

Genetic Study of Kuru

Liam James Quinn

**MRC Prion Unit at UCL
Institute of Prion Diseases
University College London**

A thesis submitted in fulfilment of the degree of Doctor of Philosophy

October 2019

Declaration

I, Liam James Quinn confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

-Liam James Quinn

October 2019

In loving memory of my father Padraic

Ar dheis Dé go raibh a anam

Abstract

Kuru was the first documented epidemic of prion disease in humans. It took place in a restricted region of the Papua New Guinea Highlands during the twentieth century, with the final cases occurring in the first decade of the twenty-first century. Over 2,700 deaths were recorded in the period 1957-2004 when surveillance was carried out in the region. The epidemic impacted greatly upon affected communities as a result of its high incidence, and its association with practices of sorcery in the region.

This project sought to further understand the genetic impact of kuru using a genotyped data from 943 individuals from 21 ethno-linguistic groups in the Eastern Highlands region of Papua New Guinea (EHPNG). Analysis was conducted to classify the population structure in the region using a suite of population genetic tools. Linguistic group membership was the strongest descriptor of population genetic structure in the region, informing experimental design in subsequent analyses. A drop in genetic diversity was observed in the most affected South Fore linguistic group during the course of the epidemic, confirming that the drastic impact on the ground left a genetic signature. A new tool (Chromomatcher) was developed and incorporated in attempts to find genetic variants under positive selection during the kuru epidemic. Several genetic variants show evidence of being under recent positive-selection. Finally, polygenic architecture of prion diseases was investigated revealing a significant polygenic architecture for sporadic Creutzfeldt–Jakob disease (sCJD) with a substantial portion shared with variant Creutzfeldt–Jakob disease. No shared architecture was observed between sCJD and kuru, reflecting the challenges of applying this technology to under-studied populations.

Impact Statement

Work performed in this thesis has enhanced understanding of the kuru epidemic, the first human prion disease epidemic. The epidemic had a drastic impact on affected communities in the region and has informed understanding of prion diseases in other parts of the world. Kuru and the associated scholarship and publicity that came with it have come to form a large part of the reputation and identity of the affected region. Work conducted in this thesis has been carried out with the intention of improving understanding of kuru via a methodical analysis free from any prior judgements that have previously circulated regarding the epidemic and the affected peoples. Work presented will hopefully contribute to removing some of the damaging ignorance and stigma which has previously been attributed to the disease and the affected communities. Importantly, future publication of work in this thesis and communication directly to the affected communities will help relieve some of the tensions and uncertainties that have afflicted the region that has endured kuru.

Understanding of kuru in the past has informed debates at parliamentary public policy meetings in the United Kingdom regarding public health management in relation to prion disease. Work in this thesis will help further understanding of the risks and potential impact of human prion disease and should inform policy approaches to be implemented with the aim of avoiding future human prion epidemics and the terrible consequences for individuals, families and communities highlighted in this thesis.

A significant portion of time and effort was expended optimising a new 'Chromomatcher' tool specifically designed to find signatures of recent natural selection in populations with complex population demography. This tool will provide future benefits to other investigators tackling problems of capturing genetic impacts of recent selection pressure in populations as complex as those in the kuru-affected region.

It is hoped that work in this thesis to apply the latest computational methodologies to understand kuru will provide a blueprint for future scholars when attempting to glean understanding of recent

epidemics from a genetic perspective. The opportunities and challenges presented by the kuru epidemic informed specific experimental designs that hoped to best leverage the data available in order to capture the genetic legacy of kuru. Contributions from studies in other fields helped inform experimental design also and motivated lines of enquiry that will also hopefully encourage future scholars to take holistic, multi-disciplinary approaches when approaching genetic analyses routed in complex cultural, epidemiological settings.

Contents

Title	1
Declaration.....	2
Abstract.....	4
Impact Statement	5
Acknowledgements.....	11
List of abbreviations.....	13
List of tables and figures	15
Chapter 1 – Thesis introduction and project aims.....	18
1.1 PD (PD)	18
1.1.1 Sporadic Creutzfeldt-Jakob disease (sCJD)	19
1.1.2 Inherited PD	20
1.1.3 Acquired PD.....	22
1.1.4 Iatrogenic CJD (iCJD)	22
1.1.5 vCJD.....	24
1.1.6 Structure of cellular PrP (PrP ^C) and PrP ^{Sc}	28
1.1.7 Strains and phenotypic variability.....	28
1.1.8 Diagnosis of PD and PD surveillance	31
1.2 Kuru.....	33
1.2.1 Kuru research	34
1.2.2 Impact of kuru on affected communities	43
1.3 Population genetic research	44
1.3.1 Population genetic research in Papua New Guinea and the wider region.....	44
1.3.2 Archaic introgression	47
1.3.3 Eastern Highlands of Papua New Guinea (EHPNG).....	50
1.3.4 The Fore people	53
1.4 Change in EHPNG	54
1.4.1 Impact of European contact and administration.....	54
1.4.2 Ipomean Revolution.....	56
1.5 Understanding population responses to natural selection pressure	57
1.5.1 Use of genetic data to understand natural selection in human populations	57
1.5.2 Use of selection techniques in Oceania and other under-studied populations	59
1.6 Understanding of genetic architecture of traits	61
1.7 Aims of the project.....	68

1.8 References	69
Chapter 2 - Population genetic analysis of EHPNG	76
2.1 Introduction	76
2.2 Materials and methods	80
2.2.1 Genetic dataset	80
2.2.2 Phasing of data	82
2.2.3 Identification of related individuals and estimates of genetic diversity within groups.....	82
2.2.4 Analysis of EHPNG population structure	83
2.2.4.1 Principal Component Analysis (PCA).....	83
2.2.4.2 Admixture analysis.....	84
2.2.4.3 ChromoPainter and FS analysis of EHPNG linguistic groups.....	84
2.2.5 Influence of geography on population structure and observed genetic diversity	86
2.2.6 Analysis of migration processes using TVD and F_{ST} between individuals	86
2.2.7 fastGLOBETROTTER.....	88
2.3 Results	89
2.3.1 F_{ST} Measurements.....	89
2.3.2 Admixture Analysis.....	90
2.3.3 Principal Components Analysis	91
2.3.4 Runs of Homozygosity (ROH)	92
2.3.5 CP Analysis and FS Clustering.....	92
2.3.6 Admixture Analysis Using fastGLOBETROTTER	101
2.4 Discussion.....	102
2.5 Summary of findings	111
2.6 References	112
Appendices.....	114
Chapter 3 - Impact of kuru on affected communities in EHPNG	129
3.1 Introduction	129
3.2 Materials and methods	132
3.2.1 Assessment of Impact of kuru on EHPNG	132
3.2.2 Analysis of <i>PRNP</i> codon 129.....	135
3.2.3 Understanding of <i>PRNP</i> 127V haplotype in kuru affected communities.....	137
3.3 Results.....	140
3.3.1 Analysis of genome-wide signatures of population response to kuru	140
3.3.2 Analysis of <i>PRNP</i> codon 129 frequencies.....	143

3.3.3 <i>PRNP</i> 127V variant does not provide striking signature of selection in genome-wide scan of kuru-affected population.....	148
3.4 Discussion.....	150
3.4.1 Change in diversity.....	150
3.4.2 <i>PRNP</i> codon 129 minor allele frequency study.....	153
3.4.3 Analysis of <i>PRNP</i> 127V	155
3.5 Summary of findings	156
3.6 References	158
Appendices.....	161
Chapter 4 – Search for variants under selection pressure during the kuru epidemic.....	169
4.1 Introduction	169
4.1.1 Complex patterns of incidence and exposure	169
4.1.2 Aetiology and architecture of kuru.....	170
4.1.3 Complex population structure	171
4.1.4 Recent and short epidemic.....	172
4.2 Materials and methods.....	173
4.2.1 Phenotypic information in the dataset.....	173
4.2.2 GWAS	174
4.2.3 XP-EHH and Relate Analysis.....	177
4.2.4 Chromomatcher analysis	178
4.2.5 Analysis of WGS data of kuru-resistant individuals	180
4.3 Results.....	183
4.3.1 GWAS	183
4.3.2 XP-EHH and Relate analysis	185
4.3.3 Chromomatcher analysis	186
4.3.4 WGS analysis.....	190
4.4 Discussion.....	193
4.5 Summary of findings	202
4.6 References	203
Chapter 5 – Exploring polygenic architecture of PD’s and kuru	206
5.1 Introduction	206
5.2 Materials and methods.....	209
5.2.1 Approach taken to investigate PRS of PD	209
5.2.2 sCJD GWAS and PRS.....	211

5.2.3 Shared aetiology between sCJD and vCJD	212
5.2.4 Shared aetiology between sCJD and an 'unrelated trait'	212
5.2.5 Shared aetiology between sCJD and kuru.....	212
5.3 Results	213
5.4 Discussion.....	219
5.5 Summary of findings	224
5.6 References	225
5.7 Appendices.....	227
6 Summary and Future Work.....	229
6.1 Summary	229
6.2 Future work.....	231
6.2.1 Suggestions for development of investigations performed in this thesis	231
6.2.2 Future avenues of kuru research	232
References	235

Acknowledgements

I would firstly like to thank and acknowledge the contribution of my supervisors Simon Mead and Garrett Hellenthal. Their guidance and support has been of immense value during the PhD. Simon, your continuous and infectious enthusiasm for all things related to this project has helped spur me on throughout. Your clear, concise advice and our long conversations about experimental design have been of unquestionable value. Your role as a principal supervisor providing support through a very tumultuous time for me and my family will not be forgotten. Garrett your generosity, clarity and patience when explaining the complex statistical underpinnings of your programmes has taught me a lot that I hope to carry forward in my career. Your example of being a methodical scientist, seeking objective explanations from experiments with healthy levels of scepticism has been of value as well.

I would like to thank the long term custodians of kuru research who have contributed to this project, principally Jerome Whitfield, John Collinge, Tracy Campbell and Michael Alpers. Jerome, in addition to the immense task of collecting biological samples from thousands of individuals in the New Guinea Highlands you have been an available source of wisdom and guidance for all things related to the history and anthropology of the region. John, your commitment to continuing research into kuru and fresh insights at lab meetings have been of great use. Tracy thank you for being the person most responsible for DNA extractions and curation of samples over several decades. I would still be in the basement at Queen Square trawling through freezers without your help. Michael, your life's work in the field of kuru research has been a great inspiration to me, and meeting you in person and talking this year was a great privilege. Thank you for taking the time to read manuscripts and offer comments in such detail. I would also like to thank the many members of the Papua New Guinea Institute of Medical Research who have provided field resources, communicated with affected communities and helped stimulate a fruitful collaboration between UK and Papuan prion disease researchers. Importantly I would like to thank members of the kuru-affected communities who have provided support and participated in this project through the numerous interviews conducted and

biological samples they have provided. The ultimate motivation of all kuru research is to aide these communities for finding answers and explanation for what was a cataclysmic period in their history.

I would like to thank numerous people in both the Human molecular genetics team at the MRC Prion Unit and the Hellenthal group at UCL genetics institute. From the MRC Prion Unit Tracy, Penny, James, and Ron for all the laboratory work that underpins any computational analysis. Holger, thank you for assisting so greatly with the Whole Genome Sequence analysis, and teaching me the quirks of database querying. Other members of the group have offered wonderful input during lab meetings, so thanks to Emmanuelle, Carolin, Luke, Ines, Chiara, Thanos and Will. From the Hellenthal group in particular I am grateful to Kaustubh, Camillo, Javier, Lucy and Saioa for PhD advice, demonstration of software applications and patiently explaining the underpinnings of their own research. I would like to thank Peter Price who made invaluable contributions in helping me test trial versions of the Chromomatcher software. Thank you to Doug Speed who provided guidance on imputation of samples during the project. Thank you to Ida Moltke at the University of Copenhagen who has offered advice on the project and kindly invited me to Copenhagen in 2017 to demonstrate use of the Relate tool.

I would like to thank Emma Jones, Lee Darwent, my brother Paul, Fernando Guntero, and Helen Speedy for proof-reading draft chapters of this thesis and providing great feedback.

On a personal level I would like to thank all my friends and family over the past few years. In particular my parents, who have instilled principles of hard work (Dad) and generosity (Mum). I would not be in this position without those virtues which you have both personified throughout my life.

List of abbreviations

BSE	Bovine Spongiform Encephelopathy
CJD	Creutzfeldt–Jakob disease
CP	ChromoPainter
CSF	Cerebrospinal fluid
EHPNG	Eastern Highlands of Papua New Guinea
FS	FineStructure
FST	Fixation index
GWAS	Genome wide association study
IBD	Identity by descent
iCJD	iatrogenic Creutzfeldt Jakob disease
ISEA	Island South East Asia
kya	Thousand years ago
MDD	Major depressive disorder
MDS	Multiple Dimensional Scaling
MRI	Magnetic Resonance Imaging
MSMC	Multiple Sequentially Markovian Coalescent
nSL	New haplotype-based statistic
PCA	Principle Components Analysis

PD	Prion disease
PNG	Papua New Guinea
PNGIMR	Papua New Guinea Institute of Medical Research
<i>PRNP</i>	Prion protein gene
Prp	Prion protein
PrPC	Cellular prion protein
PrPsc	Scrapie isoform of prion protein
ROH	Runs of homozygosity
RTQuIC	Real-Time Quaking Induced conversion
sCJD	Sporadic Creutzfeldt–Jakob disease
SNP	Single Nucleotide Polymorphism
TNG	Trans New Guinea Phylum
TSE	Transmissible Spongiform encephalopathy
TVD	Total Variation Distance
vCJD	Variant Creutzfeldt Jakob Disease
WES	Whole exome sequence
XP-EHH	Cross population extended haplotype homozygosity

List of tables and figures

Figure 1.1	Schematic of <i>PRNP</i> mutations (Page 20)
Figure 1.2	Representation of brain imaging of vCJD and sCJD (Page 25)
Figure 1.3	Route of kuru through EHPNG (Page 27)
Figure 1.4	Kuru surveillance performed by PNGIMR (Page 39)
Figure 1.5	Map of EHPNG (Page 50)
Figure 2.1	CP analysis of PNG merge data (Page 93)
Figure 2.2	MDS plot of Oceania merge data (Page 96)
Figure 2.3	Geogenetic analysis of EHPNG linguistic groups (Page 98)
Figure 2.4	Migrant analysis results of South Fore linguistic group (Page 100)
Figure 2.5	fastGLOBETROTTER output for target South Fore linguistic group (Page 102)
Figure 2.6	CP outgroup analysis heatmaps (Page 106)
Figure 3.1	Info score histogram for Imputed merge data (Page 136)
Figure 3.2	Analysis of changes in genetic diversity during kuru epidemic (Page 141)
Figure 3.3	Histogram of imputation analysis of genetic diversity analysis (Page 142)
Figure 3.4	Analyses of changes in genetic diversity using Plink 1.9 summary statistics (Page 143)
Figure 3.5	Ancient DNA analysis of <i>PRNP</i> codon 129 minor allele frequencies (Page 145)
Figure 3.6	Analysis of <i>PRNP</i> codon 129 minor allele frequencies in EHPNG (Page 146)

Figure 3.7	Time-dependent analysis of <i>PRNP</i> codon 129 genotypes (Page 147)
Figure 3.8	Relate analysis of <i>PRNP</i> 127V carriers (Page 148)
Figure 3.9	Chromomatcher clustering at <i>PRNP</i> in South Fore population (Page 149)
Figure 3.10	Histogram of Chromomatcher scores for South Fore population (Page 150)
Table 4.1	Breakdown of Epidemiological labels in PNG merge data (Page 172)
Figure 4.2	Map of PNG WGS data used in analyses (Page 180)
Table 4.3	Details of Geographical location of PNG WGS samples (Page 181)
Figure 4.4	GWAS output of kuru meta-analysis (Page 182)
Figure 4.5	Kuru meta-analysis results (Page 183)
Figure 4.6	XP-EHH and Relate analysis of kuru (Page 184)
Table 4.7	Results of XP-EHH analysis of kuru (Page 185)
Figure 4.8	Chromomatcher optimisation analysis (Page 186)
Figure 4.9	Chromomatcher analysis of kuru-resistant 'elderly women' (Page 187)
Figure 4.1	Chromomatcher analysis of high-exposure villages (Page 188)
Table 4.11	Summary of Chromomatcher high scoring regions (Page 189)
Table 4.12	Results of WGS data analysis (Page 189)
Table 4.13	WGS analysis of Chromomatcher high scoring regions (Page 190)
Table 4.14	WGS analysis of sCJD associated variants (Page 191)
Figure 5.1	sCJD GWAS (Page 211)

- Figure 5.2 sCJD PRSice analysis (Page 214)
- Figure 5.3 vCJD PRSice analysis (Page 215)
- Figure 5.4 Facial attractiveness PRSice analysis (Page 216)
- Figure 5.5 kuru PRSice analysis (Page 217)

Chapter 1 – Thesis introduction and project aims

In this chapter foundational ideas in relation to kuru and kuru research are discussed. Previous scholarship has motivated the paths taken in this thesis and areas of investigation that were explored are outlined in section 1.7. A review of current understanding in research of all forms of human prion diseases (PD) in section 1.1 provides context to the broader field that kuru research resides (section 1.2). From the outset of this project it was understood that understanding of the specific population genetics of the region of disease would be pivotal, with previous findings discussed in section 1.3. Recent events that have dramatically affected the way of life in the region other than kuru are discussed in section 1.4. The use and application of methods to test for natural selection in populations is discussed, and their particular application to novel under studied-populations in section 1.5. Finally a discussion of the importance of understanding of the genetic architecture of traits and diseases is undertaken in section 1.6, and the benefits and challenges of characterising the full spectrum of genetic contribution to disease.

1.1 PD (PD)

PD is characterised by the action of the misfolded prion protein (PrP) leading to neurodegenerative diseases that occur across multiple organisms(1). PrP is encoded by the prion protein (*PRNP*) gene. *PRNP* was located on Chromosome 20 between base pairs 4,615,068 and 4,630,233 through linkage mapping of families with prion disease. The sequence of the gene is 253 amino acids in length with reduction to 208 amino acids after post-translational cleavage of the of the amino and carboxyl terminal ends. The gene has two exons with the entire open reading frame being found in exon 2. The gene is highly conserved across organisms particularly mammals with PrP sequence identity > 50% and is present in all vertebrates. Expression is highest in the nervous system but is also expressed in other tissues throughout the body.

Transmission of scrapie prions from sheep to model organisms (1939) and kuru prions from humans to chimpanzees (1966) revealed PD's to be novel neurodegenerative disorders due to their

transmissibility(2). PD's occur in sporadic, inherited, iatrogenic and variant forms. Spread of infectious prions in Papua New Guinea (PNG) resulted in the kuru epidemic, the first observed human prion epidemic(3). Below is a summary of each of the major categories of human PD, and research findings that have led to improved understanding.

1.1.1 Sporadic Creutzfeldt-Jakob disease (sCJD)

The most common form of human PD is referred to as sporadic CJD (~ 85% of cases) as a result of cases lacking a genetic explanation and having not been exposed to prions iatrogenically or through dietary consumption(4). Incidence of sCJD is roughly one-two cases per million of population annually, and this incidence is broadly consistent across the globe in countries where rigorous CJD surveillance is present(5).

Disease course after onset of symptoms is relatively rapid for sCJD with death on average occurring six months after onset of symptoms and as quickly as 4-6 weeks, characterised by a rapidly progressive dementia(6). Patients can survive for several years in some cases, particularly where they have received assisted feeding(7). Diagnostic methods have improved in their sensitivity and specificity in recent years with improvements in magnetic resonance imaging (MRI) techniques and analysis of Cerebrospinal fluid (CSF) using Real-Time Quaking-Induced conversion (RT-QuIC) methods(6).

A large degree of the variation in clinical presentation, age of onset and disease course is modulated by genotype at the gene encoding the prion protein (*PRNP*) at codon 129(6, 8), with heterozygotes having much longer disease courses than homozygotes(7). In order to avoid misdiagnosis with inherited PD it is now common practice in many countries for potential sCJD patients to have genetic testing for *PRNP* pathogenic mutations, as some families may lack a family history due to late onset of PD, de novo mutation or previous misdiagnosis.

Surveillance has shown an increased incidence of sCJD in recent years. This may be due to improved ascertainment and changes in the age profile of populations(9). Some concerns have been raised that this elevation in incidence may be the result of novel, atypical cases of PD now being diagnosed as sCJD. A recent atypical case of vCJD with *PRNP* codon 129 heterozygosity displayed many clinical and pathological similarities to sCJD and was initially diagnosed as sCJD(10).

1.1.2 Inherited PD

Approximately 15% of annual deaths from PD are a result of inherited pathogenic mutations in the *PRNP* gene (11). Over 30 different mutations have been identified (see figure 1.1.) and classified so far (12). Different mutations can result in greatly varied clinical presentation, age of onset and duration of disease. Such variation has been noted within family pedigrees with the same *PRNP* mutation, implicating other factors that may affect disease pathogenesis(12).

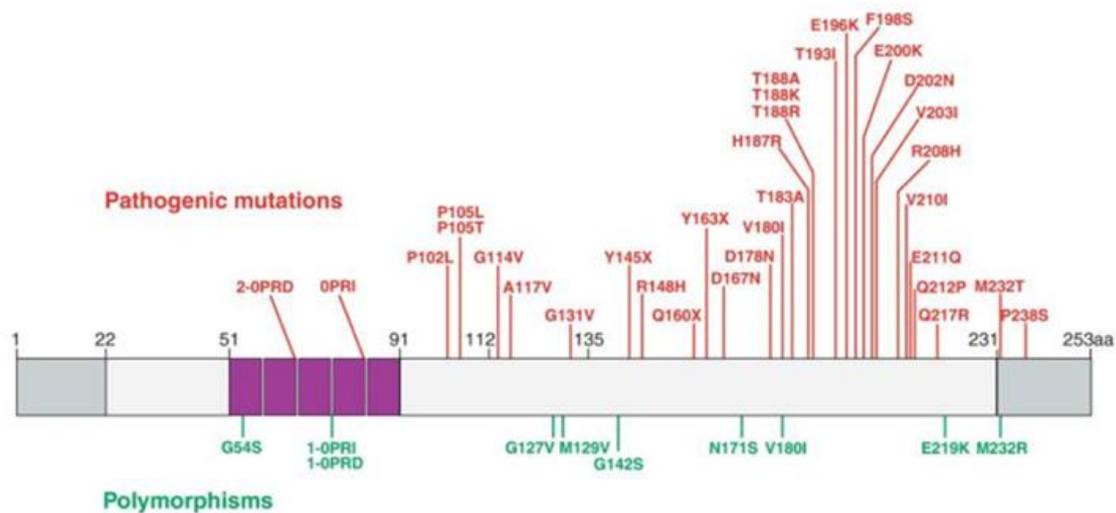


Figure 1.1. Schematic of pathogenic mutations discovered in the *PRNP* gene. Figure taken from 'Lloyd S, Mead S, Collinge J, Genetics of Prion Disease, Topics in current chemistry, 2011'.

