatitis C. Hepatology , *46*, 297–306.

Hurles, M. E., Matisoo-Smith, E., Gray, R. D. and Penny, D. (2003). Untangling Oceanic settlement: the edge of the knowable. Trends in Ecology and Evolution , *18*, 531–540.

Hurles, M. E., Nicholson, J., Bosch, E., Renfrew, C., Sykes, B. C. and Jobling, M. A. (2002). Y Chromosomal Evidence for the Origins of Oceanio-Speaking Peoples. Genetics , *160*, 289–303.

Hutt, M. (2003). Te Iwi Maori me te Inu Waipiro: He Tuhituhinga Hitori – Maori and Alcohol: A history. 2 edition, Health Services Research

centre for Kaunihera Whakatupato Waipiro o Aotearoa, The Printing Press, Wellington.

Hutter, S., Vilella, A. J. and Rozas, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics *7*, 409.

Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M. and Tuomilehto, J. (2003). Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs: A Nationwide Follow-Up Study. Diabetes *52*, 1052–1055.

Ibanez, A., de Castro, I. P., Fernandez-Piqueras, J., Blanco, C. and Saiz-Ruiz, J. (2000). Pathological gambling and DNA polymorphic markers at MAO-A and MAO-B genes. Molecular Psychiatry *5*, 105–109.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Irwin, G., Bickler, S. and Quirke, P. (1990). Voyaging by canoe and computer: experiments in the settlement of the Pacific Ocean. Antiquity *64*, 34–50.

Jacob, C. P., Müller, J., Schmidt, M., Hohenberger, K., Gutknecht, L., Reif, A., Schmidtke, A., Mössner, R. and Lesch, K. P. (2005). Cluster B Personality Disorders are Assocoated with Allelic Variation of Monoamine Oxidase A Activity. Neuropsychopharmacology *30*, 1711–1818.

Jain, A. K., Dubes, R. C. and Chen, C.-C. (1987). Bootstrap Techniques for Error Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence *PAMI-9*, 628–633.

Jin, Y., Chen, D., Hu, Y., Guo, S., Sun, H., Lu, A., Zhang, X. and Li, L. (2006). Association between monoamine oxidase gene polymorphisms and smoking behaviour in Chinese males. International Journal of Neuropsychopharmacology *9*, 557–564.

Jordan, F. M., Gray, R. D., Greenhill, S. J. and Mace, R. (2009). Matrilocal residence is ancestral in Austronesian societies. Proceedings of the Royal Society: Biological Sciences *276*, 1957–1964.

Kathiresan, S., Manning, A. K., Demisse, S., D'Agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., Arnett, D. K., Peloso, G. M., Ordovas, J. M. and Cupples, L. A. (2007). A genome-wide association study for blood lipid phenotypes in the Framington Heart Study. BMC Medical Genetics *8*, S17.

Kayser, M. (2010). The Human Genetic History of Oceania: Near and Remote Views of Dispersal. Current Biology *20*, R194–R201.

Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L. A., Moyse-Faurie, C., Rutledge, R. B., Scheifenhoevel, W., Gil, D., Lin, A. A., Underhill, P. A., Oefner, P. J., Trent, R. J. and Stoneking, M. (2006a). Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Molecular biology and Evolution *23*, 2234–2244.

Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L. A., Moyse-Faurie, C., Rutledge, R. B., Schiefenhoevel, W., Gil, D., Lin, A. A., Underhill, P. A., Oefner, P. J., Trent, R. J. and Stoneking, M. (2006b). Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific. Molecular Biology and Evolution *23*, 2234–2244.

Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J., Whittaker, P., Collins, A., Morris, A. P., Bentley, D., Cardon, L. R. and Deloukas,

P. (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. Human Molecular Genetics *13*, 577–588.

Kim, S. and Misra, A. (2007). SNP Genotyping: Technologies and Biomedical Applications. Annual Review of Biomedical Engineering *9*, 289–320.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V. and Gottesman, M. M. (2007). A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science *315*, 525–528.

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. Japanese Journal of Genetics *66*, 367–386.

Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M. and Tokunaga, K. (2008). Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. Molecular Biology and Evolution *25*, 1750–1761.

Kirby, D. A. (2000). The New Eugenics in Cinema: Genetic Determinism and Gene Therapy in "GATTACA". Science Fiction Studies *27*, 193–215.

Koenig, B. A., Lee, S. S.-J. and Richardson, S. S., eds (2008). Revisiting Race in a Genomic Age. Rutgers University Press, London.

Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. PNAS *99*, 803–808.

Lea, R. A., Benowitz, N. and Griffiths, L. R. (2006). Pharmacogenetics of Nicotine Replacement Therapy in New Zealand. In International Congress of Human Genetics The American Society of Human Genetics. Free Paper #0352.

Lea, R. A. and Chambers, G. K. (2007a). Monoamine oxidase, addiction, and the "warrior" gene hypothesis. The New Zealand Medical Journal *120*, 5–10.

Lea, R. A. and Chambers, G. K. (2007b). Pharmacogenetics in Admixed Polynesian Populations. In Pharmacogenomics in Admixed Populations, (Suarez-Kurtz, G., ed.), chapter 11, pp. 164–179. Landes Bioscience Austin, Texas.

Lea, R. A., Hall, D. A., Chambers, G. K. and Griffiths, L. R. (2006). Tracking the Evolutionary History of the Warrior Gene Across the South Pacific. In International Congress of Human Genetics The American Society of Human Genetics. Poster #1329.

Lee, S.-L., Höög, J.-O. and Yin, S.-J. (2004). Functionality of allelic variations in human alcohol dehydrogenase gene family: assessment of a functional window for protection against alcoholism. Pharmacogenetics *14*, 725–732.

Lewis, A., Miller, J. H. and Lea, R. A. (2007). Monoamine oxidase and tobacco dependence. Neurotoxicology *28*, 182–195.

Liu, I.-C., Blacker, D. L., Xu, R., Fitzmaurice, G., Tsuang, M. T. and Lyons, M. J. (2004). Genetic and environmental contributions to age of onset of alcohol dependence symptoms in male twins. Addiction *99*, 1403–1409.

Maclean, K. (2006). Humour of gene names lost in translation to patients. Nature *439*, 266.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics *37*, 413–417.

Marshall, S. J., Whyte, A. L. H., Hamilton, J. F. and Chambers, G. K. (2005). Austronesian prehistory and Polynesian genetics: A molecular view of human migration across the Pacific. New Zealand Science Review *62*, 75–80.

Mathew, C. G. (2008). New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nature Reviews Genetics *9*, 9–14.

Mattick, J. S. (2007). A new paradigm for developmental biology. The Journal of Experimental Biology *210*, 1526–1547.

McKinnon, M., ed. (2003). Bateman New Zealand Historical Atlas: Ko Papatuanuku e Takoto Nei. David Bateman Ltd, Auckland, New Zealand.

McLauchlan, G., ed. (1984). Bateman New Zealand Encyclopedia. David Bateman Ltd, Auckland, New Zealand.

McMillen, P., Kalafatelis, E. and de Bonnaire, C. (2004). The Way We Drink: The current attitudes & behaviours of New Zealanders (aged 12 plus) towards drinking alcohol. Technical report Alcohol Advisory Council.

Merriman, T. and Cameron, V. (2007). Risk-taking: behind the warrior gene story. The New Zealand Medical Journal *120*, 59–62.

Messas, G. P. and Filho, H. P. V. (2004). The role of genetics in alcohol dependence. Revista Brasileira de Psiquiatria *26*, 54–58.