PRNP mutations that have been shown to cause inherited PD can be broadly placed into three categories; protein truncating mutations, octapeptide repeat expansions and missense

mutations(13). Protein truncating mutations which cleave the *PRNP* gene product appear to be pathogenic if present after codon 130 in the single coding exon of *PRNP*(14). Theories have emerged hypothesising that this dependency on location of protein truncating variants is based on the naturally occurring *PRNP* being GPI anchored to the cell surface membrane, with late codon cleavage allowing a substantial amount of prion protein (PrP) to remain unanchored to the cell surface and with more freedom to participate in PrP recruitment. Octapeptide repeat expansions have been shown to be pathogenic when more than 4 additional repeat units are present in individuals and some have suggested this pathogenicity may be related to inherent PrP N-terminus toxicity(12), with mutated versions not having this toxicity mitigated by normal processes of regulation. The largest number of *PRNP* disease causing mutations come in the form of missense mutations where a change in nucleotide results in a change of amino acid sequence of PrP. Multiple theories have abounded as to the pathogenic mechanism for these mutations and PD(15). These include mutations impacting the native structure of PrP, affecting conformation, hydrophobicity and effects on usual glycosylation processes.

Estimates on penetrance of the various inherited pathogenic *PRNP* mutation have been enhanced recently due to the appearance of large datasets of whole genome sequence (WGS) data from thousands of individuals in multiple populations. This data has permitted improved estimates of population allele frequencies for these mutations(16). It is now understood that many previous estimates of penetrance were overestimated largely due to biases as a result of presentation of diseased individuals compared to non-symptomatic carriers of mutations(17). Challenges remain for giving precise estimates of penetrance for rare diseases like PD due to continuing biases for recruitment of individuals in sequencing projects, precise surveillance of PD individuals, relatedness of individual carriers in pedigrees affecting estimates and issues of population stratification.

1.1.3 Acquired PD

Although over 99% of cases of PD have come in the form of sporadic cases without explanation or inherited PD, there have been a number of cases of PD that have been the result of acquiring prion infection as an effect of environmental exposure.

1.1.4 Iatrogenic CJD (iCJD)

From 1958 until 1985 in the UK 1,849 individuals of short stature in childhood received treatment with cadaver-derived human growth hormone(18). In 1985 four cases of CJD were reported in the United Kingdom, Australia, United States and Europe of individuals who had received human growth hormone (HGH) treatment previously(18). The treatment with cadaver derived HGH was stopped immediately due to public health concerns of transmission of infectious prions through this route and replaced completely with newly available recombinant HGH. Despite this cessation, cases have continued to appear as a result of to the extensive incubation periods typically exhibited with iCJD(18). Worldwide, over 200 individuals have died of iatrogenic CJD. In the UK alone as of July 2018, 80 individuals have died of a total of 1,849 treated with cadaveric derived HGH. This represented a 4.5% risk of developing the disease(19). All cases of iCJD in the UK have been linked to a single preparation method of HGH known as the Hartree-modified Wilhelmi procedure (HWP) (18). Multiple preparation methods were used in different countries, and it is believed a size exclusion chromatography process in these other methodologies reduced transmission of infectious prions(20). Additionally, differences have been seen internationally in terms of variation in incubation of iCJD in the UK as a result of *PRNP* codon 129 genotype that is not seen elsewhere(18). This is believed to be due to possible PrP strain variation (see section 1.1.7 for discussion of prion strains) being exhibited in UK cases compared to international cases(18).

A 2015 paper reported amyloid beta protein (Abeta) pathology (parenchymal deposits and cerebral amyloid angiopathy) in four of eight individuals who died of iCJD raising concerns that other protein based neurodegenerative diseases may be transmissible through surgical processes(19). A 2018

paper aimed to assess whether this association had a biological basis. Samples of HWP HGH that had been stored for over 30 years were examined and showed the presence of Abeta and Tau aggregates that could potentially act as seeding material for the emergence of this pathology of iCJD patients(20). This material was then inoculated into mice brains directly and pathological examination of these mice showed the presence of Abeta pathology. This pathology was not shown in mice that were inoculated with recombinant HGH. It has been suggested that iCJD patients who had this pathology may have eventually developed cerebral amyloid angiopathy or Alzheimer's disease later in life if they hadn't developed iCJD beforehand.

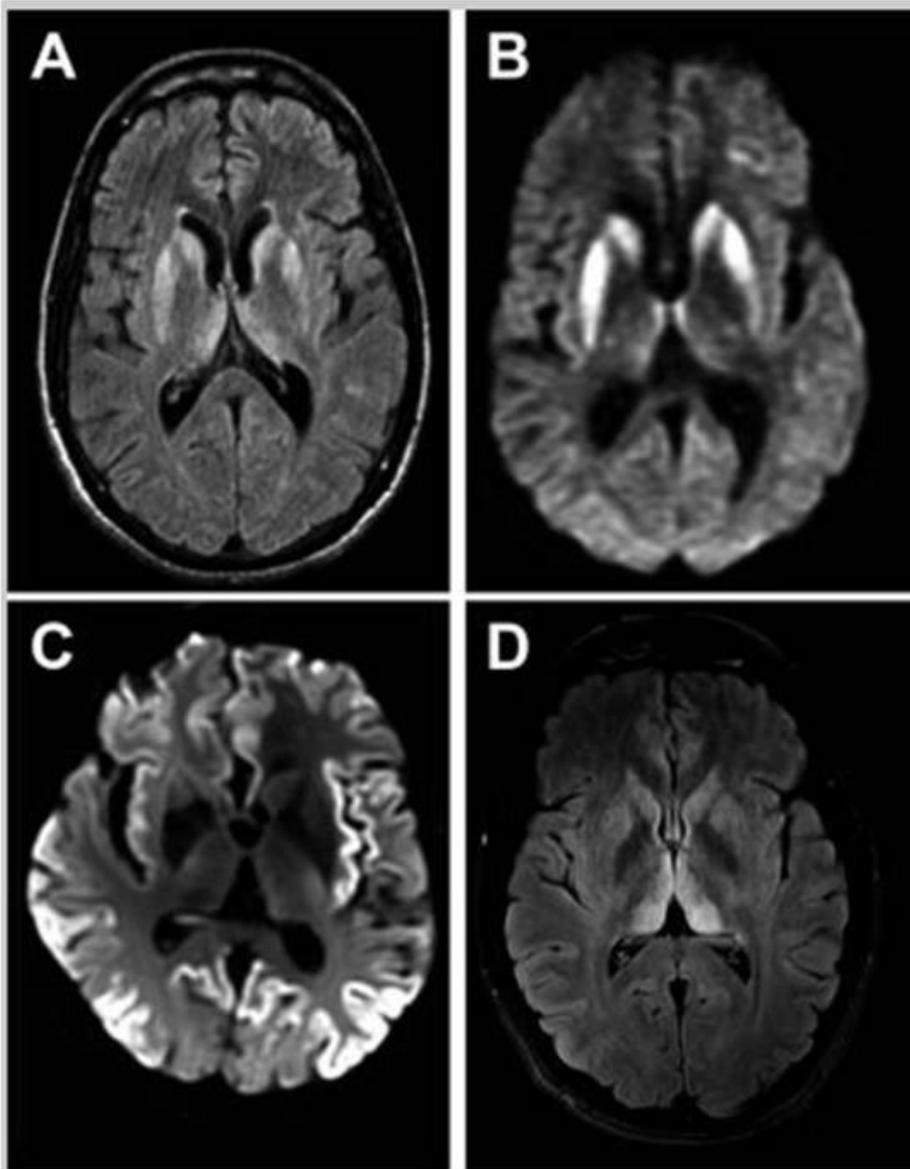
Another source of iCJD is the use of transplanted dura matter as grafts in neurosurgical procedures. Dura mater forms part of the meninges, the protective casing of the brain and spinal tissue. It is made up of connective tissue and historically grafts have been sourced from deceased persons(21). More than 200 cases of iCJD due to transplant of dura matter material have been observed globally due to iCJD surveillance programs(22). Over 50% of the cases that have been observed occurred in Japan, with this likely due to differences in surgical practices, with over 20,000 grafts being performed annually(21). The overwhelming majority of cases have been linked to exposure to a particular source of dura mater referred to as the Lyodura preparation, made in Germany(22). Six of the seven reported UK cases have been traced to this preparation and over 100 Japanese cases. The peak of the exposure to this preparation was between the years 1983-1987 with an overwhelming number of cases being individuals who underwent surgery in this period(21). There has been an observed reduction in cases of individuals who underwent surgery after this period, possibly due to stricter donor criteria and more stringent decontamination laboratory procedures(23). Of the seven cases observed in the United Kingdom until 2000, no spatial correlation was found between patients, making contamination of surgical instruments at a single site unlikely(21).

Clinical manifestation has been linked to the site of graft placement. Clinical presentation and autopsy findings show the most marked pathological changes occurring in transplant sites(21). Strain

examination through western blotting revealed type 1 PrP^{Sc} (infectious form of the prion protein, first studied in sheep with scrapie, an ovine form of PD), similar to sCJD rather than type 4 more common in variant CJD (vCJD) (see strains and phenotypic variability section 1.1.7 below)(21).

1.1.5 vCJD

A new form of acquired PD in humans appeared in 1996 when a British government minister confirmed cases of a fatal neurological disease observed in young people was the result of dietary exposure to infectious prions(24). These infectious prions entered the food chain as a result of an epidemic of Bovine Spongiform Encephalopathy (BSE) in the British cattle herd(24). The announcement by the British government led to public concerns that potentially millions of UK citizens had been exposed to deadly prions and a huge epidemic of this new variant CJD (vCJD) was imminent due to the extent of exposure in the British population. The BSE epidemic resulted in millions of cases of BSE in less than a decade. It has been estimated that 1,000,000 BSE infected cattle entered the food chain in the United Kingdom in the years 1980-86(25). Multiple theories have abounded regarding the cause of the BSE epidemic. It was blamed on agricultural practices in relation to animal feed, with cattle being given feed containing infectious prions. This was the result of recycled remnants of livestock including parts of the central nervous system used in the manufacture of the livestock feed. One theory originally offered was that meat and bone meal (MBM) being used as feed at the time was contaminated by sheep scrapie material(24). This was dismissed when it was shown that strains of BSE prions (see strains section below 1.1.7) and strains of scrapie prions are distinguishable(26). Another theory stated that the outbreak was initiated by a 'sporadic-like' form of PD (this sporadic form of BSE has not been fully established in practice) that occurred in cattle and was amplified through the feeding process(27). The British government took steps to mitigate the risk of further BSE infected cattle entering the food chain with the slaughter of over 5,000,000 cattle suspected of having BSE(28). Neighbouring countries instituted a complete ban on the importation of British beef in 1996(28).



Typical MRI features of Creutzfeldt-Jakob disease (CJD). (A and B) Sporadic CJD showing typical basal ganglia signal return on fluid-attenuated inversion recovery (FLAIR) (A), which is more obvious on diffusion-weighted sequences (B). (C) Diffusion-weighted imaging sequence showing striking cortical ribboning with normal basal ganglia in sporadic CJD. (D) Variant CJD showing pulvinar sign on the FLAIR sequence. Figure and legend taken from 'Mead S, Rudge P, CJD mimics and chameleons, Practical Neurology, 2017'.

vCJD has a markedly different clinical presentation compared to sCJD with a longer course (average 14 months) and different neuropathological presentation. vCJD is noted for its lengthy incubation periods between exposure to beef prions and manifestation of symptoms. Additionally, a prodromal psychological period can be seen, associated with severe depression and anxiety. The ultimate

stages of vCJD are similar to those of sCJD with the development of ataxia, involuntary limb movement and dementia(25).

Neuropathology differences (gliosis, plaques, site of brain) are believed to be a result of both the peripheral source of infection for vCJD compared to sCJD and also different strain properties (see later)(27).

Cases of vCJD reached a peak in the year 2000 and cases have only occurred once or twice annually since 2011. To present, 178 individuals have died in the United Kingdom from vCJD(5). There has been a handful of cases of vCJD in 11 other countries as a result of exports of British beef. To date, all cases of vCJD with the exception of a single definite and one possible case, had *PRNP* codon 129 genotypes of MM(29, 30). This demonstrates the strong protective effect of the *PRNP* codon 129 genotype, believed to be caused by a combination of the heterotypic protein-protein interactions in a heterozygote individual, and the incompatibility of the valine allele with the vCJD strain. It has been hypothesized that there will be a second wave of cases of individuals with different genotypes at this locus in the future(31). The total number of cases so far has been lower than feared initially, this is largely felt to be due to the 'transmission barrier' that exists between the bovine source of infectious prions and infected humans(27).

vCJD is distinct from sCJD in that it is associated with prion infection and abnormal PrP deposition in peripheral lymphoreticular tissues like tonsil, spleen and appendix tissue. Studies of tissue from appendix and tonsil removed surgically have revealed the presence of abnormal PrP immunohistology in 16 of 3,200 appendices examined(32). This has led to predictions that one in 2,000 UK individuals currently may be asymptomatic carriers of vCJD prions(32). Public health concerns have arisen regarding the possibility of transfer and infection of these prions between individuals as a result of blood donations, and organ and tissue transfers(32).

The impact of BSE and vCJD went far beyond the deaths of the individuals who developed vCJD. Over 5,000,000 cattle were slaughtered as a result of developing BSE or as part of measures to prevent BSE infected cattle from entering the food chain(28). These measures had a severe impact on the beef industry which was further affected after the European wide ban on imported British beef in 1996(28). This also led to increased strain between Britain and its neighbours diplomatically(28). Developments after the BSE/vCJD crisis included re-evaluation and increased regulation of the food, cosmetic and pharmaceutical industry. This impacted the availability and prices of products with downstream effects on employment in various sectors and inflation in the wider economy(28). Trust and effectiveness of scientific advice to government was damaged as a result of the crisis, with mechanisms to communicate key scientific ideas relating to public health risks being called into question(33).

There was erosion in public trust in the government with its handling of the crisis. This arguably contributed to the historic defeat of the Conservative government in the 1997 general election. Prior to the official government announcement in 1996 revealing the link between BSE and cases of vCJD, there had been numerous denials of such links between BSE and cases of vCJD. The Phillips report on BSE and vCJD in 2000(33) highlighted a number of failings in the government at various levels. This included a lack of coordination between departments and a culture of secrecy. There was a comprehensive review of practices at all levels of government, food industry and agricultural sectors. Criticism ranged from inadequate responses at the Prime Ministerial and cabinet levels of government down to practices and adherence to regulations at slaughter houses in Britain during the crisis. The consequences of the BSE/vCJD crisis highlight the potentially severe downstream impact public health concerns as a result of PD. This stresses the importance of vigilant adherence to health and safety regulations, an affective and coordinated response to any arising crisis, management of risks, and clear, transparent communication of the issues to the public at large(33).

1.1.6 Structure of cellular PrP (PrP^C) and PrP^{Sc}

After the discovery that abnormal forms of PrP comprised the infectious agent of TSEs, the precise three dimensional structure of PrP^{Sc} has remained elusive and much debated(34, 35). Full understanding of the structures would help better understand and illuminate processes of PrP propagation and how mutations in *PRNP* lead to clinical differences between individuals.

Initial attempts to characterise PrP^C were based on predictive modelling approaches, with many predicting that PrP^C structure is dominated by alpha-helices. When the PrP^C structure was analysed by Nuclear Magnetic Resonance imaging in transgenic mice, it revealed the presence of the three alpha helices and a short domain of beta sheet structure(36). Early predictive models of PrP^{Sc} pointed to a structure dominated by Beta sheet domains(36).

Ascertaining the structure of PrP^{Sc} has been more challenging due to its tendency to form insoluble amyloid fibrils and aggregates. This insolubility makes PrP^{Sc} unsuitable for characterisation by traditional means of molecular characterisation(37). A low resolution structure was eventually proposed after cryomicroscopic analysis of a non GPI-anchored form of PrP^{Sc}. The proposed structure consists of two intertwining fibrils composed of a beta sheet structure. This proposed structure has been referred to as a 'beta-solonoid'(38). Although how this structure is involved in conversion of PrP^C into PrP^{Sc} is not fully understood and is doubted by many experts in the field, it is felt that hydrogen bonding between the upper and lower rungs of the beta-solonoid will act as joining points for newly converted PrP^C.

1.1.7 Strains and phenotypic variability

Substantial variation has been observed between PD in terms of clinical presentation, histological lesion profiles and incubation times(39). Variation has also been observed within the same PD and shown to be independent of host genetic background through experimentation with transgenic mice(40). A large component of this variation is believed to be in part due to the presence of different PrP^{Sc} "strains". The term "strain" has been used to draw comparison with strains of other

infectious agents including influenza or tuberculosis. The mechanisms through which prion strain properties are conferred are currently unknown. It has been theorised that strain properties are caused by different conformational forms of the PrP^{Sc} that impact on disease spread and pathogenesis(12).

The concept of strains has been long held in the study of diseases spread through viral and bacterial pathogens(41). Understanding of the specific and original form of pathogen allows tracing of the disease to its original source and prevention of future outbreaks. In the 'protein only' model of PD this concept of strain variation is conveyed through different conformational forms of PrP^{Sc} that exist. These different strains consistently retain their properties through prion propagation processes in an organism, and after passage to different organisms(42).

Some of the clinical and histopathological variation observed in prion disorders is also attributable to host background genetics, and route of transmission of. In particular genotype at *PRNP* codon 129 has been demonstrated to play a significant role in strain variation. Inoculation of transgenic mice has shown the consistency of strain transmission through repeated inoculation in identical genetic backgrounds(40).

Conformation selection of particular PrP^{Sc} strains has been shown to play an important role in disease pathogenesis. It is believed that at stages of disease progression multiple PrP^{Sc} strains will exist, with only a subset being favourable for interaction with host PrP^C. The reduced incubation periods after repeated inoculations of transgenic mice in an 'adaptation' process reflect this, where favoured conformational forms are concentrated at each end point(43). Conformational selection plays an important role in infectious prions crossing the species barrier as seen in vCJD with only particular capable of interacting with host PrP^C. There is also interaction of strain type with host *PRNP* codon 129 genotype for vCJD. Only two cases of vCJD have been reported with a heterozygous genotype at this locus(10).

Classification of different prion strains in an internationally recognised system has been hampered by technical hurdles and disagreements over nomenclature. Agreement on to what degree slight biochemical differences as a result of distinct PrP^{Sc} conformation should be considered separate strains is a challenging issue and the wide variety of methodological and chemical means of classifying and obtaining different strains create challenges in developing an agreed upon system(42). A widely accepted classification system is based upon differential fragmentation of forms of PrP^{Sc} after digestion with protease enzymes, with different PrP^{Sc} conformations resulting in different points of PrP^{Sc} being cleaved. This has resulted in four major strains being recognised (according to the London classification system) with strain four being only found in vCJD and the others in varying ratios in other PD's. Complications exist with PrP cleavage also being affected by pH of the containing solution, and still limited understanding of attachment of metallic ions at the termini of the PrP^{Sc} unit.

Further understanding of strain effects and distribution of different PrP^{Sc} strains in the various PD's will help advance knowledge of PD aetiology, incubation and transmission. Discovery of peripheral PrP^{Sc} in asymptomatic carriers of vCJD strains will allow modelling of future cases of vCJD and permit assessment of risk of transmission due to blood transfers between individuals. Currently the causes of sCJD (the most commonly presented PD) are still not fully understood and the possibility of prion strains playing a role must be taken into consideration in addition to the roles of somatic pathogenic mutations and random stochastic neurological processes. Advancements in imaging that now allow three dimensional analysis of prions through electron microscopy(44) will also allow more precise and consistent molecular characterisation of the infectious prion material.

Attempts to understand the original source of the kuru outbreak (section 1.2.1) have relied upon understanding the strain properties of the infectious agent. Processing of brain homogenate from three individuals who died of kuru with protease enzymes showed similar fragmentation patterns to sCJD and not vCJD despite the more similar mode of transmission and dietary route of infection(45).

This is reflected in similar incubation period times in mice infected with infectious kuru brain homogenate material to those with infectious sCJD material(46). It is now believed that the original source of kuru was a case of sCJD in an individual in EHPNG and consumption of infectious prions at the mortuary ceremony upon this individual's death.

1.1.8 Diagnosis of PD and PD surveillance

PD's are ultimately confirmed with neuropathological examination. Until recently, clinical features of patients with PD could give clinicians some degree of confidence in believing a case of PD to exist but would ultimately need to be confirmed post mortem. Genetic testing for inherited forms of PD is now a part of the diagnostic process and genetic screening is now common practice for all individuals suspected of PD. The absence of one of the over 30 mutations in the *PRNP* gene can help rule out inherited PD as a cause, including in individuals without a family history of PD where a de novo mutation may have occurred. Additional methodologies have been trialled to quicken PD diagnosis which would allow improved care and counselling for patients and families of PD. MRI scans have shown distinctive findings for PD with individuals with vCJD showing a marked 'pulvinar' sign in about 90% cases(25). Distinctive MRI scans also have been noted for individuals with iCJD. Cerebrospinal fluid (CSF) can be examined for proteins released by damaged neurons and glia in response to neurodegeneration (eg. 14-3-3, S100b proteins) and for the presence of abnormal PrP that can seed recombinant protein in the real-time quaking induced conversion (Rt-QuIC) reaction. Whilst other CSF based measures to detect PD have been developed, all of these tools currently lack the required specificity, accuracy and accessibility to be comprehensive diagnostic tool for PD.

Current diagnostic approaches for PD in the United Kingdom are based on a system of 'definite', 'probable' and 'possible' diagnosis for each of the PD categories. Definite sCJD is confirmed with the presence of progressive neurological syndrome accompanied by either biochemical assay or neuropathological confirmation. Definite iCJD is confirmed with the identification of an iatrogenic risk factor such as past growth hormone treatment or neurosurgery. Definite genetic PD's are

confirmed by identification of one of the known mutations in the *PRNP* gene. Definite vCJD diagnosis is confirmed by the presence of a progressive neurological disorder with a course longer than six months and without any identifiable iatrogenic risk exposures or mutations in the *PRNP* gene. The development of a blood test for PD is an important goal that does not exist at present. The earliest possible diagnosis would lead to improved care and counselling for patients and potential new drugs to treat PD would be most effective prior to the onset of symptoms and neurodegeneration. Additional benefits of earlier and accurate diagnosis of PD include reduced costs to health services, particularly if improved outcomes through effective treatments can be achieved.

Comprehensive surveillance of PD's commenced in the United Kingdom in 1990 and was instituted in most other European countries by 1996. This was largely in response to the BSE epidemic in UK cattle in the 1980's and 1990's, and to check if any dietary exposure to BSE prions in populations would result in increases incidence of PD(25). Primary objectives of PD surveillance include centralised coordination and reporting of incidence of PD's, consistent and accurate classification of PD's, and consistent follow up guidelines for the 'definite', 'probable' and 'possible' cases that are diagnosed. Surveillance guidelines place specific emphasis on potential geographical clustering of PD, regulations for historical blood donations of individuals diagnosed with PD, paediatric surveillance and ascertainment of occupational exposure to PD.

The comprehensive structures and systems in place to ensure systematic surveillance of PD and its oversight across the globe requires considerable resourcing. The demands of PD surveillance involve interaction and guidance at the level of individual neurologists and neuropathologists who make routine referrals coordinated with international reporting systems of PD. Following the peak and subsequent decline in cases of vCJD since the year 2000 there have been some calls to reallocate resources away from PD surveillance(25). Others have argued that public health threats in the light of recent research demonstrating the potential prion-like transmissibility of other more common neurodegenerative diseases make maintaining the well-established surveillance practices of prion

surveillance relevant and important. Fears of potential carrier status of individuals for vCJD as the result of findings of PrP^{Sc} in peripheral lymphoreticular tissue have raised additional public health threats. These include transmission dangers due to surgical and other medical procedures.

Dismantling of the present-day existing surveillance structures would be a straight forward task but re-instating them in light of any new, future public health concern would be more challenging. The BSE/vCJD crisis highlights the unexpected and drastic impacts that such crises can generate and in disease surveillance can mitigate these impacts.

1.2 Kuru

Kuru is a fatal neurodegenerative condition (OMIM: 245300) observed in the Highland areas of Papua New Guinea (PNG) in the mid twentieth century with cases appearing until recently(47). The disease was concentrated in the Fore linguistic group in the Eastern Highlands region, with cases also occurring in neighbouring groups with whom the Fore were known to inter-marry(48). The disease has predictable stages of progression leading ultimately to death in all afflicted individuals(49).

The disease course lasts from a period of three months to a year(50). Disease progression has been classified into three major stages; ambulant, sedentary and terminal(51). The initial ambulant stage is associated with head and joint pain and some ataxia. After a period of some months, the disease progresses to the sedentary phase when walking can only be achieved with assistance from a physical aid. It is the appearance of tremors at this stage from which kuru has derived its name in the Fore language meaning shivering or shaking(51). During this stage, reflexes diminish to the point where the individual is unable to stand without support and eventually requires assistance to sit up. In the final terminal phase the sufferer is bed ridden, incontinent, and the eventual loss of swallowing reflex leads to difficulty eating. In many cases individual deaths occurs as a result of respiratory failure and septicaemia(52). Other comorbidities of kuru include a significant proportion of individuals who develop pneumonia at the terminal phase(51). Neuropathology demonstrated

that neurodegeneration was centred on the cerebellar portions of the brain and not the basal ganglia as seen in vCJD(53).

1.2.1 Kuru research

Kuru was first encountered by patrol officers of the Australian Administration in the early 1950's(54). Early reports by patrol officers in the region noted the devastating impact of the disease, which included a lack of young women available for marriage(55). The endemic practice of sorcery (kuru sorcery in particular), and its destabilising impact in the region was observed. The anthropologist Ronald Berndt reported retrospectively in 1958 on his time among the North Fore in 1953 about the presence of kuru. Berndt believed that the disease was a psycho-somatic response to the arrival of Western people, goods and practices in the region(56). The first medical examination of a kuru patient was performed by Vincent Zigas in 1955 who described the clinical presentation in the female patient as being due to an unexplained 'hysteria'(57).

After arriving in the Eastern Highlands in 1957 Carleton Gajdusek and Vin Zigas published the first case report of kuru, describing it as a completely novel disease(58). The disease was notable for its high incidence among the Fore people, its apparent clustering in families, sex imbalance in cases and the lack of immune response observed in sufferers that would be expected for infectious pathogens(58).

The clustering of cases of kuru in families and its geographic restriction to the Fore and neighbouring communities led to hypotheses of a genetic cause of kuru(3, 54, 55, 59). One such theory predicted the presence of a common genetic variant that was widespread in the affected region. This variant would have a high attack rate and was dominant in women. Genetic hypotheses led to some fears of spreading of such a 'kuru gene' beyond the Eastern Highlands as males were now being recruited for coastal labour exchange schemes(54). A proposal to quarantine the Fore region to prevent this was rejected by administrative authorities of the region(54). Among the arguments against a complete genetic explanation for the appearance of kuru was its high incidence, with over 1,000 cases in the

Fore region in the years 1957-62 (when surveillance was first conducted), and the high incidence in cases in years prior (understood to have happened through interviews with local residents)(49). The likelihood of a gene variant with such powerful negative effects on individuals being maintained in a population for a prolonged time was unlikely. Only if a variant conferred a balancing advantageous benefit may this have been the case. No such positive factor was evident to support such a balancing selection hypothesis. The complete genetic hypothesis was further damaged by the subsequent ethnographic and epidemiological work that was conducted in the region.

Anthropologists Ronald and Catherine Berndt spent some time in the Eastern Highlands region, and whilst observing the striking manifestations of the disease in individuals, attributed its cause to a psychosomatic response to the effects of sorcery practices in the area(49). Patrol officers also noted the practice of endocannibalism as part of mortuary rituals in the Fore region(54). Initial goals for epidemiological work were to clearly define the geographical region of kuru, which involved a comprehensive survey of villages in which cases had occurred. Investigators visited 177 villages and 155 of these villages went on to harbour individuals with kuru subsequently(49). Anthropologists Shirley and Robert Glasse were based among the Fore in 1962. They were tasked to live amongst the Fore and in addition to studying patterns of inheritance and gender relations conduct an extensive survey of kuru. This focused on creating pedigrees of individuals who had attended mortuary feasts. Their analysis, inconsistent with a simple genetic aetiology, showed that many of the individuals previously thought to have been related were in fact non-kin, but were considered family members through complex systems of shared residence and acceptance of non-kin into families(51).

Additionally, interviews with community members about historical cases of kuru revealed a recent and dynamic history for the incidence of kuru. Their work revealed that the first case of kuru was believed to have taken place in the village of Uwami in the Keiagana linguistic region close to the turn of the twentieth century and then spread to the North Fore region and not reaching the southern extremities of the South Fore region until the early 1930's(51). This recent predicted origin of kuru provided a strong argument against a genetic-only hypothesis as any genetic variant of such

recent origin would not likely increase in frequency to such a level to explain the observed incidence of kuru at the height of the epidemic.

The practice of mortuary anthropophagy in the kuru affected region was noted in reports by observers in the region in the early 1950's(54). Epidemiological and ethnographic work now focused more heavily on this as a possible explanatory mechanism for the spread and incidence of kuru.

Epidemiological work conducted by Michael Alpers and Shirley and Robert Glasse regarding consumption practices and participation rights revealed a striking correlation between the age and sex patterns of participants in mortuary feasts and incidence observed for kuru. Adult males did not participate in the consumption of human flesh due to beliefs of its polluting potential in Fore epistemology(49, 51). Children of both sexes were potentially exposed to infectious material when fed during the preparation process(51).

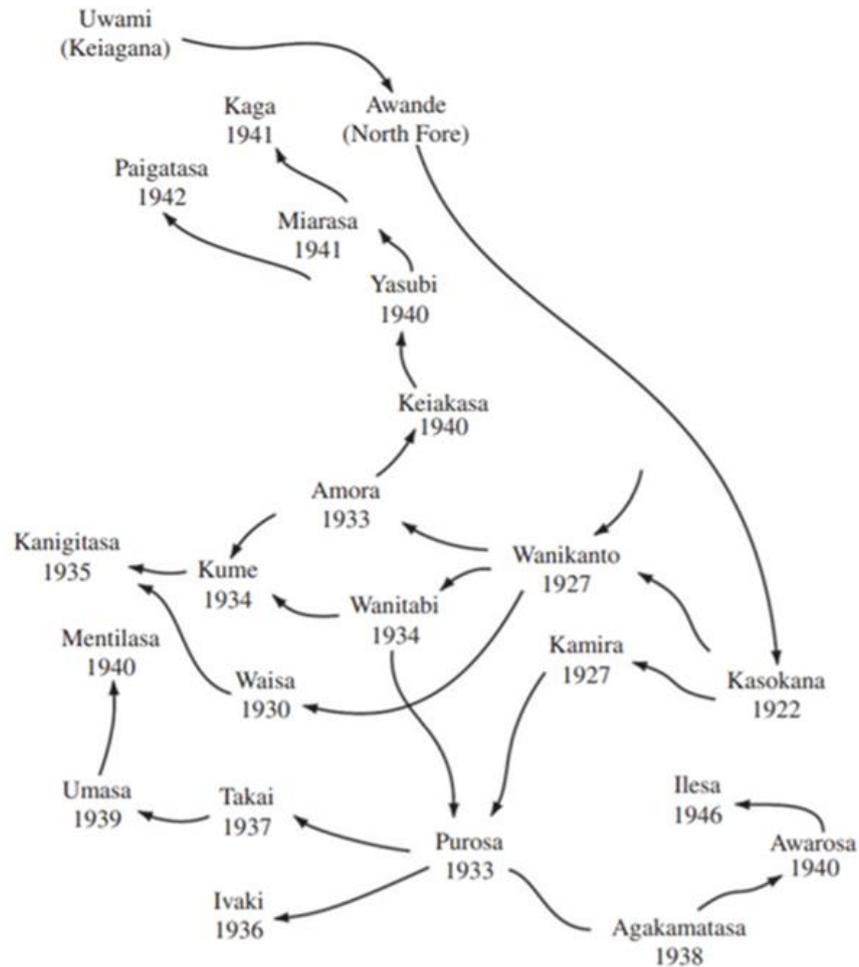


Figure 1.3. Proposed route of kuru through EHPNG. Figure taken from 'Lindenbaum S, Kuru sorcery: disease and danger in the New Guinea Highlands, 2016'.

The correlation between participation at mortuary feasts and incidence of kuru provided strong evidence for a transmissible agent for kuru (no human neurodegenerative disease had shown to be transmissible in nature). Subsequently a link was suggested by Hadlow in 1959 based on the similarities in neuropathology and clinical presentation between kuru and scrapie in sheep(60). This disease has existed for centuries and experimentally shown to be transmissible(61). In 1965 successful inoculation experiments in two Chimpanzees with brain homogenate from a young girl who died of kuru showed kuru to be caused by an infectious pathogen(62). The delay of clinical presentation of symptoms until(63) two years after initial inoculation, represented a prolonged incubation period that had not been observed for conventional viral and bacterial infections. Kuru also elicited no immune response from sufferers associated with traditional infectious agents. The

prolonged incubation periods also explain the difficulty in making an immediate connection between participation in mortuary feasts and kuru. A successful Chimpanzee inoculation of brain material from a CJD patient showed kuru to be part of a class of transmissible neuropathies. This work would ultimately lead to Carlton Gajdusek being awarded the Nobel prize for medicine in 1976 and triggered the search to characterise the nature of this novel pathogen(64).

The first ten years of intensive kuru research beginning with the first publication of a kuru case to the successful inoculation of kuru material and demonstration of its transmissibility had essentially resolved the major questions in relation to the mechanism of spread of kuru. Additionally, a marked drop in incidence of kuru was noted by the mid to late 1960's as a result of complete cessation of mortuary feasts by 1959. A complete absence of cases of kuru in children born after 1959 was observed through continued kuru surveillance confirming hypotheses of a shift in average age of onset predicted by the link to mortuary feasts(50). A lack of any cases in individuals born after 1960 in addition to the lowered incidence in males ruled out the possibility of vertical transmission(65). Investigators were confident that mortuary feasts were not happening surreptitiously due to the extensive community participation that feasts required. Relatives and kin would travel from long distances to partake and investigators believed that this would have been noticed. New work was now underway to understand the specific biochemical nature of this agent and its mechanism in disease processes. Many questions still remained regarding the mysterious variation in incubation periods displayed and for how long cases would still appear.

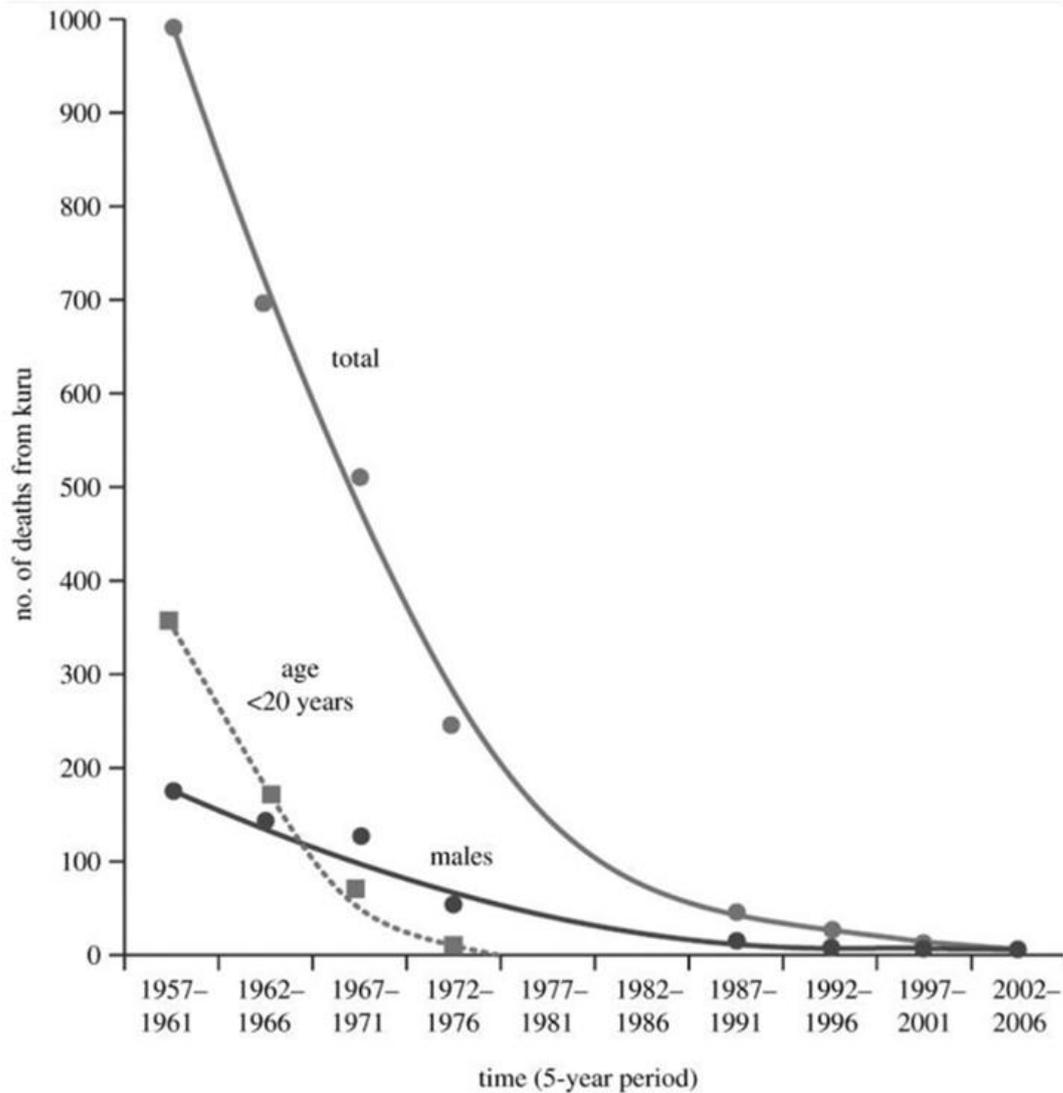


Figure 1.4. Proposed route of kuru through EHPNG. Figure taken from 'Alpers M P, Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end, Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2008.

Work to identify the causative agent of kuru and other transmissible encephalopathies culminated in the identification of a single protein agent initially named Protease resistance protein (PrP) due to its resistance to digestion with protease enzymes. Identification was permitted after a series of rigorous passage experiments using hamster model organisms(66). Prion infected brain homogenate material was gradually filtered and passaged to other animals until only this proteinaceous agent remained and infectivity was retained. The study author Stanley Prusiner would go on to coin the term 'prion' to describe this infectious protein agent, a departure from previous lexicon which included 'slow virus' and 'infectious amyloid'(67). This novel infectious agent did not appear to possess any genetic,

nucleic acid material. This led to a hypothesised 'protein-only' hypothesis for replication of an infectious agent involving recruitment of host encoded healthy PrP and conversion to diseased form through a templating process, a mechanism originally hypothesised by Griffiths in 1967 (Nature 1967). The gene that encoded PrP was identified and named *PRNP*. Mutations in this gene were identified and linked to cases of CJD that aggregated in families, (see section 1.1.2) further confirming the centrality of this gene and its protein product to PD processes.

The kuru epidemic was looked at with renewed interest in the 1990's due to the unfolding BSE/vCJD crisis in the United Kingdom (see section 1.1.5). Policy makers and experts looked to kuru, as the only example of an orally transmitted PD epidemic to model and predict the vCJD epidemic unfolding in the United Kingdom. Goals of this analysis were to predict the number of people infected and the number of eventual cases of vCJD(50). A collaboration was established between the UK based Medical Research Council (MRC) Prion Unit and the Papua New Guinea Institute of Medical Research (PNGIMR) in an attempt to gain a better understanding of both the British and Papuan epidemics(54).

An analysis of 11 kuru patients after 1996 who all had incubation periods in excess of 36 years showed eight of them to have heterozygous genotypes at *PRNP* codon 129(68). This indicated a link between a heterozygous genotype at this locus and prolonged incubation periods. New genetic research of individuals from the kuru affected region revealed a distinct lack of homozygote genotypes at *PRNP* codon 129 from what would be expected from Hardy-Weinberg equilibrium prediction of genotypes(69). This supported the hypothesis that the previously noted predisposition to sCJD was similarly relevant for kuru. All 11 cases observed between 1996-2004 were born before the 1950's, demonstrating that incubation periods could extend across much of a human lifetime. All 11 individuals resided in the South Fore linguistic region, confirming the staggered cessation of the practice of mortuary feasts. These were ended in the North Fore earlier due to earlier establishment of colonial administration and enforcement of prohibition of feasts. Genetic contributors to these

prolonged incubation periods beyond *PRNP* codon 129 genotype have not been established in kuru or other human PD. There has been an observed association of non *PRNP* variants and incubation period in more highly powered mice experiments(70).

Analysis of the strain type of kuru showed it to possess the same strain properties as sCJD and not vCJD(45). This ruled out a zoonotic source of the virulent strain of kuru that spread during the epidemic. The original source was more likely to be an individual who died from sCJD and who was consumed in a mortuary feast that was practiced at the time of death. Examination of a kuru patient also showed a lack of deposition of PrP^{Sc} in peripheral lymphoreticular tissue that is seen for vCJD(68), highlighting the heterogeneity displayed for the various human PD's.

Genetic analysis of kuru cases, exposed populations and unexposed individuals to kuru disease revealed an aberrant proportion of *PRNP* codon 129 heterozygotes in elderly women born prior to 1950 exposed to kuru(69). There was not a clear association of *PRNP* codon 129 homozygotes and the 152 kuru cases. It is believed that a lack of genotype association at this locus in cases of kuru was due to the nature of sampling of individuals. The majority of kuru case individuals were obtained after the peak of the epidemic and this cohort would have been enriched for heterozygote genotypes providing prolonged incubation periods (and possibly other protective genetic factors).

The prolonged incubation periods observed in kuru (in some cases over 50 years), and the stratification of incubation and attack rate related to *PRNP* codon 129 genotype in kuru studies led to concerns that the full extent of the vCJD epidemic in the United Kingdom would not be fully felt for some time. The vCJD epidemic in the United Kingdom could in fact be 'multiphasic' depending on *PRNP* codon 129 genotypes in the exposed population(54), and as of yet unidentified other genetic contributory factors contributing to incubation period.

A novel variant in *PRNP* at codon 127 in 12 women out of the 275 who were exposed to kuru disease in the dataset and not in any of the 152 cases of kuru was discovered. The variant was not found in

individuals analysed in various other non-kuru affected linguistic groups and geographically concentrated in inhabitants of villages in the Purosa valley at the epicentre of the kuru epidemic. Microsatellite data from carriers of the variant showed evidence that this variant was recent in origin due to the inferred extended length of the haplotype that this variant resides on. This variant demonstrated an episode of recent powerful selection due to a population response to a devastating epidemic. Inoculation experiments of transgenic mice expressing a form of *PRNP* with this novel *PRNP* codon 127 variant showed complete resistance to all forms of PD(71), confirming this variant's protective function against onset of kuru. The first genome wide association study (GWAS) of all human PD's was conducted in 2012(72). This did not reveal variants outside of *PRNP* that were enriched in elderly women highly exposed to the kuru pathogen compared to the healthy unexposed population. No statistically significant variant was found for variants associated with age of onset of kuru when this was measured as a quantitative trait(72).