Morell, V. (1993). Evidence Found for a Possible 'Aggression Gene'. Science *260*, 1722–1723.

Moss, H. B., Chen, C. M. and Yi, H. (2007). Subtypes of alcohol dependence in a nationally representative sample. Drug and Alcohol Dependence *91*, 149–158.

Murray-McIntosh, R. P., Scrimshaw, B. J., Hatfield, P. J. and Penny, D. (1998). Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. Proceedings of the National Academy of Sciences USA *95*, 9047–9052.

Nagatsu, T. (2004). Progress in Monoamine Oxidase (MAO) Research in Relation to Genetic Engineering. Neurotoxicology *25*, 11–20.

Neale, B. M. and Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. American Journal of Human Genetics *75*, 353–362.

Nei, M. and Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences USA *76*, 5269–5273.

Nordquist, N. and Oreland, L. (2006). Monoallelic Expression of MAOA in skin fibroblasts. Biochemical and Biophysical Research Communications *348*, 763–767.

Nurnberger, J. I. and Bierut, L. J. (2007). Seeing the Connections: Alcoholism and our Genes. Scientific American *296*, 46–53.

Osier, M. V., Pakstis, A. J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L. O., Bertranpetit, J., Bonne-Tamir, B., Lu, R. B., Kidd, J. R. and Kidd, K. K. (2002). A global perspective on genetic variation at the ADH genes reveals unusual pattern of linkage disequilibrium and diversity. American Journal of Human Genetics *71*, 84–99.

Pearson, P. L. (2006). Historical development of analysing large-scale changes in the human genome. Cytogenetic and Genome Research *115*, 198–204.

Pickrell, J., Clerget-Darpoux, F. and Bourgain, C. (2007). Power of Genome-Wide Association Studies in the Presence of Interacting Loci. Genetic Epidemiology *31*, 748–762.

Prescott, C. A., Caldwell, C. B., Carey, G., Vogler, G. P., Trumbetta, S. L. and Gottesman, I. I. (2005). The Washington University Twin Study of Alcoh-

olism. American Journal of Medical Genetics Part B (Neuropsychiatric Genetics) *134B*, 48–55.

Pritchard, J. K. (1999). Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. American Journal of Human Genetics *65*, 220–228.

Pritchard, J. K. and Donnelly, P. (2001). Case-Control Studies of Association in Structured or Admixed Populations. Theoretical Population Biology *60*, 227–237.

Pritchard, J. K., Stevens, M. and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics *155*, 945–959.

Province, M. A. and Borecki, I. B. (2008). Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. In Pacific Symposium on Biocomputing pp. 190–200, Pacific Symposium on Biocomputing.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

Rajeevan, H., Osier, M. V., Cheung, K. H., Deng, H., Druskin, L., Heinzen, R., Kidd, J. R., Stein, S., Pakstis, A. J., Tosches, N. P., Yeh, C. C., Miller, P. L. and Kidd, K. K. (2003). ALFRED – the ALlele FREquency Database – update. Nucleic Acids Research *31*, 270–271.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes *4*, 137–138.

Rosenberg, N. A., Li, L. M., Ward, R. and Pritchard, J. K. (2003). Informativeness of Genetic Markers for Inference of Ancestry. American Journal of Human Genetics *73*, 1402–1422.

Rossaak, M. and Pitto, R. P. (2005). Osteomyelitis in Polynesian children. International Orthopaedics *29*, 55–58.

Russell, P. J., ed. (1998). Genetics. 5th edition, Benjamin/Cummings, California, USA.

Russell, S. J. and Norvig, P. (2003). Artificial Intelligence - A Modern Approach. 2nd edition, Prentice Hall.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwaitkowski, D., Ward, R. and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Sabol, S. Z., Hu, S. and Hamer, D. (1998). A functional polymorphism in the monoamine oxidase A gene promoter. Human Genetics *103*, 273–279.

Saccone, N. L., Kwon, J. M., Corbett, J., Goate, A., Rochberg, N., Edenberg, H. J., Foroud, T., Li, T.-K., Begleiter, H., Reich, T. and Rice, J. P. (2000). A Genome Screen of Maximum Number of Drinks as an Alcoholism Phenotype. American Journal of Medical Genetics (Neuropsychiatric Genetics) *96*, 632–637.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mulkin, J. C., Mortimore, B. J., Wiley, D. L., Hunt, S. E., Cole, C. G., Coggil, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, K. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. and Altshuler, D. (2001). A

map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.

Sarfati, D. and Scott, K., eds (1999). Taking the Pulse: The 1996/97 New Zealand Health Survey chapter 5, pp. 69–86. Wellington, New Zealand: Ministry of Health.

Schuckit, M. A. (2009). An overview of genetic influences in alcoholism. Journal of Substance Abuse Treatment , *36*, S5–14.

Shand, B., Elder, P., Scott, R., Poa, N. and Frampton, C. M. (2007). Comparison of plasma adiponectin levels in New Zealand Maori and Caucasian individuals. The New Zealand Medical Journal , *120*, U2606.

Shepherd, C., Harbison, S. and Vintiner, J. (2004). Y STR haplotype data for New Zealand population groups using the Y-Plex 6 kit. Forensic Science International , *145*, 69–72.

Shih, J. C., Chen, K. and Ridd, M. J. (1999). MONOAMINE OXIDASE: From Genes to Behaviour. Annual Review of Neuroscience , *22*, 197–217.

Shriver, M. D., Jin, L., Chakraborty, R. and Boerwinkle, E. (1993). VNTR Allele Frequency Distributions Under the Stepwise Mutation Model: A Computer Simulation Approach. Genetics , *134*, 983–993.

Slack, A., Nana, G., Webster, M., Stokes, F. and Wu, J. (2009). Costs of Harmful Alcohol and Other Drug Use. Technical report, Business and Economic Research Limited, BERL House, Wellington, New Zealand.

Smith, T. F. and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. Journal of Molecular Biology , *147*, 195–197.

Son, S.-Y., Ma, J., Kondou, Y., Yoshimura, M., Yamashita, E. and Tsukihara, T. (2008). Structure of human monoamine oxidase A at 2.2-Å resolution: The control of opening the entry for substrates/inhibitors. PNAS , *105*, 5739–5744.

Spriggs, M. (2010). Commentary. In Farming and Language in Island Southeast Asia, vol. 51,, p. 245. Current Anthropology.

Sun Microsystems (2009). OpenOffice.org: the free and open productivity suite. http://www.openoffice.org/.

Sundborn, G., Metcalf, P., Scragg, R., Schaaf, D., Dyall, L., Gentles, D., Black, P. and Jackson, R. (2007). Ethnic differences in the prevalence of new and known diabetes mellitus, impaired glucose tolerance and impaired fasting glucose. Diabetes Heart and Health Survey (DHAH) 2002-2003, Auckland New Zealand. The New Zealand Medical Journal , *120*, U2607.

Tajima, F. (1989a). The effect of change in population size on DNA polymorphism. Genetics , *123*, 597–601.

Tajima, F. (1989b). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics , *123*, 585–595.

The Dominion Post (2006). Maori violence blamed on gene. The Dominion Post, Wellington, New Zealand. August 9, Section A4.

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H. M., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A. C., Ovington, N. R., Allen, J., Adlem, E., Leung, H.-T., Wallace, C., Howson, J. M. M., Guja, C., Ionescu-Tirgoviste, C., GET1FIN, Simmonds, M. J., Heward, J. M., Gough, S. C., Consortium, T. W. T. C. C., Dunger, D. B., Wicker, L. S. and Clayton, D. G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nature Genetics , *39*, 857–864.

Underhill, P. A., Passarino, G., Lin, A. A., Marzuki, S., Oefner, P. J., Cavalli-Sforza, L. L. and Chambers, G. K. (2001). Maori Origins, Y-chromosome

Haplotypes and Implications for Human History in the Pacific. Human Mutation , *17*, 271–280.

Vayda, A. P. (1970). Maoris and Muskets in New Zealand: Disruption of a War System. Political Science Quarterly , *85*, 560–584.

Walker, W. (2008). Moumoukai and Ngāti Rākaipaaka. Ngā Maunga Kōrero o Te Tairāwhiti (issue 17), The Gisborne Herald, Gisborne, New Zealand.

Wall, J. D. and Pritchard, J. K. (2003). Assessing the Performance of the Haplotype Block Model of Linkage Disequilibrium. American Journal of Human Genetics , *73*, 502–515.

Walsh, S. J., Triggs, C. M., Curran, J. M., Cullen, J. R. and Buckleton, J. S. (2003). Evidence in Support of Self-Declaration as as Sampling Method for the Formation of Sub-Population DNA Databases. Journal of Forensic Science , *48*, 1091–1093.

Wang, E., Ding, Y.-C., Flodman, P., Kidd, J. R., Kidd, K. K., Grady, D. L., Ryder, O. A., Spence, M. A., Swanson, J. M. and Moyzis, R. K. (2004). The Genetic Architecture of Selection at the Human Dopamine Receptor D4 (DRD4) Gene Locus. American Journal of Human Genetics , *74*, 931–944.

Wang, E. T., Baldi, P. and Moyzis, R. K. (2006). Darwin's Fingerprint: Accelerated Recent Adaptive Evolution in Humans. In International Congress of Human Genetics The American Society of Human Genetics. Free paper #0558.

Wang, J., Hinrichs, A., Stock, H., Budde, J., Allen, R., Bertelsen, S., Kwon, J., Wu, W., Dick, D., Rice, J., Jones, K., Nurnberger, J., Tischfield, J., Porjesz, B., Edenberg, H., Hesselbrock, V., Crowe, R., Schuckit, M., Begleiter, H., Reich, T., Goate, A. and Bierut, L. (2004). Evidence of common and

specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. Human Molecular Genetics *13*, 1903–1911.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. and Jin, L. (2002). Distribution of Reconbination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. American Journal of Human Genetics *71*, 1227–1234.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

Whaanga, J. (2008). Rhas background. Health and Ancestry Study Seminar, ESR.

Whittle, P. M. (2010). Health, inequality and the politics of genes. The New Zealand Medical Journal *123*, 67–75.

Whyte, A. L., Marshall, S. J. and Chambers, G. K. (2005). Human evolution in Polynesia. Human Biology *77*, 157–177.

Zhao, H. X., Stenhouse, E., Sanderson, E., Sopert, C., Hughest, P., Cross, D., Demaine, A. G. and Millward, B. A. (2003). Continued rising trend of childhood Type 1 diabetes mellitus in Devon and Cornwall, England. Diabetic Medicine *20*, 168–170.

Zweig, M. H. and Campbell, G. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry *39*, 561–577.

Appendix A

ADH Paper

A study of the alcohol dehydrogenase genes (see Chapter 2) was printed in the Journal of Human Genetics in February 2007 (Hall et al., 2007). The print version of this thesis reproduces the paper here, but this has been removed from the electronic version at the request of the journal editors.

Appendix B

MAOA Controversy

On Thursday 10th August 2006, while at the International Congress of Human Genetics (ICHG) the author was made aware of some of his research on the *MAOA* gene being reported in a daily newspaper in New Zealand (The Dominion Post, 2006). His research supervisor, Dr. Rod Lea, spent the rest of the day talking to a number of different media groups in New Zealand, trying to provide the public with a more balanced account of their research group's *MAOA* gene research.

The media reaction to Dr. Lea was surprising, considering that he had not mentioned the prevalence of *MAOA* gene variants except as a brief aside during his Tuesday, August 8th talk on nicotine replacement therapy at ICHG (Lea et al., 2006). It was after that talk that he had been initially approached by the media about his research. Most of the media attention, however, seemed to be due to the abstract of a poster at the same conference, with a title of "Tracking the Evolutionary History of the Warrior Gene Across the South Pacific" (Lea et al., 2006).

The low-expression 3-repeat variant of *MAOA*-uVNTR was labelled as a "Warrior gene" during the American Association of Physical Anthropologists Meeting (Gibbons, 2004), a label that was used in this poster, and also by Dr. Lea during his talk at the Congress. Despite the bad experiences in

the scientific community in labelling genes or variants with seemingly informative (at the time of discovery) or humourous names (see Maclean, 2006), this practice still continues. In fact, a genetic variant of AVPR1a has been labelled in a Nature news article as a "ruthlessness gene" (Hopkin, 2008), which evokes similar images to that of the Warrior gene label attached to the *MAOA* promoter variant.

The discussion time for the *MAOA* poster was about 4 hours after Dr Lea's talk, and at that time the author (being the researcher who generated the presented data) was standing next to the poster, ready to talk about the results of that research. However, he was not knowingly approached by anyone from the media about his results, which would be expected if the media concern had actually been about this research.

The controversy over the research entered the New Zealand public domain in the Dominion Post (the local newspaper for Wellington, New Zealand) on the 9thAugust with a rather incendiary title, "Maori violence blamed on gene" (The Dominion Post, 2006). Dr. Lea accepted an offer of an interview on Campbell Live, a New Zealand current-affairs news show, during which he was able to take some of the heat out of the debate.[†] Responses from other scientists followed, with Merriman and Cameron (2007) and Crampton and Parkin (2007) expressing their viewpoints in the March 2nd, 2007 edition of The New Zealand Medical Journal. While these responses were valid in rejecting a direct causative link between *MAOA* variants and aggressive behaviour, they did not appear to consider the role that the media had played in getting this suggestion of a causative link into the public arena. Lea and Chambers (2007a) reported preliminary results of this *MAOA* gene research in the same journal, together with some comments on the responses of Merriman and Cameron (2007) and Crampton and Parkin (2007). Discussion on this issue still continues (see Whittle, 2010).

[†]`http://www.3news.co.nz/TVShows/CampbellLive/Stories/tabid/817/articleID/12000/cat/100/Default.aspx`

B.1 Cultural Selection in Human Populations

Evidence for recent selection in the human genome was presented at the last International Congress of Human Genetics (Wang et al., 2006). In his talk, Robert Moyzis speculated that a developing culture may shape the genome. Comparing the recent evolution of the human genome with results from corn and other domesticated species introduces the possibility that we have "accidentally" domesticated ourselves.

His theory was based on a large number of variations in the human genome that were in quite a high frequency – in particular, a high frequency of common bad diseases. These frequencies are not predicted by neutral theory, and appear to be driven because of natural selection, probably happening about 40,000 years ago. An LDD test was used (outlined in Wang et al., 2004) that identified about 2500 sites that exhibited a selection pattern in the human genome, a much greater number than had been expected. The selective sites were mostly related to pathogen response, neuronal function, the cell cycle and DNA function.

These mutations, surprisingly, are still substantially polymorphic in the human population. It was noted that even a modest selection coefficient of 5% (i.e. individuals with a selected-for variant produce 5% more grandchildren than those without the variant) should drive a variant to fixation over about 200 generations. Looking at the population growth curve, it was suggested that the population was expanding faster than what was required in order for these variants to become fixed – exponential population growth has meant that fixation of alleles is less likely to occur.