Recent analysis of the different mortuary rites and participation of mortuary feasts in the kuru affected region have resolved previously unanswered questions regarding the initial spread and epidemiology of kuru(48). Kuru was believed to have started in Uwami in the Keiagana linguistic group and then to have spread to the North Fore village of Awande and then into the South Fore and not reaching the Gimi until the 1940's(48). Reasons as to why the greatest intensity of cases was in the South Fore and that the epidemic did not spread further into the peripheral linguistic groups was put down to mortuary practices. The peripheral non-Fore linguistic groups realised the hazards of eating kuru dead during mortuary feasts and ceased this practice(48). Customs regarding transumption of the dead were consistent across linguistic group rather than varying depending on smaller social units such as clan or hamlet(48). There was flexibility of customs at linguistic boundary regions with the Gimi border community Hepavina found to be adopting Fore transumption practices(48). The final case of kuru among the South Fore was observed in 2009, as opposed to 1993 for the Gimi and 1996 for the North Fore. Temporal differences in the cessation of kuru is believed to be caused by a contribution of a cultural axis reflecting imposition of colonial control and

also variations in mortuary practices instigated by the communities themselves(48). Within linguistic groups aberrations in practices in relation to mortuary feasts could be attributed to migrants from other region transferring their own cultural and mortuary practices.

1.2.2 Impact of kuru on affected communities

Kuru surveillance began in 1957 with monitoring of cases of kuru within the kuru region(49). There was great variation temporally, spatially and in the intensity of kuru disease. At the height of the epidemic there were 200 cases per year in the late 1950's(49). There was a perceptible drop in incidence during the 1960's attributed to the cessation of mortuary feasts (see kuru research section 1.2.1). In the worst affected villages in the Purosa valley, there was a noticeable lack of young woman in villages(59). Over 2,400 people died of kuru in the peak 20 years during the epidemic, in communities that had standing populations of approximately 40,000 individuals(49). This dropped to a rate of six per year in the late eighties and early nineties and single cases at the turn of the millennium(49). The final case of kuru was observed in 2009 and the end of the kuru epidemic was declared, with over 2,700 individuals dying from the disease after 1957(48).

At its height, the kuru epidemic was a dominant presence in the Fore region. It was the leading cause of death, particularly amongst adult women(51). Its effects were observable not just medically, but also in terms of its impact on social relations, marital practices, conflict, conflict resolution and migratory patterns(54, 59). There was great strain on families with care being required for sufferers almost constantly in the later stages of the disease, and the loss of mothers leaving young infants without traditional caregivers that placed great strain on extended families required to care for them(51). Women played vital roles in communities and tended the extensive vegetable gardens that surrounded villages(54). Amongst the Fore, kuru was attributed to the practice of kuru sorcery. This was a surreptitious practice where individual sorcerers would carry out, or be employed to carry out kuru sorcery methods against specified enemies. Kuru bundles would be created using twigs, mud, and physical material from the specified victim and then be placed in the forest(51).

Accusations of suspected sorcery were a great source of violent conflict between communities in the affected region(51). Desperation to find a treatment for this incurable disease led to families and communities taking desperate measures including visiting shamans and curers often in distant and remote villages in other linguistic groups that required arduous journeys and carrying of patients of rough and rugged terrain(51).

1.3 Population genetic research

1.3.1 Population genetic research in Papua New Guinea and the wider region

Since the first meeting of Europeans with the peoples of the Eastern Highlands in the 1920's there has been great interest into developing an understanding of the cultures, histories and origins of its peoples. This work has been performed in multiple fields including linguistics(73, 74), archaeology(75) and anthropology(76, 77). It became quickly apparent that the New Guinea Highlands represented a unique region in the world due to the extreme diversity of its cultures and languages.

Archaeological evidence points to the first presence of human settlements in the PNG Highlands at least 40,000 years ago (kya), followed by a prolonged period of isolation(78). At the time of original settlement, sea levels were lower than at present and New Guinea was part of a larger landmass known as the 'Sahul' which also comprised the Australian mainland and Tasmania(79). Genetic studies using the genomes of modern individuals, including modern PNG Highlanders and Aboriginal Australians predict a historic model of settlement of the Sahul(80). This model infers an ancestral split time of 42kya for the ancestors of modern day East-Asians and Europeans and then a split between Papuans and Aboriginal Australians 37kya. This split between aboriginal Australians and Papuans is many years before the disappearance of the land bridge between Australia and Papua New Guinea, reflecting ancient population structure in the region.

The island of New Guinea is renowned for its linguistic diversity. Of the approximately remaining 6,000 languages spoken on Earth over 900 are spoken on the island of New Guinea(73, 81). In the Highlands the average population per language is 2,500 and these languages can transition greatly in relatively short geographic distances(82). Broadly speaking, Papuan languages are classified as languages in the Melanesia region that are not Austronesian. Linguists have been unable to trace a common ancestor language for all Papuan languages unlike Austronesian languages, pointing to an ancient origin for such a proto-Papuan ancestral language.

The island of New Guinea is home to the third largest language family in the world, the Trans New Guinea Phylum (TNG) with a geographical extent across the island of New Guinea and extending to nearby islands including Timor, west of PNG(81). Over 300 languages are members of the family (up to 480 depending on inclusion criteria used). The radiation of these languages is believed to be due to the independent development of agriculture in the PNG Highlands centred on the cultivation of Taro(83). The epicentre of this radiation is believed to be close to the Central and Eastern Highlands of PNG as there is a high diversity of TNG languages spoken there and archaeological evidence pointing to ancient (up to 9kya) development of Taro cultivation in these regions(75).

The agricultural revolution that occurred in New Guinea independently of other similar processes occurring elsewhere is believed to have had a great impact on social, political, and ecological development of the region(75). Variations in population density, population expansions, hierarchical political structures, social customs and conflict have been linked to variation in uptake and practice of Taro cultivation(75). A general Western Highlands to Eastern Highlands cline has been noted in terms of the intensity of Taro cultivation(75), forest clearance, population density, and the presence of hierarchical political systems(75). The higher density of land more suited to Taro cultivation in Western points of the Highlands is believed to be the primary driver of these differences.

Many Austronesian language speaking groups reside along the New Guinea coastline. These peoples harbour ancestry as a result of the Austronesian expansion(84). The Austronesian expansion has

been associated with the movement of goods including dogs(85), rice based agriculture(86) and the distinctive Lapita pottery(87) that was transported in outrigger canoes that allowed more reliable movement over large bodies of water than previously. Linguistic(83) and genetic evidence(88) point to the origin of this expansion from island Taiwan, resulting in the spread of Austronesian language speaking groups across the Southern Pacific, Near and Remote Oceania and as far west and Madagascar. Analysis of non-recombining Y chromosome and mitochondrial DNA confirm the matrilineal practices of Austronesian peoples(88). Recent evidence has revealed that this westward migration has left a genetic signature on modern day Somalis in North-East Africa and Yemen in the Arabian peninsula, demonstrating the vast range of this expansion(89). Within PNG, coastal and island groups exist that harbour extensive Asian ancestry as a result of an admixture event between the previous inhabitants of Melanesia and incoming Austronesian peoples that began around 3,400 years ago. This appears to have been a dynamic process with varying amounts of Austronesian ancestry in different parts of Melanesia. Findings based on genetic evidence from ancient DNA obtained from individuals who lived in Vanuatu in the South Pacific 3,100-2,700 years ago show the expansion into the South Pacific to have proceeded in phases, with different groups of individuals with varying proportions of Papuan and Austronesian ancestry at different stages(90, 91). Interpretations of these findings place emphasis on Melanesia and PNG being a key point in onward Austronesian expansion into near and remote Oceania(92).

Extensive genetic differentiation between groups in PNG was reported in a study of 380 individuals from PNG from 85 different language groups from the PNG Southern Lowlands, Northern Lowlands, Bismarck Archipelago and PNG Highlands(84). No signature of an Asian ancestry component observed in coastal populations has been observed in highland populations, confirming a lack of observable genetic signature in highland populations as a result of the Austronesian expansion. The PNG Highlands groups show remarkable genetic differentiation between each other. The extent of this differentiation is not observed in other global sites where an agricultural Neolithic transition has occurred(84). This genetic differentiation is shared between coastal groups as well supporting the

idea that cultural and linguistic factors drive group differentiation more than geographical barriers. It has been suggested that this lack of genetic homogeneity in the PNG Highlands that is observed in Europe and Sub-Saharan Africa is due to the lack of a subsequent technological revolution such as the Bronze age in Europe and the Iron Age associated with the Bantu expansion in Sub-Saharan Africa(84). Analysis of WGS data from both PNG Highland and coastal individuals using the multiple sequentially Markovian coalescent (MSMC) approach calculated an estimated split time separating coastal and highland groups at approximately 20 thousand years ago (kya), and separation of Highland groups has occurred in the last 10kya. There are observations of a marked population increase in the last 10ky which has been attributed to the Neolithic transition associated with Taro cultivation(84).

1.3.2 Archaic introgression

Modern day Papuans are among the groups with the highest proportion of Archaic hominin derived DNA in their genomes(93). Hominins are species of the homo genus who are all more closely related to one another than to chimpanzees. Humans are the only remaining living hominin species but findings have revealed a wide range of different hominin species to have existed in the past(94). Approximately 2% of the DNA of modern Papuans (and all individuals with non-African ancestry) is derived from an introgression event that affected the ancestors of all present day Non-Africans(95). This admixture event is believed to have occurred 47-65kya(96). Additionally, after the discovery of archaic remains in a Siberian cave named Denis after a local holy man who resided there, DNA extraction revealed the discovery of a new archaic group that was named Denisova(93). Comparison with genomes of modern-day individuals showed present day Melanesians (including Papuans) to harbour approximately 5% of their DNA as a result of admixture with a Denisova-like ancestor(93). The vast distance between the site of the Denisovan remains and present day Melanesians has led to the conclusion that Denisovans inhabited huge swathes of present day Asia prior to their extinction. Lower proportions of Denisovan derived DNA have been suggested for other Asian populations that do not have Melanesian ancestry in some proportion. This includes south Asian groups, including

Tibetans who harbour a form of the *EPAS1* gene that is believed to be Denisovan derived and to provide benefits relating to living at higher elevations such as the Tibetan plateau(97). It has been argued that the variation in Denisovan ancestral components observed in Asia could be due to different pulses of introgression occurring at different periods in human history(98).

Attempts to further characterise patterns of archaic introgression in the genomes of present day individuals has led to the discovery of sites of increased archaic derived DNA which have experienced periods of positive selection after hybridisation with humans(98). Sites implicated with archaic introgression have been linked to traits such as altitude adaptation(97), lipid metabolism(98) and immunity(99). Additionally 'archaic deserts' have been observed at sites where there is a depletion of archaic derived DNA beyond what would be expected by chance fluctuations due to demography(98). Given the lower effective population size estimates of Neanderthal and Denisovan archaic groups, a higher load of deleterious mutations would be expected and negative selection against these would be likely after hybridisation. This has been supported by a general reduction of archaically introgressed DNA in gene coding regions of individuals with Archaic derived DNA(98). Additional depletion has been observed at the X-chromosome, which has previously been associated with reduced male fertility in hybrids. Archaic deserts in genes highly expressed in testes tissue have also been observed that support this hypothesis of reduced male fertility in hybrids(98, 99).

A recent publication using 161 WGS samples from individuals in Island South East Asia (ISEA) and Papua have now revealed a complex history of Denisovan introgression regionally within Asia, ISEA and Papua(99). Analysis of 'highly certain' Denisovan haplotype blocks in the genomes of these present day individuals has revealed the presence of three distinct Denisovan lineages that were separated by one another and from the sequenced Altai Denisovan individual by hundreds of thousands of years. The date of separation of a new lineage named D_1 from the Altai individual was estimated at 283kya and new lineage D_2 363kya using coalescent modelling. This deep in time divergence is sufficiently close to the Altai Neanderthal/Altai Denisovan split that these lineages

could be considered almost sister species. Geographical analysis of these lineages in modern day individuals show east Asians to be completely lacking in lineage D_1 , Papuans and neighbouring islands to have both lineages D_1 and D_2 reflecting geographical separation of these lineages as well as a temporal one. The restriction of the D_1 lineage to east of the Wallace line also makes possible that Denisovans were capable of traversing long tracts of open water that would have existed at the time of their settlement east of Wallacea. Attempts at dating the various introgression episodes using a series of simulations pointed to three discrete episodes with lineage D_2 introgressing most distantly in the past with the ancestors of present day ISEA and Papuans 46kya then lineage D_1 introgressing into Papuans 30kya and finally lineage D_0 that most closely resembles the sequenced Altai Denisovan individuals introgressed into the ancestors of modern day individuals from Siberia and East Asia ~15kya. Additional analysis of D_1/D_2 ancestry proportions between mainland Papua and New Britain reveal differences that cannot be explained by subsequent processes of drift alone but instead a subsequent, more recent series of introgressions approximately 16kya after the split between the ancestors of New Britain and mainland Papua. Again, this points to a very complex model for ancient population history and interaction with distinct hominin archaic hominin groups who had been resident and genetically distinct from one another for some time, perhaps a period greater than the first appearances of anatomically modern humans in the fossil record. The late introgression dates for the D_0 lineage into East Asians and D_1 into mainland Papuans show the late existence of these hominin groups, thousands of years after the believed extinction Neanderthal hominins in Eurasia. Analysis of adaptive legacy of the introgression of genetic material from divergent Denisovan lineages composed of Denisovan derived haplotype blocks revealed ontological enrichment for genes linked to immunity, adipogenesis and smooth muscle cell proliferation. Use of new haplotype-based statistic (nSL) techniques that examine enrichment of derived allelic variants in preference to ancestral alleles as evidence of selection were used. These tests showed an enrichment of several genes in the top 1% highest frequency haplotype blocks in Papuans in the top 5% of the nSL distribution for genes linked to functions in immunity and diet.

1.3.3 Eastern Highlands of Papua New Guinea (EHPNG)

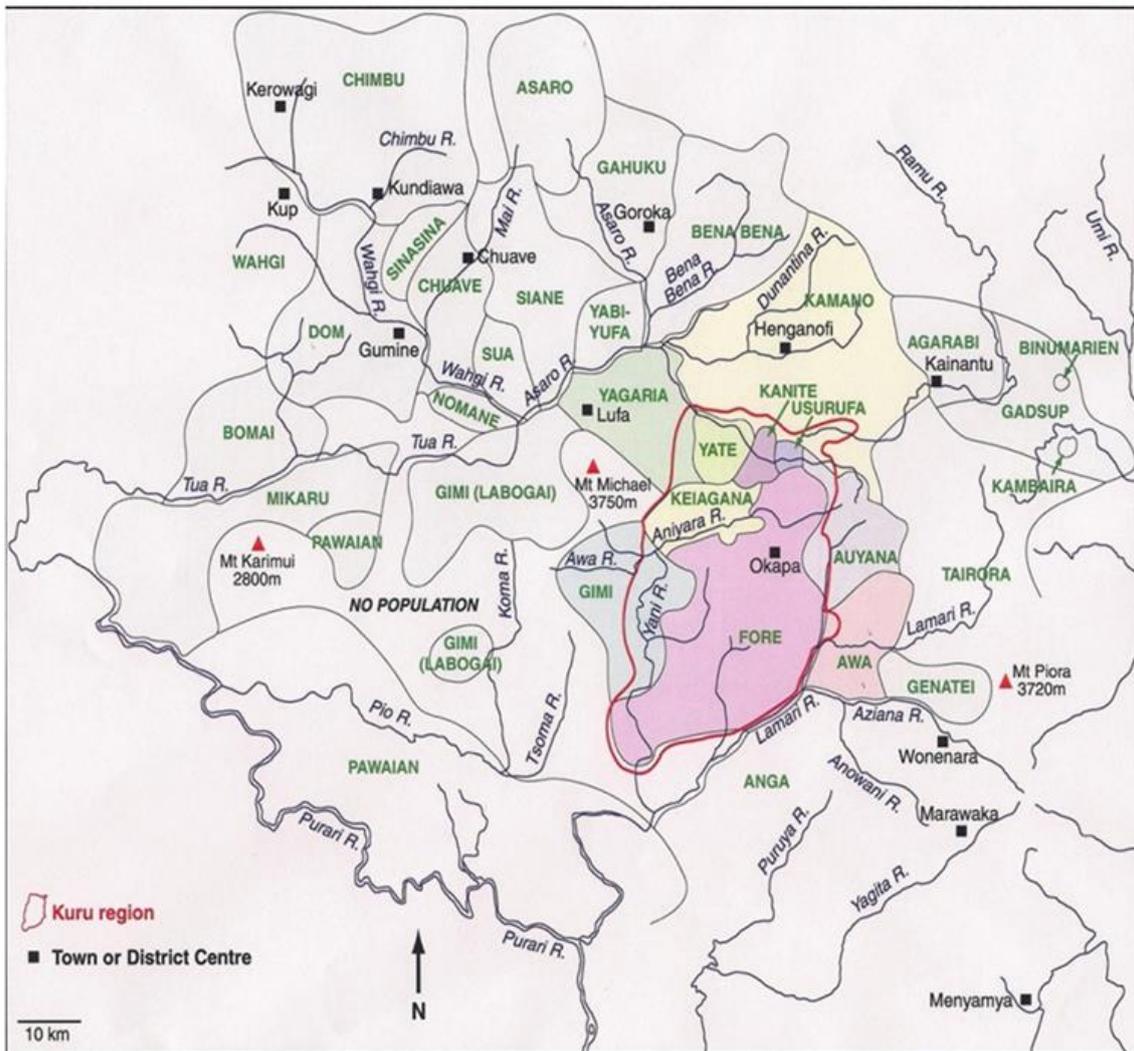


Figure 1.4. Map of EHPNG region showing placement of linguistic groups with the kuru-affected region circled in red.

EHPNG form a part of the wider Highlands community within PNG. It is home to 37 different linguistic groups in the region(100). The administrative capital of the region is Kainantu in the Kamano linguistic group. EHPNG communities were the first groups to make contact with European gold prospectors and missionaries who first ventured into the territory. This region is home to speakers of languages that form part of the Trans-New Guinea family(83). The majority of groups in the region speak either the central or eastern branches of this family, with the Anga and Pawaian language groups more difficult to classify and forming stock isolates(83). A general cline has been

observed in population densities observed in the region, with groups towards the south living at lower population densities to groups in the north and in particular the north-west.

Variation has existed in organisation of communities in the region, but in general in pre-contact EHPNG group organisation was centred on clan membership(101, 102). Collections of clans could form temporary fluid alliances for purposes of warfare and ceremonial pig exchanges(54). The formation of alliances was incredibly complex and not strictly restricted to individuals' hamlets of residence(103). Individuals within a hamlet may have had social ties or kin in other hamlets that were in conflict with individuals in their hamlet of residence. Disputes would often occur over land disputes, pig theft(54), sorcery accusations(51), and mistreatment of wives married into separate families(103). Decision making processes reflected the egalitarian nature of EHPNG groups with no individuals capable of coercing other individuals into action as a result of an elevated social position(104). Individuals could be influenced by persuasion or by strength of personality of individuals, but no individuals had elevated status formally recognised(103).

Social ties were maintained through complex processes of reciprocal exchange. In the Mount Hagen area of the EHPNG a ceremonial ritual known as Mokka was documented with hundreds of participants witnessing ritualistic exchange of goods including kina shells and pigs(105). Processes of reciprocal exchange and obligation could bind individuals from distant clans over long periods of time. Groups in the region had patrilocal patterns of residence with brides moving to the village of residence of their husband(105).

Studies of kinship practices in several groups in the region have revealed a common pattern of flexible and shallow kinship structures(77, 106). Individuals would generally share kinship based on association to a shared paternal ancestor but this ancestral reference point would tend to be shallow in time. Individuals or groups of individuals who would arrive in a region would often be accepted into groups openly, perhaps retaining their original residence as a kinship term but in subsequent generations this remote origin and lack of biological, genealogical kinship would be

forgotten. It is believed that this flexible kinship may be a response to the need to form large groups against the threat of neighbouring enemies. Flexible kinship systems are also believed to have aided the relatively recent territorial and population expansions that were the result of the introduction of the sweet potato (see section 1.4.2).

The carving of the region into a multitude of clans with complex and volatile relationships made travelling over long distance hazardous(105). This danger in addition to the extreme terrain in the EHPNG may have been impediments to processes of migration for individuals and groups of individuals in the region. A study of the Tairora linguistic group based on interviews with residents of numerous hamlets in the region demonstrate dynamic processes of migration historically(104). The model suggested for the Tairora postulates regular short distance movements of individuals between neighbouring clans as a result of non-kin adoption processes, marriage, residing with age mates in other villages and changing residence as a result of disputes and warfare. Additionally, less frequent mass migrations of many individuals would occur as a result of villages being routed as a result of warfare. Groups would often migrate large distances in order to escape conflict and further upheaval in their current area of residence and take up residence often in different linguistic groups entirely. The flexible kinship systems and the food surpluses evident in the EHPNG at the period of European contact would have helped facilitate acceptance of these groups of long distance migrants. Instances of these long distance migrations have been found in studies of the Awa(107), South Fore(103) and Kamano(101) linguistic groups.

A previous genetic study of linguistic groups in EHPNG was conducted using blood groups and serum protein polymorphisms acting on 20 linguistic groups in the Eastern Highlands(108). The measures of genetic distance between the groups corresponded to geographical distances, and coefficients of kinship reflected the observed population densities. Levels of heterozygosity were not found to vary as a function of population size and density. The findings of this study were based on a limited number of protein polymorphisms that could not be considered to be truly neutral markers. In

population genetics, variants not under selection are preferred, as changes in allele frequency caused by selection pressure can greatly obscure any signals due to historic demographic change.

A study of the Pawaian linguistic group who live in the southern extremities of the region in forests at extremely low population densities, confirm genetically their isolated pattern of residence and interaction with other groups in EHPNG(109). The findings of this study are based upon only three class 1 human leucocyte antigens and their relative frequencies compared to other highland groups. Inferences in this study including the degrees of admixture with other highland groups might be investigated more thoroughly with a more comprehensive dataset.

1.3.4 The Fore people

Approximately 15,000 Fore speaking peoples occupy the southern extremities of EHPNG. They have been divided into 2 separate groupings based on linguistic differences, with the North Fore speaking the Ibusa dialect and the South Fore speaking Atigina, Purosa and Ilesa dialects(103). The region came to prominence in both medical, ethnographic, epidemiological circles due to its association with kuru.

Like other groups in the region the Fore held a common belief system centred around the 'five souls' of an individual(110). Fore purifying rituals had the requirement that men have restricted contact with 'polluting' elements. This precluded regular contact with women or feminine physical material. Adult men among the Fore would live in separate men's' houses from women and children. This cosmology translated into Fore fascination and concern with the presence and threat of sorcerers. Sorcerers would often use polluting elements such as hair, and menstrual or faecal matter from prospective targets and use them in ceremonies. The degree of fear of sorcery was so great that Fore individuals would be very guarded with their physical material, resulting in deep drop trenches acting as latrines and ensuring that any physical possessions or personal material would not remain in other villages after visits. The culture of sorcery that pervaded the Fore was exceptional by the

standards of EHPNG. Sorcery accusations would often result in conflicts and much energy was dedicated to investigating and resolving sorcery accusations.