B.2 The case for Maori Cultural selection

It is reasonable then to assume that there has been some cultural selection in the Maori population. Only the bravest among the Polynesian population

would have decided to take the initial long journeys in large canoes across the Pacific, and this inclination to take risks would have been fortified by land wars and revenge battles after settlement and population expansion in New Zealand (Whittle, 2010). This cultural selection of risk-takers may have had an effect on the genetic structure of the Maori population, and as a result, produce some selective signals that can be observed in the genome.

Considering the limited founding population size and geographic isolation, the genomic patterns of the Maori population will likely display distinctive haplotype signatures. In addition, haplotype blocks should extend over larger genetic distances and have markedly reduced diversity compared to other human populations. It is expected that selective signals in the genome will be able to be detected more easily due to these factors.

B.3 *MAOA* And Risk-taking Behaviour

Caspi et al. (2002) investigated whether a repeat polymorphism at *MAOA*-uVNTR (which alters *MAOA* expression) could predict if maltreatment during childhood was carried through to antisocial behaviour at age 26. It was found that individuals with severe maltreatment and high *MAOA* activity were able to regulate their own behaviour in a similar fashion to those with no maltreatment during childhood, but this capability of moderation was not present in individuals with low *MAOA* activity. A separate study has shown that brain *MAOA* activity correlates inversely with aggression levels (Alia-Klein et al., 2008), which supports the aggression association with low expression variants of the promoter region. The association of this region with aggressive, risk-taking behaviour indicates a plausible candidate for a gene region that has been under selection during the risky Polynesian voyages and later land wars (Whittle, 2010).

An alternative representation of the results can be considered, namely that individuals with low *MAOA* activity are more likely to continue the

customs and behaviour of their parents. This hypothesis removes the emphasis of aggressive behaviour in association with this gene, and instead considers how individuals may respond to learning and reward pathways. However, any association analyses must be interpreted with caution. Genetic contributions can never be treated as absolute, as discussed previously in Chapter 3, Section 2.2.4. Both Caspi et al. (2002) and Guo et al. (2008) point out some complexities of gene-environment interactions for *MAOA*, noting that associations can be interpreted differently depending on the environmental interactions being tested.

Appendix C

ICHG

The International Congress of Human Genetics (ICHG) 2006 was based in Brisbane, Australia, and lasted from 6thAugust to 10thAugust. There were about 1500-2000 people attending the Congress, which was held at the Brisbane Convention and Exhibition Centre. I attended 60 oral presentations, presented two posters, and was listed as a contributing researcher on two other posters.

The congress only happens once every five years, so it was great to be able to have the opportunity to attend this large conference in my first year of study. What I got out of ICHG was more of an understanding of Human Genetics as a very broad subject area, and some confidence that the research I am doing for my PhD is novel, useful, and interesting to others. I chose to go to a number of seminars that were outside my area of study, and in many cases was able to find some way to approach problems that would be useful in my own research. In addition to the attendance at talks, I was also introduced to a number of people who have been working with Rod Lea, Geoff Chambers, and myself over the course of my PhD project. Overall, the conference has helped me to get an idea of where my research lies with respect to other human genetic research, and to develop links with people who I can advance my career with in the future.

If I had only attended talks that were within the range of what I was studying, then I would likely have missed out on the talks about education. I sneaked into the last few minutes of Peter Farndon's talk, and was disappointed that I had not dropped by earlier for that. He likened the current method of teaching genetics to being given a 747 flight manual at the beginning of a flight, and being expected to understand the mechanics of flight in order to be able to understand the in-flight speed indicator. In other words, he was suggesting that we don't consider the population able to understand genetics until they have been taught everything we consider as necessary before entering research, even though there is a much smaller subset of key concepts needed for this understanding. MaryAnne Aitken said that we should never start teaching genetics with Mendel's peas, something that has very little relevance in today's understanding of genetics. She said that the students may have a knowledge of the terms through the media and TV programs such as CSI, but don't have an understanding of the underlying meaning of these terms. Joseph McInerney suggested that viewing every student as a potential bench scientist was not necessarily the best approach for teaching. Joseph also pointed out that graduate students are the worst at teaching to a high-school level because they have very specific learning, and do not tend to study outside their own course.

Genomic structure and structural variation was a common theme in many of the talks that I attended, one of the things that the current human genome sequence is not able to show. Ewan Eichler noted that most approaches cannot detect inversions or novel sequences. His group checked one person against the reference sequence, choosing 300 sites at random that had structural differences, and found that 2% of these sites were clinically relevant. He also noted that a 5Mbp region near chromosome 1 was a structural variation hotspot. Stephen Scherer highlighted that no single technology will capture all genome variance – one example of this is that SNP tests are not typically able to detect DNA duplications. Nigel Carter compared a BAC assay with an Affymetrix 500k SNP chip, and showed that

there were variants throughout the genome that were only picked up by one of these methods. The BAC assay was better at detecting duplications and complex variations, while the Affymetrix array was good at detecting deletions.

Appendix D

HapMap Training Courses

During the first week of of April, 2007, a Wellcome Trust course was held, Working With The HapMap. The course that I attended was held at the Sanger Centre in Hinxton, Cambridge (UK). There were 30 participants, coming from all over the world. It was an intensive, 4-day course that introduced participants to the HapMap project and a number of different genetic analyses that could be carried out using the publicly available data. All of the software demonstrated during the course is freely available for download. Wellcome Trust have a very open policy about their information and software, and it is generally made freely available to anyone who wants to use it.

D.1 Merging of rs Numbers

Mike Feolo gave a talk at the Cambridge course on the dbSNP database. He talked about how new Single Nucleotide Polymorphisms (SNPs) were submitted, compared against currently existing SNPs, and merged if they were already in the database. The database RefSeq(rs) numbers are updated about twice a year, and whenever there is a new build of the human genome

sequence. As a result of these builds, some rs numbers are merged to account for increased knowledge of mutations, a process that I was not aware of. This has an impact on my research, because I have received data from an Illumina 317k SNPchip, which I compare to the HapMap dataset. The *RS* numbers from the Illumina chip are for a specific dbSNP build, which may not be fully consistent with the most recent build (from which the HapMap data is retrieved). Mike has recommended that both the build number and the rs number are used when referring to a specific mutation. Sharma Buch pointed out that it was possible to ask Illumina for information about the merged rs numbers for their chips, which would allow datasets from different builds to be used together.

D.1.0.1 Haplotype Blocks and Linkage

Paul de Bakker noted that researchers have moved away from primarily using D' as a descriptive statistic in association studies, because it is not considered to be relevant for association testing. Researchers will typically use D' when investigating haplotypes and haplotype blocks, as this statistic gives an indication of the *recombinational* history between two SNPs. This apparent history can be misinterpreted due to recurrent mutations and back mutations. In a case where new mutations arise on a single haplotype background, D' will not be reduced, because these mutations do not suggest that a recombination event has occurred. On the other hand, such mutations *do* affect r^2 values, so for studies where the *mutational* histories of SNP are more important than *recombinational* histories (as with association studies), r^2 is seen as a better choice. However, it is a good idea to report both values in situations where an estimate of the degree of interaction between markers was required.

D.2 Data Mining Utilities

Mike also introduced a new program to integrate with dbSNP called Genome Workbench. The program allows zooming from chromosome level to sequence level, extraction of data, filtering, and many other tools that reduce the effort in finding genes (or mutations) of interest. Mike also introduced a new database of human variation, called dbGAP, that aimed to store all individual data – including SNPchip, biochemical, and phenotypic information – for various NCBI-funded research projects. Aggregate information will be available for most people accessing the site, but further access will require appropriate approval through a submission process to NCBI.

Albert Vernon Smith demonstrated the HapMap website,[†] showing how most tracks in the browser were customisable, and attempted to show the best possible summary of information at each level (in particular, the SNP display). It is possible to output phased haplotype tracks and tag SNPs, as well as the Haploview triangle plot. In addition, it is possible to add other data to the tracks for display in the browser. The data tracks can then be downloaded as a high-resolution SVG file for submission to journals. Another function available in the browser is the ability to download genotype-level data (including Phased Haplotype Data) for further analysis using other programs. For situations where such a download is not enough, there is a hapmart website, which allows a user to select many different combinations and filtering of data and output options. Albert also demonstrated the new version of Haploview, which supports the import of a few other data types in addition to the Linkage format previously supported. Haploview is now able to download HapMap data from the website from within the program, and is building up support for PLINK data files.

PLINK was demonstrated briefly by Paul de Bakker. It is a command-line driven software package that carries out many useful genome wide

[†] `http://www.HapMap.org`

association analyses, and has modular code that allows additional analyses to be coded in without too much effort. The package produces summary statistics, carries out association analyses and IBD estimates, and is designed to do its calculations fast.

All the software demonstrated during the course is freely available for download. Wellcome Trust have a very open policy about their information and software; all data and software are freely available to anyone who wants to use them0.

D.3 Genome Coverage

Ele Zeggini, presented a few statistics relating to genome coverage for the HapMap populations. Covering 70% of the ≈ 7 million SNPs in HapMap (i.e. capturing 70% of the total genetic variation) would require a SNP density of around 6-10kb (depending on how aggressive the tagging for SNP selection was). This results in SNPchips of around 300-500k for the whole genome – 300k for SNPs selected via a tagging mechanism, 500k for SNPs selected at random. These values happen to be about the level of SNPs available on SNPchips currently. The Affymetrix chips tend to have randomly selected SNPs, while Illumina chips are more geared towards variation identified through HapMap. This suggests that either a 500k Affymetrix chip, or a 317k Illumina chip may be sufficient for a reasonable study of genome-wide variation. While coverage for CEU, and to a lesser degree CHB+JPT, populations is fairly high, the coverage for YRI populations is much less (probably around 40%), so an idea of the study population origins is important before carrying out such investigations. For investigations that need to go into a bit more detail early on, both Affymetrix and Illumina are bringing out 1000k chips that are made up of about 500k SNPs, and 500k Copy Number Variants (CNVs).

Jonathan Marchini compared the various SNPchips with a theoretical "HapMap Chip", and demonstrated that all chips except the Affymetrix 100k had similar effective power for detecting associations (RR=1.7, MAF=[0.1,0.5]). Power reached around 80% for sample sizes of 1000, suggesting that there may be quite a bit of luck involved in finding associations in population sizes of less than 500 (where the power is around 40%).

D.4 Imputation of SNPs

Jonathan Marchini spoke about probability and Bayesian methods that could be used to extract more information from data sets. He demonstrated that it was possible to use the HapMap data to infer the confidence that an untyped SNP may have a given mutation – this process is referred to as imputation. The HapMap data give an excellent insight into the mutational relationships between SNPs, allowing the addition of imputed SNPs into a dataset to give a better idea of potential loci of disease. Jonathan carried out confirmation studies of imputed SNPs, and noted that the error rate was similar to that of the SNPchips. Of course, if an association were implied for imputed SNPs, they should certainly be typed, but this process would likely cost much less than using a more concise SNPchip for the initial typing. An extension of the imputation process is assigning confidence values of mutation to every SNP, allowing a researcher to include data that would be discarded in a standard study.

D.5 ENCODE Regions

Manolis Dermitzakis discussed the availability of data from the ENCODE regions. As part of the HapMap project, there are about 50 regions (500kb to 2Mb) of the human genome that have been sequenced in their entirety in 42

individuals. The variants discovered within these regions were then typed in all individuals in the HapMap population, providing a very detailed insight into genomic structure across human populations. It was noted that, in this set, there was on average one SNP per 300bp. This provides an insight into the average expected variation within a sequence of DNA of a given length, which would be useful in working out how much work is required within a region to categorise the total human variation.

D.6 Recombination Hotspots and Defining Haplotype Block Structure

Manolis Dermitzakis also spoke about the lengths of haplotype blocks that were discovered in the HapMap populations. Comparing blocks between populations, it was noticed that although there were areas in the genome where recombination was more likely, very few regions had a point that resulted in recombination all the time. The implications of this are that any defined haplotype block definitions will not be absolute, and any models of haplotype blocks must take into account these numerous points of "soft" recombination. Current methods for defining haplotype block regions (e.g. Gabriel, 2-gamete) result in absolute regions of recombination. In addition, there is also no provision for overlapping blocks, or blocks within blocks, both phenomenon that I have observed by looking at European data across the Alcohol Dehydrogenase (*ADH*) gene region on chromosome 4. Both of these observations would probably be able to be integrated into a recombination probability model that accounts for such recombination overlaps. The feedback from the attendees at the course was that such models were difficult to comprehend, and the GWAS method using mutational history is easier (and is currently producing many useful results).

In addition, while recombination hotspots were discussed briefly, not much was mentioned about the idea of mutation hotspots – genome loci

that were more likely to have a mutational event. The hyper-variable *HLA* region was mentioned, but reasons (structural or otherwise) behind this variability were not known by the course participants.

Appendix E

HUGO Symposium on Genomics and Ethics, Law and Society

At the beginning of November 2009, a Human Genome Organisation Symposium was held in Geneva on Genomics and Ethics, Law and Society: Sequencing of Individual Genomics – Impact on Society and Ethics (GELS). The symposium was attended by around 50 people from across the world, representing a global group of people interested in the ethical applications of genomic research. The Human Genome Organisation (HUGO) was set up to promote international collaborative effort to study the human genome, and since the completion of the human genome reference sequence, its focus has shifted to the study of issues related to knowledge of genetic data.

E.1 The Thousand Dollar Genome

The completion of the human genome project has propelled genetic researchers towards cheap, fast whole-genome sequencing for the masses. Single molecule sequencing techniques promise to be the next significant

stage in genome analysis, allowing sequencing of up to a billion DNA base-pairs per run (Liu, GELS).[†] Even with $100\times$ genome coverage (to correct for sequencing, orientation, and alignment errors), this brings a 2-week sequencing of a single human genome within reach – this is a time period that would be considered reasonable for high-resolution genetic testing in a clinical setting. Despite this speed, cost is still a fairly large barrier for direct-to-consumer testing, as a $100\times$ run on current sequencing platforms costs about $400,000 (Liu, GELS). However, the cost for an accurate full genome sequence is expected to drop quickly, bringing the target of a $1000 genome into the realm of possibility within the next five years.

E.1.1 Genome Sequencing and Informed Consent

Whole genome sequencing is happening, and at an accelerating pace. Marjolein Kriek was selected for sequencing after the completion of genetic sequences for Craig Venter and James Watson. As Marjolein is a clinical geneticist, informed consent – an understanding of all the issues involved as a pre-requirement to testing – was a given (van Ommen, GELS). Likewise, the first 10 individuals who are participating in the Personal Genome Project (PGP-10) underwent a fairly rigorous process (including a non-trivial test) to demonstrate their understanding of issues involved in sequencing an entire genome. Impressively, all these individuals have agreed to the release of their full genome sequence to the scientific community (and, by extension, the public). This immense database of personal information can describe a reasonable amount about a person, but as was relayed to Marjolein at the end of her talk, you can learn more from a person by talking to them than by reading their DNA.