1.4 Change in EHPNG

1.4.1 Impact of European contact and administration

The height of the kuru epidemic also coincided with remarkable changes in the way of life of the inhabitants of EHPNG as a result of European contact and eventual administration(103). After initial forays into EHPNG by Australian gold prospectors and Lutheran missionaries, EHPNG gradually came under administrative control by the Australian government. Patrol posts with resident officers were put in place in Kainantu in 1933 and Okapa in 1954(54). Field stores were opened in Kainantu, Okapa and Purosa. This flow of Western goods led to the gradual monetisation of the economy in the region which had traditionally functioned through complex systems of gift exchange. The arrival of Western goods also resulted in notable phenomena of cargo cults and rituals in the region, with communities ritualising their desires to acquire these previously unseen luxuries that were now appearing(51).

Modern infrastructure began to appear in the region with an airstrip at Kainantu and the creation of roads allowing vehicle access in a matter of hours between points previously inaccessible by vehicles. These points were only historically accessible through complex networks of trails with access only permitted as a result of a personal connection to the owner of the trail(103). Use of roads without permission or social tie to the owner was deterred by fears of sorcery and possible violent attacks. After initial attempts to commission roads, the construction of a road allowing 4X4 vehicle access between Purosa and Okapa was a landmark event. Previous circuitous journeys through the complex trail network could now be avoided by using this road as it was 'government owned' in the eyes of the local inhabitants. The road also facilitated an increase in the flow of new goods into the region. Construction of trade stores along the roadside were accompanied by

individuals previously resident in hamlets further afield moving their residence to be closer to this new hub of economic activity(103).

Colonial administration led to an outlawing and gradual cessation to the practices of warfare between rival clans that had had historically occurred in the region. Deaths from warfare were the largest cause of death, and its cessation meant that the region was no longer divided into hundreds of warring territories that were hazardous to traverse for individuals. The Fore appeared relieved to have the constant threat of conflict and its associated upheavals ended. Time previously devoted to forming and maintaining war alliances and mediating the numerous disputes was now available to be put to other pursuits(49). Individuals now had greater freedom of movement to move through the region and visit distant places that were unimaginable previously. Fore males used these opportunities to seek employment and would return from distant labour exchanges with new skills and knowledge. This was particularly the case with experiences cultivating cash crops such as coffee which grew easily on the mountainous slopes in Fore territory. Less than ten years after Australian administration began, Fore communities began to cultivate and sell coffee to buyers further afield in exchange for currency(54). This could be used to acquire lucrative goods, particularly steel axes, shovels, cloth and salt. Further roads were constructed autonomously by communities to facilitate improved access and contact with buyers of cash crops, and eventually resulted in even the remotest of Fore communities being a matter of hours away from roads, western goods and medicine.

Western control and administration also resulted in new medical experiences for inhabitants, both positive and negative. The arrival of coastal individuals and Europeans brought with them diseases that Eastern Highlanders had not encountered before and had no immunity(105). This resulted in outbreaks of measles and other European borne diseases(111). Treatment and initiation of vaccination programs against endemic diseases such as yaws and pot belly resulted in drastic reductions in these complaints and improvements in rates of mortality, particularly infant mortality.

1.4.2 Ipomean Revolution

Much discussion has centred around the possibility that the Highlands of PNG was undergoing an 'Ipomean revolution' as a result of the introduction and impact of the sweet potato prior to European contact(112). Estimates vary in relation to the timing of its arrival for the various communities and its route through the Highlands. It is believed to have impacted on subsistence methods, patterns of residence and social customs. These changes have come as a result of benefits gained from cultivating the sweet potato compared to indigenous staple crops such as the taro. Superior yields in sweet potato are believed to have led to a 'population explosion' and the ability to feed domesticated pigs with food surpluses(103). The sweet potato could be cultivated on mountain slopes at higher altitudes which resulted in movement of residences to higher altitudes. The population increase associated with the Ipomean revolution and opening of previous virgin lands led to cultural practices that accommodated the segmentation of groups and flexibility in adapting to new environments. The hallmarks of the arrival of the sweet potato can be clearly seen in Fore communities who were believed to have only have received the sweet potato approximately 100 years prior to European contact(103). Fore groups are noted for having a particular egalitarian social structure, flexible kinship methods and openness to new ideas and practices (which was apparent with uptake of Western goods and ideas upon contact and administration)(51). This new expansion would have led to tensions when expanding groups without previous social ties came into contact or when land reached its carrying capacity. Complex ceremonial pig exchanges that were observed during the early years of colonial contact are believed to have been a recent development and mechanism aimed at alleviating these stresses(113).

The introduction of the sweet potato has been described as placing populations in a proto-agricultural state intermediate between hunter-gather subsistence and settled agriculture(103). The arrival of the new crop allowed expansion and into new territories until ecological and demographic factors prevented further expansion and settled agriculture had to be adopted. The variation in population density, social customs and conflict between the various groupings in EHPNG have been

argued to reflect the various stages these groups were experiencing during the Ipomean revolution. There is a gradual cline in decreasing population from the Asaro (103 persons/sq. mile) and Gahuku (83 persons/sq. mile) to the BenaBena (82 persons/sq. mile) located centrally and the North (54 persons/sq. mile) and South Fore (27 persons/sq. mile) and The Anga located beyond the Lamari river(103). This trend is associated with when the crop was first adopted, with the Anga only recently beginning to use it and the Asaro believed to have used it for over 200 years. Linguistic groups in the north-west of EHPNG have far more elaborate pig exchange mechanism with the 'idza nama' feasts taking place amongst the Gahuku often having over 1,000 attendees(103). The Ipomean revolution is believed to have had such a drastic impact on the way of life in the Highlands of PNG that much of the observed residence patterns, subsistence methods and social structures that were observed upon contact were in fact relatively recent innovations and not ancient structures and traditions.

The impact of the 'Ipomean revolution', the kuru epidemic and Australian administration of the Eastern Highlands would likely have resulted in marked demographic changes in a short period of time to accompany the overhaul in the way of life that these communities were experiencing in such a short period of time. These events began at different times temporally, but with significant overlap and with contrasting demographic effects on the communities that were impacted. Impacts on patterns of genetic diversity as a result of these events due to changes in population size, migration and admixture are interesting potential avenues of research.

1.5 Understanding population responses to natural selection pressure

1.5.1 Use of genetic data to understand natural selection in human populations

Natural selection can be considered as the change in frequency of a trait in an organism as a result of a populations' response to a selection pressure. If such a selection pressure (e.g. change in environment or the presence of novel pathogen) is sufficiently powerful and persists for enough time, these changes can become fixed in a population. At a genetic level, selection is caused by the

appearance of novel genetic variants in a population as a result of random genetic mutations that impact characteristics and create variation within a population.

Genetic variation can fluctuate within a population as a result of chance selection of genetic alleles during meiosis. Changes in frequencies of alleles due to this 'genetic drift' are more pronounced in smaller populations and this genetic drift alone will result in genetic variants becoming fixed or lost in a population without any influence of natural selection(114). Over many generations the effects of genetic drift on a particular locus' frequency are stochastic and directionless and usually require many generations to reach fixation or be lost completely in outbred populations.

Natural selection can be subdivided into various modes reflecting the various genetic signatures that are present in populations. 'Positive selection' is when a novel mutation appears in a population, is positively selected due to its beneficial effect in carriers and reaches high frequency or becomes fixed(115). The directional nature of the variants effect usually leads to these variants becoming fixed in a shorter period than variants fluctuating due to genetic drift alone.

Often, selected genetic polymorphisms already exist in a population when there is a change in selection pressure resulting in one of the variants present increasing in frequency. The result is that if this variant had persisted in a population for a long time prior to this selective sweep the variant will be present on a multitude of genetic backgrounds. This will complicate attempts to find such a variant as the variety of genetic backgrounds on which it resides will dampen any signal.

Episodes of positive selection are easiest to detect when the effect of the variant is powerful, there is a lack of genetic heterogeneity with the variant having a proximal connection to the trait in question, when the variant is recent in origin and has risen quickly to high frequency, and also when the variant is present in some populations but not others.

Balancing selection occurs when genetic diversity at a locus is maintained in a population instead of a single variant rising to fixation and the alternative variant being lost. This diversity can be

maintained as the result of 'heterozygote advantage' at a locus where a heterozygote genotype confers a selection benefit for organisms compared to being a homozygote. The classical example of such a situation is in relation to malaria where a heterozygous phenotype is advantageous in locations with high exposure to the disease transmitting parasite(116). Frequency dependent selection is where genetic diversity is maintained by having diversity maintained through fitness being optimal with allelic variants being maintained at certain frequencies. This is most notable for loci that have phenotypic impacts in relation to immunity(117).

Novel genetic variation can appear in a population as a result of admixture with a previously external group(118). New genetic variation that exists in the new admixed population may be selected (negatively and positively). Analysis of regions of the genomes enriched or depleted in the DNA from a source of the admixture event beyond what would be expected by chance can provide evidence for such 'introgression'.

A major challenge for researchers when attempting to find genetic variants that have been affected by natural selection is to distinguish between changes of frequency that have occurred due to natural selection or simply due to drift alone. Results of simulations demonstrate that what may at first appear a pattern of genetic diversity at a locus due to positive selection can in fact be the result of a historic population bottleneck followed by genetic drift the population(119). Distinguishing between these scenarios is assisted with a thorough understanding of the population demography and population history of the populations under study.

1.5.2 Use of selection techniques in Oceania and other under-studied populations

The majority of studies assessing the impact of natural selection on populations have primarily been focused on populations of individuals with predominantly European ancestral history(120). Certain problems have emerged in the field as a result of this 'Eurocentric' bias. A result of this bias has been the development of technologies based on particular patterns of genetic variation that exist in these populations. This will include genotyping platforms that aim to examine genetic variation known to

exist in these populations and the creation of genetic maps based on known patterns of linkage disequilibrium(121).

Detecting natural selection in novel, understudied populations present numerous challenges both technical and theoretical. The population history of such populations tends to be more poorly understood, especially where there are no written historical records of major past demographic events that will have affected genetic diversity(122). Understanding of population demography and inference of ancestral history are often aided by the use of standardised genetic parameters.

Effective population size is an often used term that describes the idealised size of a population that is in conformity of the assumptions of the Wright-Fisher model. These assumptions include random mating and a constant population size. Inference of this parameter in real world populations and comparisons with the actual census population can provide valuable insights into the population history of a population. This is apparent for human populations as a whole, who display greatly reduced effective population sizes compared to the current global census population. This reflects a historical population bottleneck in human history followed by a population expansion. This has been used to strengthen an 'out of Africa' model of human history(123). Kinship coefficient is a measure the degree of relatedness in a population. It can be understood as the probability that two alleles selected at random are identical by descent (IBD). Understudied populations can often have extreme values for effective population size and kinship coefficients. Phenotypic measurements that are standardised in conventional studies may be more difficult to obtain or may be unobtainable in under-studied populations(124).

Evolutionary studies of understudied populations also present opportunities and advantages compared to conventional populations. Variants that may be under selective pressure and at low population allele frequencies in conventional outbred populations may have risen to higher frequencies in other populations where genetic drift may have had an additional contributory effect on the variants frequency(125). This may lead to variants that have a functional effect being

detectable in novel populations where they were previously undetectable. Additionally, novel populations with greatly diverged population history may display instances of convergent evolution where different genetic mechanisms evolve in response to the same phenotypic pressure in different populations. Examples of this include altitude adaptation(126), metabolism(127) and immunity(128).

1.6 Understanding of genetic architecture of traits

The genetic architecture of a trait or phenotype can be considered the summation of contributions of genetic factors to variability of that trait or phenotype(129). This will include single nucleotide variants, larger structural genetic variants, and interaction of these genetic variants with one another. Variation in the appearance of a trait in a population depends upon a combination of the complex and dynamic biological processes underpinning the phenotype and interaction with environmental factors(129). The broad-sense heritability of a trait (proportion of variance of a phenotype contributed by genetic variance in a population) will vary as the variation in environmental conditions varies from place to place and over time. Traits with the same amount of broad-sense heritability can also have markedly different underlying genetic architectures. This can be contrasted by the examples of Phenylketonuria (PKU) and height as traits. All population variance in PKU can be attributed to a single genetic mutation that is completely penetrant(130). Genetic variants linked to differences in height now number in the hundreds with each only explaining a small proportion of the total heritability(131). This reflects the fact that height as a trait is sufficiently distal to individual genetic variants and is instead affected by numerous and complex biological mechanisms related to hundreds of genetic variants, whereas PKU is proximally related to a single step in a biological pathway that is controlled by a single gene product. Genetic architecture is sensitive to how a trait is defined and measured. A trait such as bone density can have greatly differing genetic architecture depending on where bone density is measured in the body(113).

Full understanding of the genetic architecture of a trait will depend on the number of genetic variants affecting the phenotype, their effect sizes, and their allele frequencies in the population under study. Genetic architectures have previously been broadly defined as monogenic, oligogenic, polygenic and omnigenic(129). Improved understanding of genetic architecture in recent decades has been enhanced by improved genotyping technologies and studies having greatly increased number of participants, improving statistical power.

Benefits of understanding of genetic architecture include advancements in disease screening, diagnosis and prognosis. Identification of high risk individuals prior to onset of symptoms can allow preventative measures to be taken and savings in resources(132). An improved understanding of genetic architecture can lead to finding new drug targets in relation to a trait. Even if a variant has a low effect size in relation to a trait, targeting its protein product in a different manner may have profound impacts on the trait of interest. This has been shown with the development of drugs that cause inhibition of HMG-CoA causing downstream reduction in LDL cholesterol levels in the blood(133). This is despite the variant with the highest association with LDL cholesterol in the *HMGCR* gene only explaining 0.26% of the variance in LDL cholesterol levels(134). Further elucidation of the genetic architecture of a trait will permit improved understanding of the underlying biology of the trait of interest and its similarity to other traits and phenotypes of interest. This will allow further understanding of evolutionary processes that have sculpted human genetic variation and its importance to different phenotypes and also inform future study designs including assaying risk variants by Mendelian randomization procedures(135).

There have been concerns regarding correctly assigning rare variants in studies. Genome wide association studies (GWAS) have only been concerned with common variants ($MAF > 1\%$) to avoid this complication and to ensure sufficient statistical power is achieved for variants under study. Previous findings from GWAS results may in fact be attributable to synthetic findings as a result of being in weak linkage disequilibrium with rare variants with strong effect sizes(136). It is hoped rare

variants in the future will be better studied with a movement towards WGS data and whole exome sequence (WES) data as this becomes more affordable. Problems will still remain as some parts of the genome are less reliably sequenced than others, particularly highly repetitive regions(137). Deep imputation of genotyped datasets will also permit better understanding of the effect of rare variants on genetic architecture(138). This process requires the use of haplotype based imputation panels, which to date are heavily biased towards European and East Asian populations and will therefore have limitations (see below). This problem is exacerbated when trying to draw global conclusions about the genetic architecture of a trait based on studies heavily biased to participants with European or East Asian ancestry. Conclusions drawn from these primary studies may not be valid in other populations due to divergent ancestral histories, separate episodes of selection pressure between the populations, environmental differences between the population and how traits are measured and defined in populations(139). Understanding of genetic architecture globally will lead to improved healthcare management and allocation of resources. Benefits will be maximised when not only more detailed genetic analysis is undertaken but also when biases that currently exist in research are ameliorated.

The genetic architecture of PD's in general and kuru in particular is still not fully understood. Improved understanding of genetic architecture could be of great benefit to enhancing knowledge of the underlying biology of kuru, PD's and neurodegenerative diseases in general. Study of disease causing variants shown to be extremely rare in populations but clustering in families with inherited PD (section 1.1.2) have been shown to display varying penetrance, effects on the course of disease and age of onset(13). The only common single nucleotide variant that has so far been associated with PD's is the codon 129 polymorphism at *PRNP* (see section 1.1). The heterozygote advantage conferred at this locus is problematic for methodologies that detect association based on narrow sense, additive approaches. Care must be taken to ensure that powerful, common variants are not overlooked in assays due to complex interactions with the trait being measured.

Study of variants not observed in European and other well-studied population will provide further insights in gene function and motivate studies in under-studied populations. A variant was found in individuals in East and South Asia and shown to be protective against sCJD(140). The protective variant at *PRNP* 127V shown to be geographically concentrated to the Fore people in EHPNG was found with a focused approach on the *PRNP* gene. Discovery of disease associated variants in other loci through analysis of understudied populations would provide a major advancement in PD understanding.

Results from association analysis of other neurodegenerative diseases have greatly increased the number of loci associated with neurodegeneration phenotypes(141, 142). This has been aided by the increase in sample sizes that investigators have been able to recruit, genotype and process effectively. Given the rare nature of PD's, it will be impossible to recruit tens or hundreds of thousands of participants that is possible for common diseases, posing challenges to discovering variants and genetic loci outside of *PRNP* that play a role in the genetic architecture of this trait.

Understanding of genetic architecture in the past has been based on the compilation of statistically robust findings in relation to single genetic variant effects on phenotype manifestation. Heritability studies have previously provided estimates of heritability for many traits and the proportion of heritability that these findings represent is only a fraction of this (referred to as the 'missing heritability' problem). Development of techniques that accumulate the effect sizes additively of independent loci (the overwhelming majority of which are not individually associated with the trait of interest) to give an overall polygenic risk score (PRS) has resulted in a greater degree of this 'missing heritability' being explained. This has led some to re-evaluate the genetic architecture for some traits, particularly common diseases with it now being believed that in some cases a large proportion of the heritability is represented by the cumulative impact of thousands of loci on a multitude on complex and interrelated biological pathways that affect the trait of interest(132). A polygenic risk score is usually developed on a primary dataset of interest which takes the form of

findings from a highly powered association study and then the findings for the variants in this analysis (effect size, frequency etc.). This is then applied to individuals to estimate their polygenic risk for the trait of interest. A publication based on the use of polygenic risk scores for common diseases showed that over five percent of the control populations showed polygenic risk scores five-fold greater than the mean for the general population for coronary artery disease(143). Highlighting individuals with such elevated risks for common, treatable diseases could have great benefits for health outcomes and allocation of resources for public health bodies. Highlighting individuals with elevated risk for diseases can allow investigators to recruit individuals with such profiles into studies and drug trials before the onset of symptoms. An additional benefit to developing PRS' is that they can also allow examination of shared genetic architecture between traits of interest(144). A shared genetic architecture (or missing genetic architecture) between traits that present similarly clinically can provide a better understanding of similarities at a deeper level and inform decisions on further studies and potential drug targets for diseases.

Highly publicised research like that mentioned above has led to calls from policy makers and healthcare professionals to integrate PRS methodologies into the wider clinical setting. This has led to some concerns raised as to whether the technology is fit for purpose for a full roll out in health services(145). Some studies have highlighted the limitations of applying PRS' developed on studies with only European or East Asian participants to individuals with greatly differentiated ancestral backgrounds(146). Often predicted PRS for individuals from these backgrounds (particularly people with African ancestry) have led to completely counter intuitive predictions for PRS in these traits. In some cases (e.g. heart disease) the PRS developed on individuals with European ancestry showed that individuals with African descent having lower predicted PRS for this trait despite heart disease being more common in Afro-Caribbean communities than individuals with European ancestries(147). Some novel variants will be absent in the primary dataset that are present in individuals with African ancestries (and vice versa) due to the substantial split time since common ancestors of people with European ancestries and those of African ancestries. Factors more likely to impact the utility of these

approaches are divergent allele frequencies for many common variants in differing ancestral populations due to differential ancestral histories and episodes of selection(139). Additionally marked differences in linkage disequilibrium will vary greatly. Any linkage disequilibrium 'pruning' process to filter out non-independent loci will provide different output for distinct populations. This is apparent between European and African populations, with African populations being shown to have much shorter linkage disequilibrium blocks than Europeans(148). There are also issues regarding ascertainment bias in relation to the genotyping technologies that have been utilised to date. Many of the conventional genotyping platforms have been developed through ascertaining variants based on results from sequencing data primarily from individuals of European ancestry(149). The relative lack of utility when applying PRS scores developed in European populations to the population as a whole has led to outcry that such a technology if fully rolled out in national healthcare problems would create a two-tier service based on race and ancestral history of patients. These limitations will be greatly reduced in the future with increases in recruitment of participants with other ancestral backgrounds and a movement to WGS data to develop polygenic risk scores(150).

Applying a PRS approach to PD's including kuru has limitations compared to common diseases that have been the overwhelming focus of publications using this technology to date. It is unlikely that any polygenic risk score for PD would provide any benefit in regards to screening and diagnostics. The majority of PD's are rare in nature (the most common sCJD has a lifetime risk \sim one person in 5,000), meaning a threefold increase in an individual as a result of a PRS screen is not likely to result in any change in clinical prognosis or behaviour change for these individuals. Additionally, as of present there are no preventative treatments for PD so diagnosis and screening will not currently result in any prevention or treatment. Findings could result in earlier counselling and preparation for patients and families. Kuru in particular, an acquired PD due to a particular and unique set of cultural and environmental practices is not a current public health issue for the communities that it impacted as the epidemic has now come to an end. Development of a PRS particularly for kuru will also be

impossible due to the lack of samples available to generate a robust PRS due to the unique and restricted nature of the epidemic. A best use of these techniques to kuru and other PD's will be to demonstrate any shared genetic architecture between PD's, and between PD's and other neurodegenerative diseases.

1.7 Aims of the project

Using genotyped data from 943 individuals from 20 different EHPNG linguistic groups, the overall aim of this thesis was to understand further the kuru epidemic using the latest computational genetic techniques. It was understood that in order to do this the population genetics of the region would have to be understood and extensive effort was made to fully explore this in Chapter 2. Previous kuru research has produced a wealth of hypotheses to explore regarding the impact of the epidemic on affected communities and attempts were made to examine these in Chapter 3. Attempts to find single variants that played a role in kuru are explored in chapter 4 and the role of polygenic effects in contributing to the genetic architecture of kuru is explored in chapter 5. Specific research objectives and hypotheses are outlined at the beginning of each chapter.

1.8 References

1. Ironside JW, Ritchie DL, Head MW. PD's. Handbook of clinical neurology. 2017;145:393-403.
2. Soto C. Prion hypothesis: the end of the controversy? Trends in biochemical sciences. 2011;36(3):151-8.
3. Liberski PP, Gajos A, Sikorska B, Lindenbaum S. Kuru, the First Human PD. Viruses. 2019;11(3).
4. Mackenzie G, Will R. Creutzfeldt-Jakob disease: recent developments. F1000Research. 2017;6:2053-.
5. Unit TNCRS. CREUTZFELDT-JAKOB DISEASE SURVEILLANCE IN THE UK. 2018.
6. Pocchiari M, Puopolo M, Croes EA, Budka H, Gelpi E, Collins S, et al. Predictors of survival in sporadic Creutzfeldt-Jakob disease and other human transmissible spongiform encephalopathies. Brain : a journal of neurology. 2004;127(Pt 10):2348-59.
7. Lo RY, Shyu WC, Li H. Long-duration sCJD with PRNP codon 129 methionine homozygosity and cerebral cortical plaques. Neurology. 2006;66(12):1944-5.
8. Palmer MS, Dryden AJ, Hughes JT, Collinge J. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. Nature. 1991;352(6333):340-2.
9. Van Everbroeck B, Michotte A, Sciot R, Godfraind C, Deprez M, Quoilin S, et al. Increased incidence of sporadic Creutzfeldt-Jakob disease in the age groups between 70 and 90 years in Belgium. European journal of epidemiology. 2006;21(6):443-7.
10. Mok T, Jaunmuktane Z, Joiner S, Campbell T, Morgan C, Wakerley B, et al. Variant Creutzfeldt-Jakob Disease in a Patient with Heterozygosity at PRNP Codon 129. New England Journal of Medicine. 2017;376(3):292-4.
11. Collinge J. PD's of humans and animals: their causes and molecular basis. Annual review of neuroscience. 2001;24:519-50.
12. Mead S, Poulter M, Beck J, Webb TE, Campbell TA, Linehan JM, et al. Inherited PD with six octapeptide repeat insertional mutation--molecular analysis of phenotypic heterogeneity. Brain : a journal of neurology. 2006;129(Pt 9):2297-317.
13. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying PD penetrance using large population control cohorts. Science translational medicine. 2016;8(322):322ra9.
14. Prusiner SB. Inherited PD's. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(11):4611-4.
15. Soto C, Satani N. The intricate mechanisms of neurodegeneration in PD's. Trends Mol Med. 2011;17(1):14-24.
16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285.
17. Minikel EV, Vallabh SM, Orseth MC, Brandel J-P, Haik S, Laplanche J-L, et al. Age of onset in genetic PD and the design of preventive clinical trials. bioRxiv. 2018:401406.
18. Rudge P, Jaunmuktane Z, Adlard P, Bjurstrom N, Caine D, Lowe J, et al. Iatrogenic CJD due to pituitary-derived growth hormone with genetically determined incubation times of up to 40 years. Brain : a journal of neurology. 2015;138(Pt 11):3386-99.
19. Jaunmuktane Z, Mead S, Ellis M, Wadsworth JDF, Nicoll AJ, Kenny J, et al. Evidence for human transmission of amyloid- β pathology and cerebral amyloid angiopathy. Nature. 2015;525:247.
20. Purro SA, Farrow MA, Linehan J, Nazari T, Thomas DX, Chen Z, et al. Transmission of amyloid- β protein pathology from cadaveric pituitary growth hormone. Nature. 2018;564(7736):415-9.
21. Heath CA, Barker RA, Esmonde TFG, Harvey P, Roberts R, Trend P, et al. Dura mater-associated Creutzfeldt-Jakob disease: experience from surveillance in the UK. Journal of neurology, neurosurgery, and psychiatry. 2006;77(7):880-2.

22. Kim HL, Do JY, Cho HJ, Jeon Y-C, Park SJ, Ma HI, et al. Dura mater graft-associated Creutzfeldt-Jakob disease: the first case in Korea. *Journal of Korean medical science*. 2011;26(11):1515-7.
23. Bonda DJ, Manjila S, Mehndiratta P, Khan F, Miller BR, Onwuzulike K, et al. Human PD's: surgical lessons learned from iatrogenic prion transmission. *Neurosurgical focus*. 2016;41(1):E10.
24. Trevitt CR, Singh PN. Variant Creutzfeldt-Jakob disease: pathology, epidemiology, and public health implications. *The American Journal of Clinical Nutrition*. 2003;78(3):651S-6S.
25. Budka H, Will RG. The end of the BSE saga: do we still need surveillance for human PD's? *Swiss medical weekly*. 2015;145:w14212.
26. Bruce M, Chree A, McConnell I, Foster J, Pearson G, Fraser H. Transmission of bovine spongiform encephalopathy and scrapie to mice: strain variation and the species barrier. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 1994;343(1306):405-11.
27. Collinge J. Variant Creutzfeldt-Jakob disease. *Lancet (London, England)*. 1999;354(9175):317-23.
28. Gerodimos R. The UK BSE crisis as a failure of government 2004. 911-29 p.
29. Kaski D, Mead S, Hyare H, Cooper S, Jampana R, Overell J, et al. Variant CJD in an individual heterozygous for PRNP codon 129. *Lancet (London, England)*. 2009;374(9707):2128.
30. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A Novel Protective Prion Protein Variant that Colocalizes with Kuru Exposure. *New England Journal of Medicine*. 2009;361(21):2056-65.
31. Clarke P, Ghani AC. Projections of the future course of the primary vCJD epidemic in the UK: inclusion of subclinical infection and the possibility of wider genetic susceptibility. *J R Soc Interface*. 2005;2(2):19-31.
32. Gill ON, Spencer Y, Richard-Loendt A, Kelly C, Dabaghian R, Boyes L, et al. Prevalent abnormal prion protein in human appendixes after bovine spongiform encephalopathy epizootic: large scale survey. *BMJ : British Medical Journal*. 2013;347:f5675.
33. The L. The Phillips report on BSE and vCJD. *The Lancet*. 2000;356(9241):1535.
34. Aguzzi A, Baumann F, Bremer J. The prion's elusive reason for being. *Annual review of neuroscience*. 2008;31:439-77.
35. Mehrabian M, Hildebrandt H, Schmitt-Ulms G. NCAM1 Polysialylation: The Prion Protein's Elusive Reason for Being? *ASN neuro*. 2016;8(6).
36. Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wuthrich K. NMR structure of the mouse prion protein domain PrP(121-231). *Nature*. 1996;382(6587):180-2.
37. Acevedo-Morantes CY, Wille H. The structure of human prions: from biology to structural models-considerations and pitfalls. *Viruses*. 2014;6(10):3875-92.
38. Vázquez-Fernández E, Vos MR, Afanasyev P, Cebey L, Sevillano AM, Vidal E, et al. The Structural Architecture of an Infectious Mammalian Prion Using Electron Cryomicroscopy. *PLOS Pathogens*. 2016;12(9):e1005835.
39. Lloyd SE, Mead S, Collinge J. Genetics of PD's. *Current opinion in genetics & development*. 2013;23(3):345-51.
40. Giles K, Glidden DV, Patel S, Korth C, Groth D, Lemus A, et al. Human prion strain selection in transgenic mice. *Annals of neurology*. 2010;68(2):151-61.
41. MacCannell D. Bacterial strain typing. *Clinics in laboratory medicine*. 2013;33(3):629-50.
42. Wadsworth JDF, Collinge J. Molecular pathology of human PD. *Acta Neuropathologica*. 2011;121(1):69-77.
43. Aguzzi A, Heikenwalder M, Polymenidou M. Insights into prion strains and neurotoxicity. *Nature Reviews Molecular Cell Biology*. 2007;8:552.
44. Liberski PP. Electron Microscopy of PD's. In: Liberski PP, editor. *PD's*. New York, NY: Springer New York; 2017. p. 123-43.

45. Wadsworth JDF, Joiner S, Linehan JM, Asante EA, Brandner S, Collinge J. Review. The origin of the prion agent of kuru: molecular and biological strain typing. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3747-53.
46. Brandner S, Jaunmuktane Z. PD: experimental models and reality. *Acta neuropathologica*. 2017;133(2):197-222.
47. Collinge J, Alpers MP. Reminiscences and reflections on kuru, personal and scientific. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3613.
48. Whitfield JT, Pako WH, Collinge J, Alpers MP. Cultural factors that affected the spatial and temporal epidemiology of kuru. *Royal Society open science*. 2017;4(1):160789.
49. Alpers MP. Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3707-13.
50. Collinge J, Whitfield J, McKintosh E, Beck J, Mead S, Thomas DJ, et al. Kuru in the 21st century--an acquired human PD with very long incubation periods. *Lancet (London, England)*. 2006;367(9528):2068-74.
51. Lindenbaum S. Kuru sorcery : disease and danger in the New Guinea highlands. 2016.
52. Gajdusek DC, Zigas V. Kuru; clinical, pathological and epidemiological study of an acute progressive degenerative disease of the central nervous system among natives of the Eastern Highlands of New Guinea. *The American journal of medicine*. 1959;26(3):442-69.
53. McLean CA. Review. The neuropathology of kuru and variant Creutzfeldt-Jakob disease. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3685-7.
54. Anderson W. *The collectors of lost souls : turning Kuru scientists into whitemen*. Baltimore, Md.: Johns Hopkins University Press; 2008.
55. Anderson WH. Early perceptions of an epidemic. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3675-8.
56. Gray G. 'You are... my anthropological children': AP Elkin, Ronald Berndt and Catherine Berndt, 1940–1956. *Aboriginal History*. 2005;29:77-106.
57. Yam P. The pathological protein mad cow, chronic wasting, and other deadly PD's. 2005.
58. Gajdusek DC, Zigas V. Degenerative disease of the central nervous system in New Guinea; the endemic occurrence of kuru in the native population. *The New England journal of medicine*. 1957;257(20):974-8.
59. Lindenbaum S. Cannibalism, kuru and anthropology. *Folia neuropathologica*. 2009;47(2):138-44.
60. Hadlow WJ. Kuru likened to scrapie: the story remembered. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3644-.
61. Adams DB. Evolutionary biology and the risk of scrapie disease in sheep. *Open veterinary journal*. 2018;8(3):282-94.
62. Gajdusek DC, Gibbs CJ, Alpers M. Transmission and Passage of Experimental 'Kuru' to Chimpanzees. *Science (New York, NY)*. 1967;155(3759):212.
63. Gajdusek DC, Gibbs CJ, Alpers M. Experimental Transmission of a Kuru-like Syndrome to Chimpanzees. *Nature*. 1966;209(5025):794-6.
64. Alpers MP. Some tributes to research colleagues and other contributors to our knowledge about kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3614-7.
65. Mathews JD, Glasse R, Lindenbaum S. Kuru and cannibalism. *Lancet (London, England)*. 1968;2(7565):449-52.
66. Bolton DC, McKinley MP, Prusiner SB. Identification of a protein that purifies with the scrapie prion. *Science (New York, NY)*. 1982;218(4579):1309-11.

67. Lantos PL. From slow virus to prion: a review of transmissible spongiform encephalopathies. *Histopathology*. 1992;20(1):1-11.
68. Collinge J, Whitfield J, McKintosh E, Frosh A, Mead S, Hill AF, et al. A clinical study of kuru patients with long incubation periods at the end of the epidemic in Papua New Guinea. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3725-39.
69. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A novel protective prion protein variant that colocalizes with kuru exposure. *The New England journal of medicine*. 2009;361(21):2056-65.
70. Lloyd SE, Onwuazor ON, Beck JA, Mallinson G, Farrall M, Targonski P, et al. Identification of multiple quantitative trait loci linked to PD incubation period in mice. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(11):6279-83.
71. Asante EA, Smidak M, Grimshaw A, Houghton R, Tomlinson A, Jeelani A, et al. A naturally occurring variant of the human prion protein completely prevents PD. *Nature*. 2015;522(7557):478-81.
72. Mead S, Uphill J, Beck J, Poulter M, Campbell T, Lowe J, et al. Genome-wide association study in multiple human PD's suggests genetic risk factors additional to PRNP. *Human molecular genetics*. 2012;21(8):1897-906.
73. Ross M. Pronouns as a preliminary diagnostic for grouping Papuan languages 2005. 15-66 p.
74. Wurm SA. *New Guinea area languages and language study*. 1, 1. Canberra: Dep. of Linguistics, Research School of Pacific Studies, Australian National Univ.; 1977.
75. Feil DK. *The evolution of highland Papua New Guinea societies*. Cambridge: Cambridge University Press; 2010.
76. L. Langness L. *Sexual Antagonism in the New Guinea Highlands: A Bena Bena Example* 1967. 161-77 p.
77. Watson JB. *OTHER: New Guinea: The Last Unknown*. Gavin Souter. *American Anthropologist*. 1967;69(6):789-90.
78. Summerhayes G, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, et al. Human Adaptation and Plant Use in Highland New Guinea 49,000 to 44,000 Years Ago 2010. 78-81 p.
79. Denham T, Fullagar R, Head L. Plant exploitation on Sahul: From colonisation to the emergence of regional specialisation during the Holocene 2009. 29-40 p.
80. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538:207.
81. Greenhill SJ. *TransNewGuinea.org: An Online Database of New Guinea Languages*. *PloS one*. 2015;10(10):e0141563-e.
82. Diamond J, Ordunio D, OverDrive I. *Guns, Germs, and Steel*. 2011.
83. Pawley A. *Papuan pasts cultural, linguistic and biological histories of Papuan-speaking peoples*. Canberra: Pacific Linguistics; 2005.
84. Bergstrom A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science (New York, NY)*. 2017;357(6356):1160-3.
85. Greig K, Gosling A, Collins CJ, Boocock J, McDonald K, Addison DJ, et al. Complex history of dog (*Canis familiaris*) origins and translocations in the Pacific revealed by ancient mitogenomes. *Scientific reports*. 2018;8(1):9130.
86. *Travelling companions*. *Nature Plants*. 2016;2:16177.
87. Kirch PV. *Lapita and Its Aftermath: The Austronesian Settlement of Oceania*. *Transactions of the American Philosophical Society*. 1996;86(5):57-70.
88. Kayser M, van Oven M, Choi Y, Brauer S, Mona S, Suarkia D, et al. The Impact of the Austronesian Expansion: Evidence from mtDNA and Y Chromosome Diversity in the Admiralty Islands of Melanesia. *Molecular biology and evolution*. 2008;25(7):1362-74.

89. Brucato N, Fernandes V, Kusuma P, Cerny V, Mulligan CJ, Soares P, et al. Evidence of Austronesian Genetic Lineages in East Africa and South Arabia: Complex Dispersal from Madagascar and Southeast Asia. *Genome biology and evolution*. 2019;11(3):748-58.
90. Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, et al. Language continuity despite population replacement in Remote Oceania. *Nature ecology & evolution*. 2018;2(4):731-40.
91. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, et al. Population Turnover in Remote Oceania Shortly after Initial Settlement. *Current Biology*. 2018;28(7):1157-65.e7.
92. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature*. 2016;538(7626):510-3.
93. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010;468:1053.
94. Du A, Alemseged Z. Temporal evidence shows *Australopithecus sediba* is unlikely to be the ancestor of *Homo*. *Science advances*. 2019;5(5):eaav9038.
95. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. *Science (New York, NY)*. 2010;328(5979):710.
96. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS genetics*. 2012;8(10):e1002947.
97. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. 2014;512:194.
98. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*. 2016;26(9):1241-7.
99. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*. 2019.
100. Gajdusek DC, Alpers M. Genetic studies in relation of kuru. I. Cultural, historical, and demographic background. *American journal of human genetics*. 1972;24:Suppl:1-38.
101. Mandeville E. Agnation, Affinity and Migration Among the Kamano of the New Guinea Highlands 1979. 105 p.
102. Lindenbaum S. First impressions of the Fore. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3648-9.
103. Sorenson ER, Claessen HJM, du Torr BM, Griffith J, Hockings P, Jablonko A, et al. Socio-Ecological Change Among the Fore of New Guinea [and Comments and Replies]. *Current Anthropology*. 1972;13(3/4):349-83.
104. B. Watson J. Society as Organized Flow: The Tairora Case 1970. 107-24 p.
105. Connolly B, Anderson R. First contact. New York, N.Y., U.S.A.: Penguin Books; 1988.
106. Mandeville E. Kamano Adoption 1981. 229 p.
107. Littlewood RA. ISOLATE PATTERNS IN THE EASTERN HIGHLANDS OF NEW GUINEA. *The Journal of the Polynesian Society*. 1966;75(1):95-106.
108. Wiesenfeld SL, Gajdusek C. Genetic structure and heterozygosity in the kuru region, eastern highlands of New Guinea. *American journal of physical anthropology*. 1976;45(2):177-90.
109. Bhatia K, Prasad ML, Barnish G, Koki G. Antigen and haplotype frequencies at three human leucocyte antigen loci (HLA-A, -B, -C) in the Pawaia of Papua New Guinea. *American journal of physical anthropology*. 1988;75(3):329-40.
110. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
111. Senn N, Riddell M, Omena M, Siba P, Reeder JC, Clements CJ, et al. Measles in Papua New Guinea: an age-specific serological survey. *Vaccine*. 2010;28(7):1819-23.
112. Watson JB. THE SIGNIFICANCE OF A RECENT ECOLOGICAL CHANGE IN THE CENTRAL HIGHLANDS OF NEW GUINEA. *The Journal of the Polynesian Society*. 1965;74(4):438-50.

113. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*. 2015;526(7571):112-7.
114. Andrews TM, Price RM, Mead LS, McElhinny TL, Thanukos A, Perez KE, et al. Biology undergraduates' misconceptions about genetic drift. *CBE Life Sci Educ*. 2012;11(3):248-59.
115. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. *American journal of human genetics*. 2004;74(6):1111-20.
116. Luzzatto L. Genetic factors in malaria. *Bull World Health Organ*. 1974;50(3-4):195-202.
117. Slade RW, McCallum HI. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*. 1992;132(3):861-4.
118. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic adaptive introgression in humans. *Nature reviews Genetics*. 2015;16(6):359-71.
119. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*. 2013;193(3):929-41.
120. Hersch EI, Phillips PC. Power and potential bias in field studies of natural selection. *Evolution; international journal of organic evolution*. 2004;58(3):479-85.
121. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays : news and reviews in molecular, cellular and developmental biology*. 2013;35(9):780-6.
122. Pedersen CT, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, et al. The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics*. 2017;205(2):787-801.
123. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research*. 2011;21(6):821-9.
124. Stearns SC, Byars SG, Govindaraju DR, Ewbank D. Measuring selection in contemporary human populations. *Nature Reviews Genetics*. 2010;11:611.
125. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. *Briefings in functional genomics*. 2014;13(5):371-7.
126. Foll M, Gaggiotti Oscar E, Daub Josephine T, Vatsiou A, Excoffier L. Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. *The American Journal of Human Genetics*. 2014;95(4):394-407.
127. R. Vahdati A, Wagner A. Parallel or convergent evolution in human population genomic data revealed by genotype networks. *BMC Evolutionary Biology*. 2016;16(1):154.
128. Casanova J-L. Adaptive immunity by convergent evolution. *Nature Reviews Immunology*. 2018;18:294.
129. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*. 2017;19:110.
130. Bearn AG, Miller ED. Archibald Garrod and the development of the concept of inborn errors of metabolism. *Bulletin of the history of medicine*. 1979;53(3):315-28.
131. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-86.
132. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*. 2018;19(9):581-90.
133. Illingworth DR, Stein EA, Mitchel YB, Dujovne CA, Frost PH, Knopp RH, et al. Comparative effects of lovastatin and niacin in primary hypercholesterolemia. A prospective trial. *Archives of internal medicine*. 1994;154(14):1586-95.

134. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013;45(11):1274-83.
135. Plotnikov D, Guggenheim JA. Mendelian randomisation and the goal of inferring causation from observational studies in the vision sciences. *Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*. 2019;39(1):11-25.
136. Biedrzycki RJ, Sier AE, Liu D, Dreikorn EN, Weeks DE. Spinning convincing stories for both true and false association signals. *Genet Epidemiol*. 2019.
137. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews Genetics*. 2011;13(1):36-46.
138. Yan G, Guo T, Xiao S, Zhang F, Xin W, Huang T, et al. Imputation-Based Whole-Genome Sequence Association Study Reveals Constant and Novel Loci for Hematological Traits in a Large-Scale Swine F2 Resource Population. *Frontiers in genetics*. 2018;9:401.
139. Kuchenbaecker K, Telkar N, Reiker T, Walters RG, Lin K, Eriksson A, et al. The transferability of lipid loci across African, Asian and European cohorts. *Nature communications*. 2019;10(1):4330.
140. Hizume M, Kobayashi A, Teruya K, Ohashi H, Ironside JW, Mohri S, et al. Human prion protein (PrP) 219K is converted to PrP^{Sc} but shows heterozygous inhibition in variant Creutzfeldt-Jakob disease infection. *The Journal of biological chemistry*. 2009;284(6):3603-9.
141. Misra A, Chakrabarti SS, Gambhir IS. New genetic players in late-onset Alzheimer's disease: Findings of genome-wide association studies. *The Indian journal of medical research*. 2018;148(2):135-44.
142. Karimi-Moghadam A, Charsouei S, Bell B, Jabalameli MR. Parkinson Disease from Mendelian Forms to Genetic Susceptibility: New Molecular Insights into the Neurodegeneration Process. *Cellular and molecular neurobiology*. 2018;38(6):1153-78.
143. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*. 2018;50(9):1219-24.
144. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*. 2015;31(9):1466-8.
145. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? *Genome medicine*. 2018;10(1):100.
146. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric genetics*. 2018;28(5):85-9.
147. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? *Genome medicine*. 2018;10(1):100-.
148. Vilhjalmsdottir BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American journal of human genetics*. 2015;97(4):576-92.
149. Mackay TFC. The Genetic Architecture of Quantitative Traits. *Annual review of genetics*. 2001;35(1):303-39.
150. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics*. 2017;13(6):e1006836.