The fact that well-informed scientists have gone through (or will go through) the process to get their genome sequenced may result in other

[†]Speakers at the HUGO symposium are referenced with their last name and the tag, "GELS", as used here

less-informed people leaping into full-genome sequencing without fully understanding the consequences of their choice. A metaphor of the Judas goat was provided to demonstrate concerns about this (Knoppers, GELS) – abattoirs use a goat to lead sheep into a slaughterhouse[†]; the sheep willingly follow the leader, not fully understanding the consequences of their actions. The use of this metaphor emphasised that it is immensely important for people to understand the impact genetic sequence exposure can have, not just for them, but also for their children, parents, and other relatives.

E.1.2 Public Collaboration – Research 2.0

The speed of genomic technology advances suggests that there will be a window of only a few years before the low cost of genetic testing will put these tests within easy reach of most of the population, regardless of the amount of training and knowledge that people have. Furthermore, if the goal of an information-based economy is to get public data and results to the public, there cannot be a requirement that people have degrees in biology and/or genetics before having access to those data and results. The current outlook for the next generation of research is an Internet-driven approach, with massive collaboration and data sharing between research *participants* (Avey, GELS). If investigators are not prepared for this explosion of participant-driven collaboration, they will be in danger of getting lost in the noise of public discussion.

[†]Evocative phrases were often used in the safety of the symposium environment to remind participants about extreme views that should be considered in discussions.

E.2 Consequences of Public Release of Genomic Data

The initial impact of full-genome sequences for the people who have had their genome sequenced seems to be largely media-derived. Marjolein Kriek changed from a nobody scientist into someone who had captured the interest of lots of journalists, and no longer needed to pay anyone to get a good photo of herself (Kriek, GELS). Many expected questions were asked of her, but also a couple that she hadn't really thought about:

- Will your family still be able to eat the same stuff?
- Will you be able to choose your partner?

A genetic analysis may suggest that certain foods are not compatible with a person's genetic makeup, and possibly even that certain partners may not be the best choice for producing healthy offspring. Knowing that there is an enhanced genetic risk for a particular disease, it may not be okay to carry out actions that increase the environmental risk for that disease (e.g. a climbing career for someone with weak bones), particularly if a person wishes to keep their health insurance premiums low. This is a particularly sensitive area in the USA, where insurance may be refused if someone is considered high-risk. These personal concerns also apply to relatives to some extent. Some component of the genomic sequence will be shared, possibly also exposing some family secrets that were not intended to be made public knowledge.

E.2.1 The True Impact of Public Release

Despite concerns, the public release of full genome sequence data has, so far, not resulted in substantial setbacks for those people. Simply, everything

that could go wrong and should go wrong. . . hasn't (Knoppers, GELS). This does not mean concerns should be ignored, but as it stands, society in general seems to have a fairly reasonable and measured approach to how it deals with public exposure of genetic data.

E.3 Defining Rules for Genetic Patents

Sequence data for novel genetic variants has been used as part of patent applications in an attempt to attach a commercial value to DNA. These attempts have mostly been rejected, because genetic sequences, in themselves, cannot be used to make anything. There was a recent legal case on Expressed Sequence Tags (ESTs) – short DNA sequences that tag particular genetic regions, attempting to clarify whether or not these sequences could be patented. The patent offices cannot set the rules under which new patents are granted, but can receive new types of patents (such as ESTs) that there are no current rules for. When new types of patents are considered, the office needs to predict what might be held up in courts, even though it may be many years before applications are tested in the court system. When the ruling was finally made, It was pointed out that ESTs were only an intermediate process, and unless they were associated with a specific utility and real-world use, the patent application should be rejected. This case was brought to the courts about fifteen years after the first patents for ESTs were applied for, demonstrating how difficult it can often be to decide on what rules should be applied to new applications (Toupin, GELS).

E.3.1 The Futility of Genetic Patents

There are two main practical arguments against patents: patents hurt the innovation process, and patents hurt downstream research. Although lack of innovation seems to be a common argument by opponents of patents,

there does not seem to be any evidence of the impact of research volume on patented genes – they seem to be doing just as well as genes that have no patents (Caulfield, GELS). With regards to affecting downstream research, patents (by design) provide inventors with a monopoly to prohibit external use of patented inventions. However, in many cases researchers are knowingly breaching patent licenses in order to discover new things about patented genes. There is little evidence to suggest that patents are needed (i.e. they don't seem to promote innovation), and patents tend to be fairly expensive to obtain, maintain, and very expensive to litigate (Caulfield, GELS). In fact, fewer than 50% of all patents are maintained through their entire life (Ducor, GELS), suggesting that the licensing costs alone are enough to cause inventors to reconsider the benefits of a patent license. It seems reasonable to expect that genetic patent applications will have increasingly less worth in the future information-driven society, and research will still carry on, as before, at an exponential rate.

E.4 Privacy and Information Flow

People often consider privacy to be an absolute quality, equating it with secrecy. This is not a true view of privacy, because expecting absolute secrecy in every case means that nothing is disclosed to anyone. In this sense, privacy is an issue of respecting the social rules of information flow in particular contexts. When privacy is compromised, it means that information has been inappropriately shared, i.e. the social rules have been broken (Nissenbaum, GELS). In this light, the concept of specific features being always personally identifiable information (PII, a term often used when discussion anonymisation of data) doesn't make sense, because context matters in defining whether or not it is appropriate for information flow to happen. Social rules can be defined together with context in a specific and explicit fashion, but this does not often happen in a research

or diagnostic context. It is more common for researchers and clinicians to make an absolute statement of secrecy (e.g. a privacy clause in a consent form, or doctor-patient confidentiality) when it is obvious that information flow is a necessary part of research and clinical practice.

E.5 Genetic Determinism and Public Education

It is clear that the public do not have an intuitive grasp of what genetic tests mean. American parents have been encouraged to get genetic testing to establish what sport their child will excel in, but the research behind those tests suggest a very small ($< 5\%$) contribution to phenotypic variation (Cox, GELS). In other words, parents are led to believe that their child's future is set in stone by their genes, when other environmental factors (e.g. physical training) play a much larger role. However, this lack of understanding is not restricted to the general public; doctors and physicians (whom people are likely to trust more than anyone else in medical matters) are also caught out by misrepresentation of results (Cox, GELS). Part of this misunderstanding may be due to different terms used by researchers and clinicians (Lindpaintner, GELS). Academia generally discuss odds ratios and relative risk, and tend to be happy with odds ratios of 2 (sometimes as low as 1.15 for some genome-wide studies). Clinicians prefer sensitivity and specificity and are usually only interested when these statistics are greater than 75% (an odds ratio of about 9). A greater concern is that misunderstanding exists in the scientific research community as well. If the researchers can't agree on the importance of discovered genetic associations (including causal links), there is little hope that the general public will be able to be properly educated on the issues involved.

This raises a very important question, how should the public be educated?

E.6 Case Study – Genomic Medicine in Mexico

One of the best examples of dissemination of knowledge to the public can be found in the recent research of genomic medicine carried out in Mexico (Jiménez-Sánchez, HUGO). A research project was introduced by the Mexican congress to analyse the genetic structure of the Mexican population. In conjunction with this project, an Ethics, Law and Society research centre was opened to enable good communication and discussion with the Mexican community for genomic research. A key goal of the research project was to produce public results through an interactive database. To emphasise this, following the conclusion of the initial phase of the research, both the research paper and the database were presented to the Mexican president at the same time.

E.6.1 Returning Research to the Mexican Community

Before research began, Researchers educated community leaders about the intended goals of the research project. These leaders then presented a series of informative sessions to the public and the media. Brochures and comic books were also produced and distributed to the public. A month before participants provided blood samples, an exact copy of the consent form was posted in universities and other public spaces to make sure everyone would be aware of what they would be asked to do. Blood was taken at the universities where the genomic research was carried out, and a series of public lectures were given on that collection day to educate people about the research. Aside from general geographical location data, blood samples were completely anonymised, and participants were made aware that there was no way to get or find their sample later on in the study. Results were initially released as a research paper and interactive database. This paper was translated into Spanish, simplified into a comic book, and packages

were made to be delivered to governors and attendees of public lectures in which the results were reported.

E.7 Genetics in Africa

All evidence points towards a human origin of Africa, with a series of migrations from there to other areas in the world. It follows from this that African populations are the most diverse and genetically evolved group of people in the world (Daar, GELS). However, studies of African populations tend to be quite limited in scope, and leave behind few benefits for the study participants. Apart from a study of the malarial genome (and possibly the HapMap project), there do not seem to be any ongoing projects in Africa that are likely to provide advantages to the community. To add further strain to the research capabilities of Africa, researchers will typically only come to Africa if they are invited, and only 30-50% of professionals trained in Africa are retained (Ramesar, GELS). The lack of studies on African populations is a shame, especially because of the benefits from carrying out genetic research in Africa in tandem with other countries. Due to the increased genetic diversity of African populations, disease-associated variants may be found in higher frequency in Africa (Olopade, GELS). This makes associations easier to validate and investigate further, because the number of people with a particular variant (or combinations of variants) will be higher.

E.8 Poster Presentations

Most of the attendees who were not giving a talk at GELS were presenting a poster. A total of sixteen posters were presented at the symposium, most covering topics relating to the impact of genomic technologies on society.

For example, Christen Rachul presented a poster on how racial terminology was portrayed differently in peer-reviewed articles, press releases, and newspaper articles. Press releases and newspaper articles often exclude references and context from terms used in articles, and simplify language to terms that often have more emotional impact to readers. Billie-Jo Hardy presented a poster on genomic sovereignty, the idea that populations should have a right to management of their own genetic samples and associated information. The decisions of community leaders define where that information lies on a continuum from a global public good (free access for everyone) to a commodity (limited access at a price).

My poster presentation was about a genomic test that we have designed to estimate Maori-European admixture (genetic mixing of multiple populations) in a Maori tribe, Rakaipaaka.[†] The population profile was part of a larger "Rakaipaaka Health and Ancestry Study" to investigate ways in which the current and future health of the community can be improved. Many attendees at the conference were interested in the migration history of Maori, a recent (1000-500 years ago) migration to a land that had no previous established human population.

E.9 Speakers at GELS 2009

- Stylianos E. Antonarakis, *The Medical Genome*
- Charles Auffray, *Redefining Intellectual Property in the Transition from Genomics to Systems Medicine*
- Linda Avey, *Personal Genetics*
- Alastair V. Campbell, *What is Special about 'Genetic Privacy'*
- Timothy Caulfield, *Are Gene Patents the Problem?*

[†] `http://gringer.org/gels_poster.pdf`

- Ruth Chadwick, *Redefining Privacy, Choice, and the Internet*
- David Cox, *Genetic Determinism and Real Life*
- Abdallah Daar, *Genomics Initiatives in Developing Countries*
- Philippe Ducor, *'Open Access' Aspects of DNA Patenting*
- Gerardo Jiménez-Sánchez, *Genomic Medicine in Mexico*
- Klaus Lindpaintner, *Future of Health Care Industry*
- Edison T. Liu, *Genomic Sequencing*
- Jeantine Lunshof, *Redefining Privacy*
- Bartha-Maria Knoppers, *Personal Genomics and Privacy*
- Mark McCarthy, *What Will New Sequencing Technologies Deliver for Science and Society*
- Partha Majumdar, *Ethical Dilemmas in the Conduct of Genetic Research*
- Helen Nissenbaum, *Privacy, Technology, Policy, and the Integrity of Social Life*
- Olufunmilayo Olopade, *Advances in Breast Cancer*
- G. J. B. van Ommen and Marjolein Kriek, *Aims, Outcomes and Experiences of a Dutch Female Sequence*
- Raj S Ramesar, *Will Africa be Relegated to the Role of a Neglected Parent?*
- James Toupin, *The Development of the Law of Gene Patenting*

Appendix F

Recombination Simulation

In order to demonstrate recombination in presentations and reports, it was useful to carry out a basic simulation to generate a visual representation of recombination through successive generations. The algorithm written for this purpose is an R script that generates a list of "Haplotype Blocks", together with a number between 0 and 1 indicating the location at which the block begins (or, alternatively, the location of the recombination points). An example of this is shown in Table F.1, which indicates the block structure of the final recombinant chromosome of Figure 1.5 in the Introduction (reproduced here as Figure F.1). Not including the beginning and end of the chromosome, there are 10 recombination points.

The recombination simulation function takes four variables as inputs: the first two variables being the parental chromosomes from which to generate the recombinant chromosome, and the second two being the minimum and maximum number of recombination events. The output of the function is the recombinant chromosome, as a data frame containing a list of blocks together with the recombination point at which the blocks *start* (as shown in Table F.1). The actual number of recombination events is determined by a single unweighted random sample (i.e. the `sample` function of *R*) from a list of numbers between the minimum and maximum