Chapter 2 - Population genetic analysis of EHPNG

2.1 Introduction

In this chapter work is presented that was undertaken with the aim of characterising the population structure of EHPNG. This was motivated through a combination of the great interest to further understanding of these populations from a genetic perspective and importantly to improve experimental design and permit accurate inferences in relation to genetic epidemiology analysis of the kuru epidemic. Hypotheses tested in this chapter were;

- Linguistic group membership is the best descriptor of population genetic structure of EHPNG from a population genetic perspective.
- Individuals exist within linguistic groups that have distinct ancestry, providing evidence of migration of individuals between linguistic groups as documented in other fields of study.
- Genetic admixture events in the past between populations in EHPNG are likely to have left a genetic signature on modern day inhabitants of the region.

Further information is provided below to contextualise the work undertaken and the specific hypotheses tested.

EHPNG represent one of the final regions in the world to have come under colonial administrative rule(54). Inhabitants of the region predominantly lived in isolation from the rest of the world with the only interaction coming through lengthy, indirect trade networks that navigated the complex terrain between the coastal and highland regions(105). Western observers held the belief that the highlands of New Guinea were likely uninhabited until the 20th century(54). Exploration by Christian missionaries and gold prospectors in the first decades of the 20th century however revealed the highlands to contain densely populated valleys that were home to close to approximately million people(82). These groups were marked by their complex cultural and trade systems(59), cosmology(151), linguistic diversity(73, 74) and independent adoption of agriculture(75). The demographic and genetic properties of such populations are of great interest to scholars in

understanding the processes that could maintain a region in relative isolation from the outside world yet so rich in cultural, geographic and linguistic diversity (See section 1.3.1 for more information on the history of PNG).

Previous genetic studies have shown that the PNG Highlands display high levels of genetic differentiation between the various groups that populate the region(84). The number of distinct languages spoken in the highlands as a whole is believed to be in the hundreds (see section 1.3.1 for more on the PNG Highland linguistics). In EHPNG, an area approximately half the size of Wales in the United Kingdom (EHPNG 11,157km² and Wales 20,735km² respectively), 37 languages are spoken(108). The PNG Highlands was also the epicentre of an independent agricultural revolution based on taro cultivation. Further analysis and understanding of processes that have contributed to population structure and genetic diversity in the PNG Highlands will enhance comparisons between EHPNG and other more studied centres of agricultural innovation.

At the time of colonial rule, groups in EHPNG were administered according to linguistic categorisations made by the colonial Australian authorities(100). When individuals were asked by administrators and investigators about their perception of the geographical limits of their communities they were often not aware of the extent of the region where their own language was spoken(51). There were no written languages at the time of contact or records kept(51). Knowledge of past events was usually maintained through oral traditions(51). Dates were not kept via a calendar system and points of time in the past were usually made by making reference to well-known past events including major conflicts between groups, volcanic eruptions(152) or floods(54).

Communities in EHPNG have practices of exogamous marriage where women would move to the village of residence of their new husbands(101). There was variation regarding whether these exogamous marriage practices would cross linguistic speaking borders, with marital exchange occurring between Fore communities and the Keiagana(100), but never between the Fore and Anga(49). Goods would also flow across the EHPNG region over linguistic boundaries as well as

different belief systems and ideas(100). Consanguineous marriage was common in the region historically(54).

Prior to European contact, communities were assembled into hamlets which would form part of larger clans(65, 101, 104). Members of clans would often share affiliation to one another through a shared patrilineal ancestor(101). These patrilineal lineage groups were not hierarchical, the time depth of the shared origin was shallow and of groups were not rigid, with flexibility and the possibility to move between groups(101, 104). It is not clear whether description of groups at the hamlet, village, clan, or linguistic group level best represent stable subpopulations of individuals that optimally characterise genetic diversity in the region.

In addition to how groups are organised socially, politically, and linguistically, geographical features including distance and topography will likely play important roles in the patterns and distribution of genetic diversity and diversification on display. EHPNG is notable for the widespread steep terrain reaching peaks as high as Mt Michael at 3,750 metres above sea level, to valleys below 1,000 metres above sea level. Patterns of residence due to geography and ecology may also have been shaped by the presence of large river systems, which at times may be challenging to cross. The need to grow crops in different conditions and malarial conditions at lower elevations may also have stimulated movements of people historically.

Genetic admixture refers to the presence of DNA in a population derived from two distinct non-interbreeding populations. Episodes of admixture have had large effects on shaping human population genetic history and population structure(153-155). Genetic methodologies are able to detect and accurately date historically known episodes of admixture and have also inferred admixture episodes that were not present in the historical record(156). Most of these inferred admixture events have occurred at a continental scale between populations separated previously on large spatial scales and were also genetically distinct. Coastal regions of PNG harbour admixed populations that were formed as part of the Austronesian expansion with many speaking

Austronesian languages(84). There is no evidence of this migration impacting the ancestry profiles of inhabitants of the Highlands of PNG despite the presence of goods such as tobacco(157) and sweet potato(158) which have external origins. Undoubtedly there have also been episodes of admixture within the Highlands of PNG and populations within EHPNG are likely to have been formed through admixture events. In EHPNG, the degree of diversification between groups within such small geographical areas raises the possibility of testing for undocumented past admixture events.

Until now, a region showing such high genetic differentiation like EHPNG has not been studied using modern genome-wide methods in densely sampled datasets. This study has permitted investigation of processes such as local patterns of historical migration between groups, the effect of geographical features and local cultural differences, all of which have played a role in shaping the genetic differentiation observed today.

The contribution of migration to genetic diversity in EHPNG is still poorly understood. Observers noted the presence of features that aided possible migration in the region and others that have impeded it. Stimulating factors include donation of land to new arrivals(104), membership of clans based on observance of shallow ancestral lineages , and the practices of non kin adoption, refugees from warfare(100, 103), exogamous marriage and an age-mate system that allowed non-kin to move between villages in the region(59, 106). Impeding factors are the degree of difficulty in traversing the landscape, both in terms of the extreme geography and the carving of the region into hundreds of warring territories that would be hazardous to pass through. Kuru spread through southern linguistic groups in EHPNG as a result of people moving residence and taking their customs to new settlements. Improved understanding of migration may permit better understanding of and support for work from other fields about the movement of kuru in the region during the epidemic

Using a dataset of 943 individuals from 21 different linguistic speaking groups in EHPNG the genetic structure of EHPNG has been analysed. Goals include improving understanding of what processes have shaped observed genetic differentiation including geographical distance, linguistic and cultural

differences, processes of migration and degrees of group separation in the region. In other populations that have been sampled this densely, a lack of genetic differentiation between groups has made it problematic to make inferences on these processes based on genetic data and analysis. The amount of differentiation observed in EHPNG and PNG Highlands more generally, present opportunities to provide a holistic understanding of the population structure in a region that remained largely isolated from external influences until the 20th century. How groups and individuals organised themselves in such a setting can provide insights into to how historically communities organised themselves prior to the homogenizing influences of technology, empire and population mixture.

Importantly, EHPNG was the epicentre of the kuru epidemic that reached its peak in the mid-20th century(48). A detailed understanding of the population structure in the region in both kuru affected and non-affected groups will provide the basis for improved in depth analysis of the genetic response of these groups to kuru exposure which is a major aim of this thesis.

2.2 Materials and methods

2.2.1 Genetic dataset

2.2.1.1 Primary dataset

Laboratory studies were approved by the Papua New Guinea Medical Research Advisory Committee, and by the local research ethics committee of the Institute of Neurology.

Full participation of the communities involved was established and maintained through discussions with village leaders, communities, families and individuals. The field studies followed the principles and practice of the Papua New Guinea Institute of Medical Research (PNGIMR) which included individual oral consent from all participants before any samples were obtained.

The ongoing collaboration between the PNGIMR and the Medical Research Council Prion Unit (MRCPU) was initiated in 1996. The collaboration began to focus on reinvigorating kuru surveillance

that had originally begun in 1957, and additionally to collect biological samples from kuru cases, individuals exposed to kuru, individuals born after 1960 not exposed to kuru, and samples from individuals from EHPNG populations who had no kuru exposure. Information was obtained about the individual's village of residence, date of birth, and sex. After initial processing these samples were transported to the MRCPU in the United Kingdom. Blood samples from 4,456 individuals were subsequently taken from individuals from these various cohorts over the next ten years. Of these 1,594 have been genotyped and different types of genetic analyses have been undertaken. Samples were selected for genotyping with an aim to prevent the accumulation of unnecessary first degree related individuals in the genetic dataset. In addition to the samples obtained through the PNGIMR MRCPU collaboration, genotype data was obtained from serum samples previously used in investigations by the National Institute of Health in Bethesda Maryland in the 1970's. This collection comprised individuals who died of kuru and also individuals from the Pawaian linguistic group.

2.2.1.2 Batch bias

Samples were genotyped at various phases of the collaboration during and after the primary sample collection. The first 557 samples were genotyped on the Illumina 670 genotyping platform. The following 1,037 samples were genotyped on the Illumina Omni Express genotyping platform four years apart (746 samples in 2012 and 291 samples in 2016). 83 individuals were genotyped on both the Illumina 670 and Illumina Omni Express platforms to permit downstream analysis of possible biases introduced by the different genotyping platforms. The arrays comprise 748,000 and 678,000 variants respectively. These platforms were designed to detect polymorphic variants in continental populations, making ascertainment of variants unique to these isolated populations a possible concern. To mitigate this problem, haplotype based approaches that are less prone to errors due to ascertainment were used.

2.2.1.3 Merger with publicly available data

The primary dataset was merged with publicly available data in 3 separate merges. The primary Illumina 670 and Illumina Omni data was merged with phase 3 data from the 1000 Genomes Project(159) and data from a previous publication with 380 participants from the PNG and neighbouring islands(84). After all quality control procedures undertaken in PLINK 1.9(160) (allele frequency > 0.01 , missingness < 0.01 , HWE > 0.00001 , identification of related individuals please see below 2.2.3) there remained an intersection of 122,663 variants and 1,462 individuals, this dataset is referred to as the '*PNG merge*' subsequently. A second merge was performed only using the 1,037 Illumina Omni Express samples from the primary dataset; these were merged with data from various Oceanic and South Asian samples from publicly available data including the 1000 Genomes Project again (161). This merge, after the same quality control procedures as above, resulted in 149,113 variants and 1,202 individuals. Subsequently this dataset is referred to as the '*Oceania merge*' (See appendix 1 for breakdown of samples in *Oceania* and *PNG merge*). The data used in analyses in this chapter can be found in Appendix 1.

2.2.2 Phasing of data

Autosomal chromosomes were jointly phased for all individuals in each of the merges using SHAPEIT with default parameters and the linkage disequilibrium-based genetic map build 37(162).

2.2.3 Identification of related individuals and estimates of genetic diversity within groups

Pairwise estimates of relatedness were performed using the PLINK1.9 –genome tool. This gives pairwise estimates of 'PIHAT' which can be seen as analogous to a pairwise Identity by State (IBS) estimate. For every pairwise estimate over 0.1875 (roughly analogous to a first cousin relationship) the individual with the inferior genotyping rate was removed. The profile of these PIHAT estimates were also used to further understanding of the within group genetic diversity displayed by each group and fineStructure (FS)(163)cluster (see below).

2.2.4 Analysis of EHPNG population structure

A primary aim of the study was to understand which factors were most important in explaining patterns of observed genetic diversity. In order to gauge the contributions of linguistic group membership, village of residence and geographical parameters to this diversity multiple analyses were performed. Additionally, analysis was performed to test if groups in the region had distinct ancestral histories by comparing groups in EHPNG to other populations outside of EHPNG that were available in the merged genetic datasets.

2.2.4.1 Principal Component Analysis (PCA)

To visualize genetic distances among sampled individuals, PCA was performed on each dataset using the PLINK 1.9 command “—pca”, when using all samples or after removing samples from some groups to reduce the impact of sampling bias from groups with a large number of samples. Nine individuals were taken from each linguistic group to ensure equality in sample sizes. Prior to PCA, data were pruned with PLINK 1.9 using the function “indep-pairwise 50 5 0.5”. PCA analysis was performed on a subset of individuals from the *PNG merge* dataset from 20 EHPNG linguistic groups to focus on population structure within EHPNG. In order to remove the impact of sampling bias 10 individuals were selected from each linguistic group and a maximum pairwise relatedness of 0.1875 inferred by the plink genome tool ‘PIHAT’ was used (See appendix 3). Analysis of what each linguistic group contributed to the variation in each eigenvector was performed, the squared value of individuals eigenvalues were compiled for each linguistic group for each eigenvector and tabulated.

An analysis of individuals from 20 villages from three kuru-affected linguistic groups (Gimi, North Fore, South Fore) with nine individuals from each village was performed to see if village residence could visually explain genetic variation at this reduced geographical scale. Additional PCA analysis was performed on the *Oceania dataset* to see how EHPNG is placed in a regional context and if all groups clustered together which would support the hypothesis of shared common origin of EHPNG groups compared to outside groups. Separately, PCA was applied to check for batch effects between

the various stages of genotyping samples and between the different genotyping platforms used during the duration of the project. Samples genotyped on the Omni dataset in 2012 were compared to those in 2016 and then all Omni individuals with individuals genotyped on both Illumina Omni and Illumina 670 platforms.

2.2.4.2 Admixture analysis

Supervised admixture analysis was performed using the Admixture tool(164). Number of clusters per analysis was predefined for K2-10. Again to control for sampling bias ten unrelated individuals were used per each of the linguistic groups. An analysis using the *Oceania merge* of EHPNG individuals to understand the distinctiveness of the genetic structure of EHPNG populations, placed on a regional scale. Admixture analysis was also performed on samples within EHPNG to assess if any structure could be defined using this method (See appendix 6 for admixture plots).

2.2.4.3 ChromoPainter and FS analysis of EHPNG linguistic groups

ChromoPainter (CP) analysis was performed on each of the dataset mergers to identify which groups share most recent ancestry with our EHPNG individuals. In brief, CP is a 'painting' technique that compares haplotype patterns within a target chromosome to those within a set of sampled "donor" chromosomes(27). In a genetic region, if a target's haplotype patterns are more similar to a particular donor relative to the other donors, this suggests the target shares a more recent ancestor with that donor relative to the others for that genetic region. CP provides a 'painting profile' for each target individual reflecting the amount of genome-wide DNA for which the individual is inferred to share a most recent ancestor with each donor individual.

CP analysis was performed on 1,462 individuals from the *PNG merge*. Following a previous publication (28), two CP model parameters were first estimated, the switch ("-n") and emission ("-M") rates, using 10 Expectation-Maximisation (E-M) iterations ("-i 10 -in -iM") when applying CP to 1,462 target individuals and 22 chromosomes. These estimated values were fixed and separately painted each of the 1,462 individuals as a target when using the other 1,461 individuals as donors.

Finestructure (FS) was used to group individuals into genetically homogenous clusters based on the CP output. Importantly, these groupings are free from any bias due to a priori classifications of individuals, e.g. based on linguistic classifications (27). Following the recommended FS approach described by a previous publication(27), a normalisation parameter 'c'(0.70) was inferred and performed two million iterations of Markov-Chain-Monte-Carlo (MCMC) , sampling an inferred clustering every 10,000 iterations after a burn in of one million iterations. Starting from the single MCMC sampled clustering with highest posterior probability, 100,000 additional hill-climbing steps in FS were then performed to find a nearby state with even higher posterior probability. The results of this hill-climbing approach grouped these 1,462 individuals into 105 clusters.

Using a greedy approach, the FS tree that produced 105 initial clusters was collapsed to a lower number of clusters by ascending levels in this phylogenetic tree. Many of the tree splits at lower points in the tree have very small likelihood separations in the clustering allocation. Merge of some of these clusters can allow a more comprehensive assessment of population structure without loss of much information. The approach to collapsing the initial FS data was a similar one taken by a study of population structure in the British Isles(165). When the FS tree is collapsed to 25 clusters there is clear recapitulation to linguistic group membership (Appendix 11).

A similar process was performed with the Oceania merge where all 1,208 individuals from 53 different populations acted as both recipients and donors in CP analysis. FS analysis was performed on the output to again check grouping of individuals into genetically homogenous clusters free from bias. To assess the effect of village of residence as a suitable descriptive category for describing population structure in EHPNG an alternative analysis was performed using individuals from 32 villages from the North Fore, South Fore and Gimi linguistic groups. For the remaining linguistic groups in the analysis only single villages were sampled.

To infer about shared common ancestry between groups in EHPNG, using the *PNG merge* all 943 EHPNG individuals were used as recipients and painted against a panel of 15 other groups from non

EHPNG PNG populations. This included highland, coastal and neighbouring island populations. The same approach was used in the *Oceania dataset* where all 664 EHPNG individuals were painted as recipients and individuals from 13 more distant populations acted as donors compared to the PNG merge analysis.

2.2.5 Influence of geography on population structure and observed genetic diversity

Pairwise geographical distances were calculated in R package ‘sp’ (30) after latitudinal and longitudinal coordinates were supplied by PNGIMR fieldworkers who collected the data. For each dataset merge, Pearson’s correlation coefficients were calculated between genetic and geographic distance across all EHPNG individuals.

In addition, following reference 31, genetic distance measure Total Variation Distance (TVD) was assumed to be related to geographic distance via the following exponential function:

$$F(D_{ij}) = \frac{1}{\alpha_0} \exp(-(\alpha_1 D_{ij})^{\alpha_2})$$

α , i and j parameters were inferred using grid search for the parameters that gave the smallest mean-squared error between TVD and geographic distance in the entire dataset. This modified distance measure is referred to as the ‘geogenetic distance’, which was then compared to TVD to visualize patterns of genetic differentiation that deviate from expectations due to geographic distance.

2.2.6 Analysis of migration processes using TVD and F_{ST} between individuals

Differences between the CP painting profiles of individuals were calculated using the Total Variation Distance (TVD) measure as described by Leslie et al (165).

$$TVD_{XY} = 0.5 \sum_{k=1}^K (f_k^X - f_k^Y).$$

where f_k^X and f_k^Y are the average genome-wide proportion of DNA that individuals from the recipient groups X and Y, respectively, match to donor group $k \in [1, \dots, K]$ as inferred by CP. These distances were also calculated for each of the linguistic groups, population labels, villages and FS clusters produced in the CP analyses. This was done by aggregating the copying profile of the constituent members of the linguistic groups, population labels and FS clusters. K can change depending on how individual copying profiles are formed. They can be based on population labels where an individual's CP copying profile will be based on the amount of material copied from donors from K different populations or by donors grouped into K by FS cluster membership.

The TVD measure was used to detect individuals that could be considered recent migrants to their village of residence or the descendants of recent migrants to their village of residence. The approach is based on the conclusion from CP analysis (see results) that linguistic group membership is the best overall descriptor of population structure and ancestral history. The hypothesis and approach was based on the assumption that any individual who has an ancestry profile that is markedly different from other individuals from the same linguistic group, measured by TVD, can be considered to represent an individual of different original precedence or having parents of different precedence. An aggregated painting profile for each linguistic group or copying vector was created by averaging the copying vector of all individuals in each linguistic group. Each individual's CP painting profile was then compared to each of the aggregated linguistic group painting profiles (Appendix 5 for samples in analysis). Long-range migrants were classified as individuals who had a copying profile most similar to a different linguistic group as measured by TVD and this group was geographically separated by at least one other intervening linguistic group region. An exception was made for individuals from groups neighbouring the Anga and Pawaian who had profiles more similar to these groups given the divergence of these groups from others observed in prior analysis of the genetic structure of populations in the region.

2.2.7 fastGLOBETROTTER

In order to test hypotheses of discrete episodes of population admixture, fastGLOBETROTTER(14) was used. fastGLOBETROTTER assumes a 'pulse' model whereby admixture occurs instantaneously for each admixture event, followed by the random mating of individuals within the admixed population from the time of admixture until present-day. fastGLOBETROTTER uses the painting profiles from CP analysis described above as input. Two paintings are performed for each analysis with all groups being painted as donors and recipients to understand the genetic makeup of admixing sources. The second painting involves the hypothesised target population of admixture acting as a recipient with potential admixing sources acting as donors. Inference of CP model parameters '-n'- and '-M' were performed for each target population in analyses as was done in previous CP analysis of EHPNG as a whole mentioned above.

The first test was whether highlands populations of PNG had experienced discrete episodes of admixture in the past that affected all highland communities as a whole leaving a genetic signature in the genomes of present day inhabitants of the Highlands of PNG. The most likely source of discrete admixture to have occurred in recent millennia would have been a result of the Austronesian expansion that has impacted near and remote Oceania over the past three millennia(7, 14, 22, 33-35). Chromosome painting was performed on the *Oceania dataset*, with a first painting used to describe the admixing sources using all groups as recipients and donors comprising all groups in the *Oceania dataset*. The second painting used to estimate dates of admixture pulses in the target highland population used highland population individuals as recipients and PNG coastal Austronesian groups and other Austronesian groups in East Asia, Near and Remote Oceania acting as donors.

The second test was whether the South Fore linguistic group could be described as descending from a mixture of genetically differentiated sources who intermixed (i.e. admixed) over one or more narrow time periods, and – if so – to date when this mixture occurred. The substantial sampling of

individuals from the South Fore population allowed examination of possible admixture episodes in this individual linguistic group. This analysis was aiming to discover pulses of admixture *within* EHPNG rather than being the result of an arrival of an admixing source previously external to the highlands of PNG. Individuals who were classified as migrants in the migrant analysis described above were removed from the analysis to allow insights into deeper-in-time processes that affected modern group composition. If individuals who recently arrived due to processes of migration in the most recent generations were included it would reduce power to detect pulses of admixture that occurred in the South Fore population further back in time. Again 2 CP analyses were performed, but for this experiment on the *PNG merge* dataset. The first with all EHPNG groups acting as donors and recipients and the second painting with the target South Fore population as a recipient group and other EHPNG groups acting as donors. Again ‘-n’ and ‘-M’ model parameters were estimated.

2.3 Results

2.3.1 F_{ST} Measurements

F_{ST} differences between the 22 linguistic groups within EHPNG showed greater on average differences (0.021, SD 0.012) between groups of individuals from nations in Europe as greatly separated as Finland and Spain (0.012)(84) in an area half the size of Wales (11,100 km²). Average F_{ST} between the Anga and all other groups (0.042, SD 0.013) was high and consistent between other groups (SD 0.0045) suggesting a completely distinct ancestral history for this linguistic group. A similar pattern appears for the Pawaian (mean F_{ST} 0.036, SD 0.0058) and Tairora (mean 0.029, SD 0.0073) linguistic groups as well, again suggesting relatively distinct ancestral histories for members of these linguistic groups. F_{ST} plots revealed a correlation between pairwise TVD of linguistic groups and straight-line distance (R^2 0.19). When the Anga and Pawaian groups are removed from this analysis, this correlation increases (R^2 0.21) (Appendix 9).