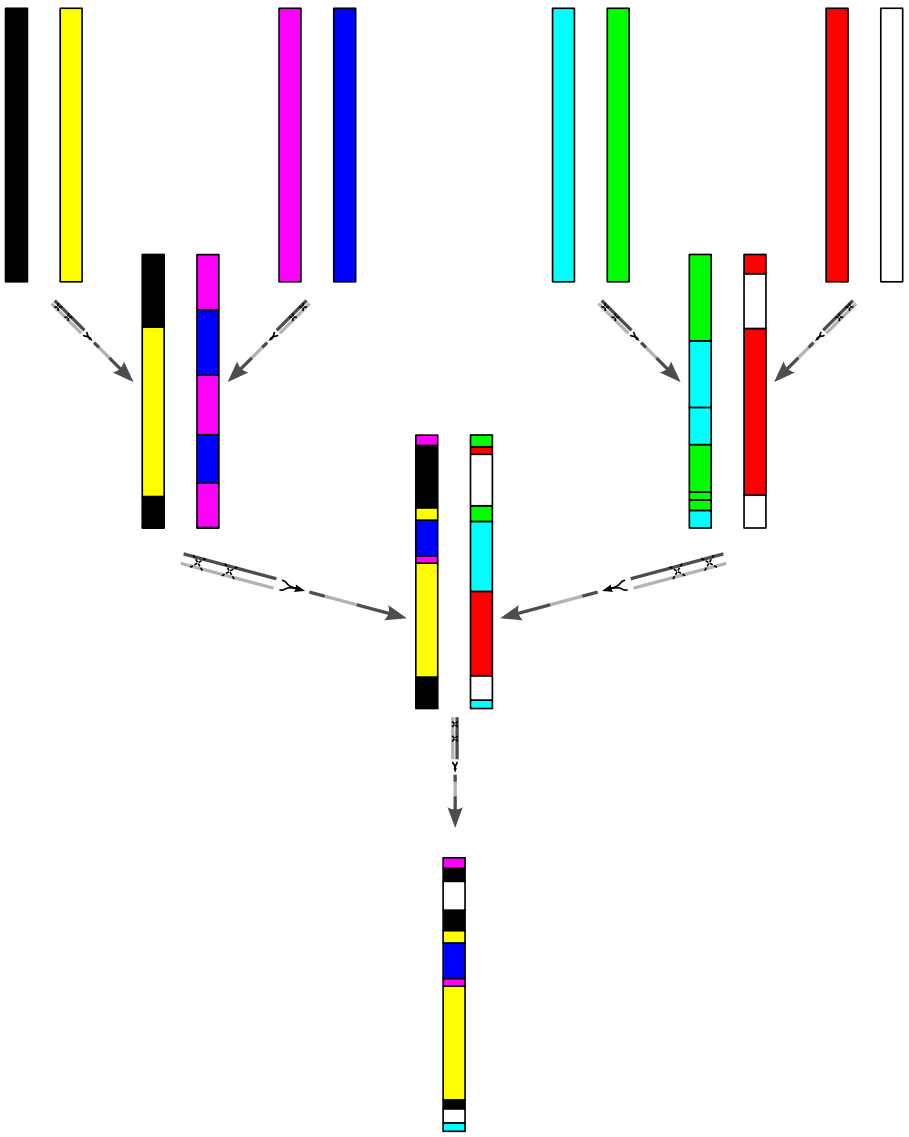

Figure F.1: The genetic ancestry of a single chromosome is complex, the result of multiple recombination events that happen at each generation. In this figure of simulated recombination, black lines indicate recombination points (see Figure 1.4). The final chromosome shown in this figure contains a genetic history that is derived from six of the original eight ancestral chromosomes.

Block	Points
cyan	0.00000000
white	0.03082244
black	0.08267944
red	0.11491636
magenta	0.53165148
blue	0.55864589
red	0.68870415
black	0.73414129
white	0.80855758
black	0.91409931
magenta	0.96201204

Table F.1: Haplotype blocks simulated after 4 generations of recombination.

number of specified recombination events. The location of recombination points is determined by the generation of a random number having uniform distribution over the interval $(0, 1)$ (i.e. the `runif` function of *R*).

The function selects at random a chromosome to start extracting sequence from, and stores the part of that chromosome up to the first recombination point. At each subsequent recombination point, the function changes chromosome, creates a recombination event, then continues extracting sequence from the other chromosome:

```
recombine <- function(ChrX, ChrY, minPoints = 3, maxPoints = 4){
  result <- NULL;
  recombPoints <- sort(runif(sample(minPoints:maxPoints,1)));
  currentChr <- sample(c(0,1),1);
  startPos <- 0;
  newPoints <- NULL;
  for(endPos in c(recombPoints,1)){
    if(currentChr == 0){
      stopX <- c(ChrX$Points[-1],1);
      curPoints = which((ChrX$Points<=endPos) &
        (stopX >= startPos));
      newPoints <- ChrX[curPoints,];
```

```
    } else {
      stopY <- c(ChrY$Points[-1],1);
      curPoints = which((ChrY$Points<=endPos) &
        (stopY >= startPos));
      newPoints <- ChrY[curPoints,];
    }
    newPoints$Points[1] <- startPos;
    result <- rbind(result,newPoints);
    startPos <- endPos;
    currentChr <- (currentChr + 1) %% 2;
  }
  return(result);
}
```

It is important to note that this algorithm has been designed for demonstration purposes only and should not be used to infer anything about the nature of recombination that has already happened in a real-world situation. In particular, this simple recombination algorithm makes an assumption about recombination that is incorrect – it assumes that the probability of a recombination event is uniform across the entire chromosome. This assumption can be corrected by providing a distribution indicating the probability of recombination across the entire chromosome, and modifying the random number generation to take account of this non-uniform probability distribution.

Appendix G

Databases

I designed and built a database based on the Rakaipaaka Beneficiary Electoral Roll forms, which has been used by Te Iwi o Rakaipaaka (TIORI) as an electronic record of membership. I also cleaned up the questionnaire database, making it easier to extract answers to specific questions, and update the layout of the database in the future. My work on these databases has been useful for the implementation of a new sample storage database at ESR (which has been integrated with the Rakaipaaka database and the questionnaire database). This database includes genetic and biochemical data from analysed samples, and has made future data processing and analysis much easier. Part of my work has also involved helping to design and develop this database, increasing its utility for other people.

In September 2006, I travelled to Nuhaka and spent two days TIORI with their membership database. The genealogical information from this database may be used in the future to determine how well genomic information correlates with recorded ancestry. Work carried out included setting up their network to enable easier communication between computers, and organising domain name registration and forwarding for rakaipaaka.iwi.nz and rakaipaaka.co.nz.

Additional work for the Health and Ancestry Study (RHAS) has included the conversion of survey forms from Microsoft Word files to Microsoft Infopath documents. The Infopath format allows for the entry of data using a computer, producing a well-structured text file (XML) as output that can then be integrated into the ESR database. I have converted both the general Health and Medical Research Questionnaire (which includes questions about Employment, cigarette and alcohol consumption, drug use, family medical history, exercise, and eating habits) and the consent form for RHAS (relating to genetic testing, and feedback about study-derived results) into the Infopath format.

G.1 Genealogical Construction

Using the genealogical information from the RHAS database, a consensus family tree was constructed with 926 individuals linked together via marriage or ancestry, and an additional 206 individuals who could not be placed on the larger linked tree.

G.1.1 Transliteration

The most interesting initial outcome of this was the observation of names that were transliterated between different languages. The largest disagreement among the individuals who reported ancestry was in the name of Yoachim (or Johann) Schmidt (one of the many different names he was given). This ancestor had a Prussian origin, but his name has been modified to numerous English and Maori equivalents by his descendants. The English-sounding first names that have been given to him have been Jack Hachem (with possible alternative spellings of that second name being Hacham, Hakken and Hakon), with a surname of Smith. The Maori interpretation for the first name has been Haki (or just Ki), with a surname of

Mete. There seems to be no further support in the English or Maori names for the first name being Johann, which would probably be transliterated to English as John (rather than Jack) or in Maori as Hoani (or Hone). However, there is still a possibility of Johann being an alternate (but more rarely used) name for the same person. Looking at the extremes of these name variations, the same person is referred to both as Yoachim Schmidt and as Ki Mete – it would be difficult, in the absence of intermediary evidence, to work out that these two names referred to the same person.

Some of the English/Maori variations that were observed during an attempt at creating a consensus family tree for Rakaipaaka can be found in Table G.2 and Table G.3. Note that these are the author's own inferences, and may not necessarily reflect the typical translations used for English/Maori name conversions.

Maori	English
Ahenata	May
Ani	Annie
Arihi	Alice
Ema	Emma
Haki	Jack
Hoani	John
Hone	John
Te One	John
Keita	Kate
Matenga	Martin
Mere	Mary
Oriwa	Olive
Paora	Paul
Pira	Bill
Piripi	Phillip
Pita	Peter
Rangi	Henry
Raniera	Daniel
Rawiri	David
Rewi	Dave
Ripeka	Rebecca
Ruihi	Lucy
Taare	Charles
Tame	Tom
Te Rina	Lena
Timoti	Timothy
Wiremu	William

Table G.2: English/Maori transliteration of first names

Maori	English
Harete	Hallet
Huka	Hook
Mete	Smith
Pakai	Park

Table G.3: English/Maori transliteration of surnames