Analysis of F_{ST} of 20 villages in the kuru restricted region revealed differences between villages that are more difficult to resolve than differences between linguistic groups. Average F_{ST} between villages

in the analysis was 0.007 (SD = 0.0047) compared to 0.022 between linguistic groups. This is to be expected as the sampling area of the 20 villages is restricted to a subset of EHPNG region. This must be caveated with the understanding that many of the pairwise F_{ST} measurements were negative meaning no difference could be obtained. This was the case for villages in immediate proximity including villages on different sides of linguistic boundaries like Kamata in the South Fore region and Awande in the North Fore region. The F_{ST} difference between the villages of Kalu in the North Fore and Ivaki (0.020) and Kalu (0.020) in the south-western extremes of the South Fore was much larger than the F_{ST} difference between South Fore and North Fore regions as a whole (0.0046). This does reveal that many subtle differences in relatedness over smaller distances, including across linguistic boundaries will be lost when simply assessing structure at the linguistic group level.

2.3.2 Admixture Analysis

Admixture analysis of 20 EHPNG linguistic groups with K number of clusters set to 2-10 supports the relevance of linguistic group classification as a genetic descriptor. When K=2 it appears to broadly describe a split in Eastern and Western linguistic groups in EHPNG. At K=3-4 the population isolates Anga and Pawaian appear. K=5-10 shows the emergence of distinct clusters of linguistic groups reflecting their different ancestral histories. When K=10, 8 of the clusters are dominated by individual linguistic groups with the remaining 2 clusters being dominated by combinations of individuals from North-Western populations.

Admixture analysis of 20 villages in the Gimi, North Fore and South Fore linguistic groups show less structure and differentiation than the analysis based on linguistic group classification. At K=2 there is a general distinction between individuals from more northerly villages dominating one cluster and individuals from more southerly villages dominating the other cluster. With additional clusters a village defined structure does not emerge with individuals from various villages in the region being strongly associated with each cluster. Eventually at K=6 there is no discernible structure with many individuals not being associated with a particular cluster but contributing to several clusters. This

suggests that village of residence is not as strong a predictor of genetic population structure compared to broader linguistic group classification (See appendix 7 for admixture graphs).

2.3.3 Principal Components Analysis

Principal components Analysis (PCA) of all EHPNG samples reflected the strength of linguistic grouping in describing genetic diversity in EHPNG. Each progressive principal component appears to separate individual linguistic groups from the rest of the data. Appendix 3 shows PCA of 18 EHPNG linguistic groups and an analysis of a contribution of each linguistic group to each eigenvector of the PCA analysis. Subtle effects of linguistic group membership can be observed, with the majority of eigenvectors appearing to involve separation of linguistic groups (Eigenvectors 2,4-10,12,14,15) while other eigenvectors appeared to be based more on general geographical clines existing in the region (Eigenvectors 1,3,11,13).

PCA analysis of 20 villages in the South Fore, North Fore and Gimi linguistic groups did not show a clear structure in comparison to the wider linguistic group classification (Appendix 8). Principal components 1 and 2 did not cluster individuals in their villages of residence clearly. The spatial orientation of loose clustering of village residents did not conform greatly with their spatial orientation within the region. That being said, the two villages from the Gimi and North Fore linguistic groups group together and the villages at the geographical extremes of this sub-region were at the margins of the plot with groups in the centre clustering together. When the outputs are averaged and contributions to the principal components are measured per village there is no outstanding group that dominates the first principal component, although individuals from the village Ilesa at the extreme east of the region contribute 26.6% of the variation to the second principal component. This analysis suggests that developing clear distinction between groupings of village members at a village level is much more challenging than at linguistic group level.

PCA analysis was performed on the *Oceania* merge dataset, this analysis clustered EHPNG samples with other PNG samples, separated from more distant Oceanic populations in the analysis.

2.3.4 Runs of Homozygosity (ROH)

Runs of Homozygosity from all samples from EHPNG showed elevated levels of ROH compared to continental populations in the analysis (See appendix 4 for tables of homozygosity). When compared to other groups in the dataset EHPNG have the greatest enrichment in ROH. EHPNG members of the dataset having on average 14.13 (SD 4.96) ROH longer than 1MB. This is longer even than well-established population isolates like the Batwa Pygmy people (mean 13.2, ROH >1MB, SD 11.03) and data from other groups outside EHPNG but from PNG (mean 9.33, ROH >1MB, SD 5.71). Within EHPNG there was variation between groups, with the highest number of ROH being in the Anga (24.65, SD 7.56) and the lowest being in the Labogai (8.55, SD 3.23). This may reflect the population densities and degree of isolation and genetic drift the different groups have experienced. When average number of ROH is plotted against population density, according to a previous study of these populations, there is no significant relationship between these variables in the analysis (R^2 0.013; $p = 0.31$).

2.3.5 CP Analysis and FS Clustering

Concerns relating to ascertainment bias and the lack of polymorphic variants in PNG populations in the data sets have placed an emphasis on a haplotype based approach in characterising the population structure in EHPNG. The extensive sampling in this study and distinctive patterns of high linkage disequilibrium mean that long haplotype blocks were inferred in EHPNG which contained sufficient polymorphic variants to allow resolution of population structure, comparable to other studies using the same methodology. The use of data from other PNG regions generally had lower resolution. This was acceptable however as this data was largely included comparatively to further understand structure in EHPNG which is the focus of this thesis.

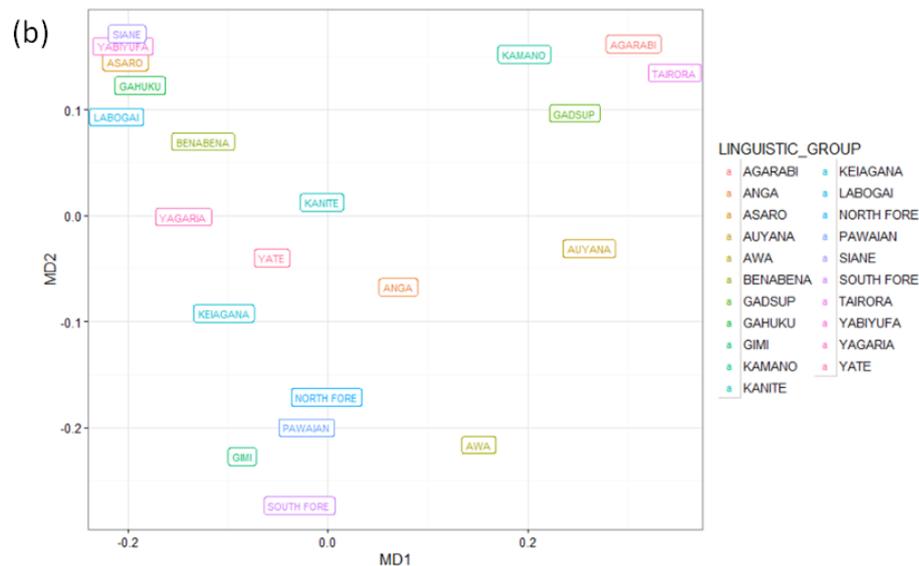
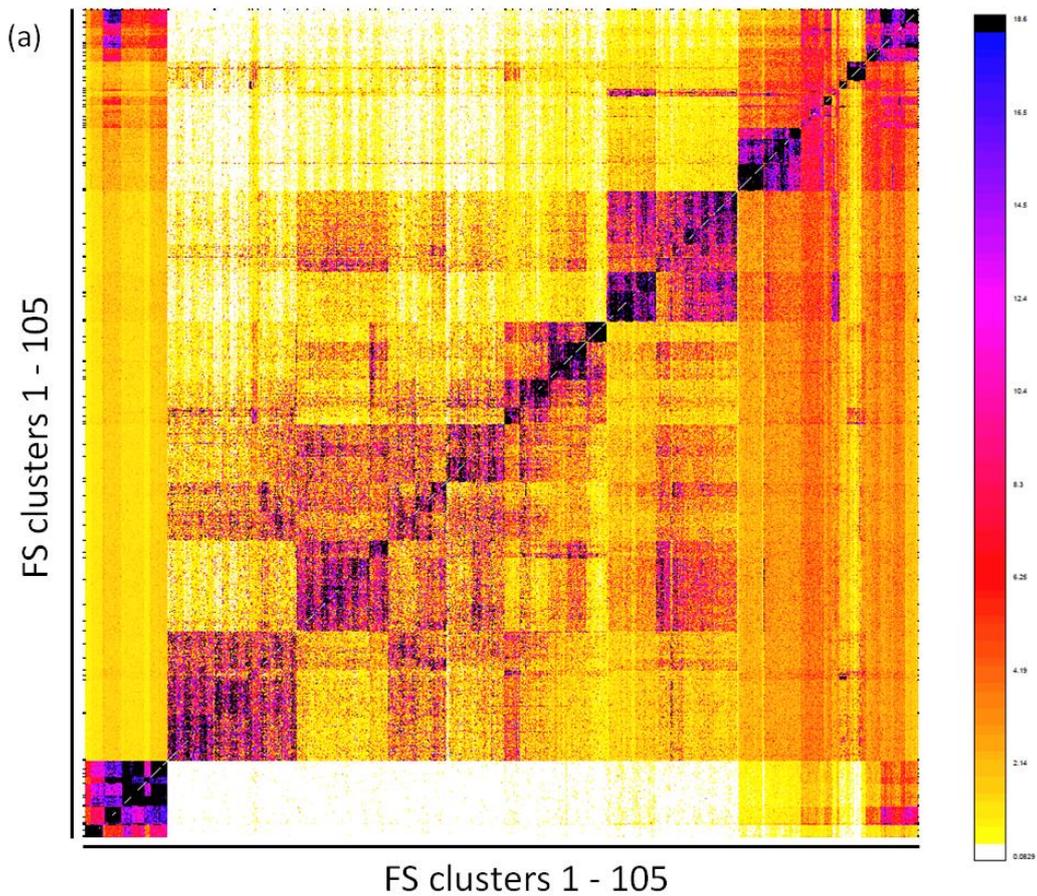


Figure 2.1. (a) heatmap of CP analysis of PNG merge (1,462 individuals, 49 populations). Depth of shading reflects inferred relatedness between pairs of individuals. Individuals are grouped into FS-derived genetic clusters, individuals are ordered on each axis identically. The darkly shaded blocks reflect the strong population structure best represented by linguistic group membership. (b) MDS plot using CP output from EHPNG individuals. Individual's CP output has been averaged for each linguistic group in EHPNG to give averaged CP copying vectors. Manhattan distances between these vectors are then used to create a matrix that is reduced using MDS. The placing of the linguistic group data corresponds remarkably to the geographical spacing of linguistic groups in EHPNG (see map of region figure 1.5.) (c) Overleaf, a table of the full FS clustering the heatmap axes are based upon.

A heatmap of individuals from EHPNG under CP analysis shows clear population structure when individuals are grouped together by linguistic group classification alone (Appendix 5). The heatmap also groups together linguistic groups in broad western, southern and eastern blocks. The most distinctive linguistic groups in the heatmap are the Anga and Pawaian groups with deep blocks of shading and little copying from other individuals in the analysis. Streaks in the heatmap appear to reflect individuals who have distinct ancestry profiles to other individuals from the same linguistic group possibly reflecting individuals who have migrated to regions from other linguistic groups. The analysis with additional samples from other PNG Highland regions, PNG coastal groups, Oceanic and Asian samples also showed this structure within EHPNG. To ascertain if this placement of groups in linguistic categories and then broad geographical categories based on a priori knowledge resulted in a possible confirmation bias, FS clustering of CP output was performed and again showed groups being best categorised by linguistic labels at a genetic level. The analysis placed EHPNG individuals into 51 separate clusters. 16 of these clusters were comprised of individuals from the same linguistic groups, 17 clusters were comprised of more than 95% of individuals from just 2 neighbouring linguistic groups and the remaining 18 clusters were mixed, with members from more than two usually neighbouring linguistic groups. Individuals from the Anga and Pawaian linguistic group are placed in clusters only comprising other Anga or Pawaian individuals. Figure 2.1 (a) shows a CP heatmap with individuals grouped according to FS cluster assignment. Ascending the FS tree resulted in merger of clusters with greatest affinity and further demonstrates the affinity of individuals to share similar ancestry profiles with individuals from the same linguistic group. This can best be seen in a comparison of the CP heatmap ordered by linguistic group membership with a CP heatmap ordering based on FS classification of individuals (Appendix 5).

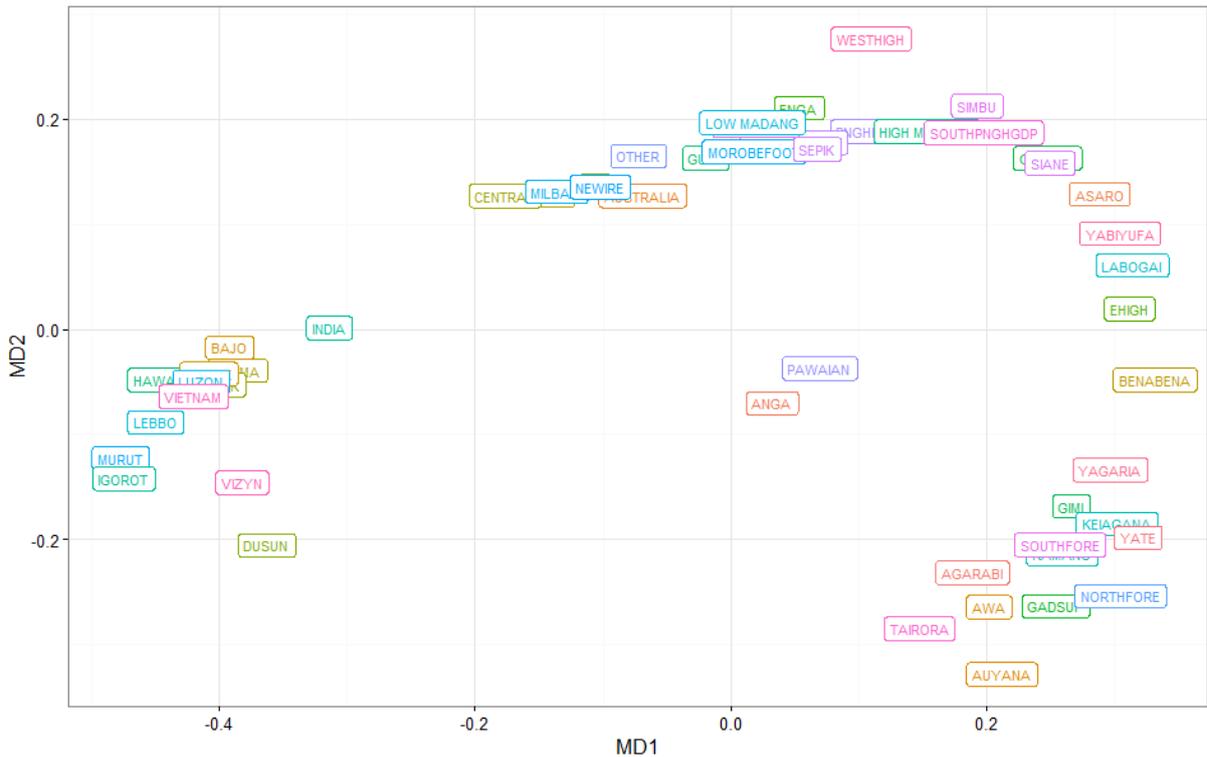


Figure 2.2 MDS plot of CP output of an analysis from the *Oceania merge* dataset, This is based on conversion of the original matrix to create CP copying profiles for each of the groups, This plot is based on multiple dimensional scaling (MDS) of a pairwise distance matrix (TVD) of these group copying profiles. Non PNG populations separate distinctly in MD1 and MD2 separates the highland groups. The Anga and Pawaian groups appear distinct from all other highland groups in this analysis.

Multi-Dimensional Scaling (MDS) of CP output (Figure 2.1b) places linguistic groups in positions in scales one and two that when plotted reproduce the spacing and separation of the linguistic groupings geographically in EHPNG.

To assess structure at the level of village of residence using CP output, individuals from South Fore, North Fore and Gimi villages were extracted as these linguistic groups had more than a single sampling village location. A reduced heatmap was plotted placing in individuals in their villages of residence on the heatmap. There were no indication of high relatedness between individuals from the same village on this heatmap. In fact the strongest distinction appeared to be between all villages in the North Fore linguistic group and all other villages in this sub-region.

CP analysis of EHPNG individuals being painted by a panel of non EHPNG PNG Highland individuals and PNG coastal groups as donors shows uniform copying profiles across all linguistic groups except

the Anga and the Pawaian. When a different donor panel comprising individuals from outside of PNG was used there appeared to be uniform painting profiles across all linguistic groups including the Anga and the Pawaian.

TVD averaged between members of each of the linguistic groups plotted against straight-line geographic distance shows a linear relationship between these 2 variables (R^2 0.32). For linguistic groups with more than one sampling location the averaged position was used. Repeated analysis, this time removing the Anga linguistic group shows a slightly increased correlation (R^2 0.33). The same analysis performed for villages in the kuru region with more than three samples shows an even larger correlation (R^2 0.41, $p < 0.001$). When done at the individual level this relationship between variables is weaker (R^2 0.18, $p < 0.001$).

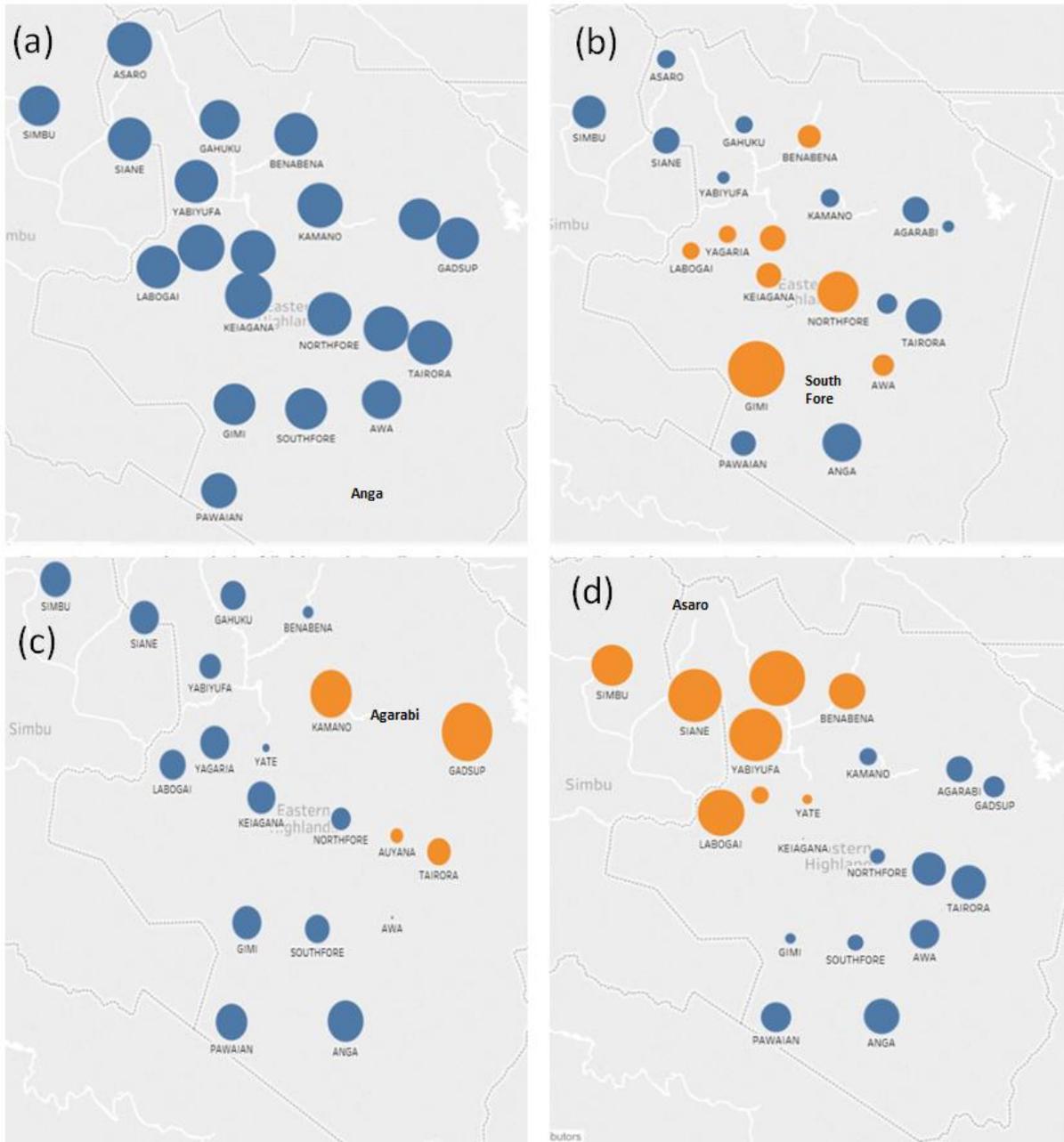


Figure 2.3 Geogenetic analysis of (a) Anga, (b) South-Fore, (c) Agarabi, and (d) Asaro linguistic groups. Colour of shading for each linguistic group reflects if group is more closely related (orange) or more distantly related (blue) than predicted through geographical distance. Size of the circles reflect the value of this distance. This further confirms linguistic group designation as the best predictor of overall population structure observed in the region. The Asaro belong to the Gahuku-Bena-Bena sub-family of the East-Central family and have affinity for other members. The Agarabi belong to the Gadsup-Auyana sub-family of the Eastern family. The Anga are considered a separate stock language and the South Fore are part of the Fore-Gimi subfamily of the East-Central family.

Analysis of relationship between geographic distance and TVD via the derived exponential function revealed relationships between groups that conform to previous linguistic classifications more than geographical distance measures. There is a general trend observed for groups to have general agreement between expected geogenetic distance calculated by the exponential function and observed TVD genetic distance. Aberrations from these expectations are informative and are reflective of relative placements of languages within sub-branches of the Trans-New-Guinea language group. The Anga and Pawaian appear distant to all groups far greater than expected reflecting their placement as separate stock-languages. Groups within the different Central and Eastern sub-branches of the East New Guinea Stock of the Trans-New Guinea language show reduced genetic distance than that predicted by geographic distance.

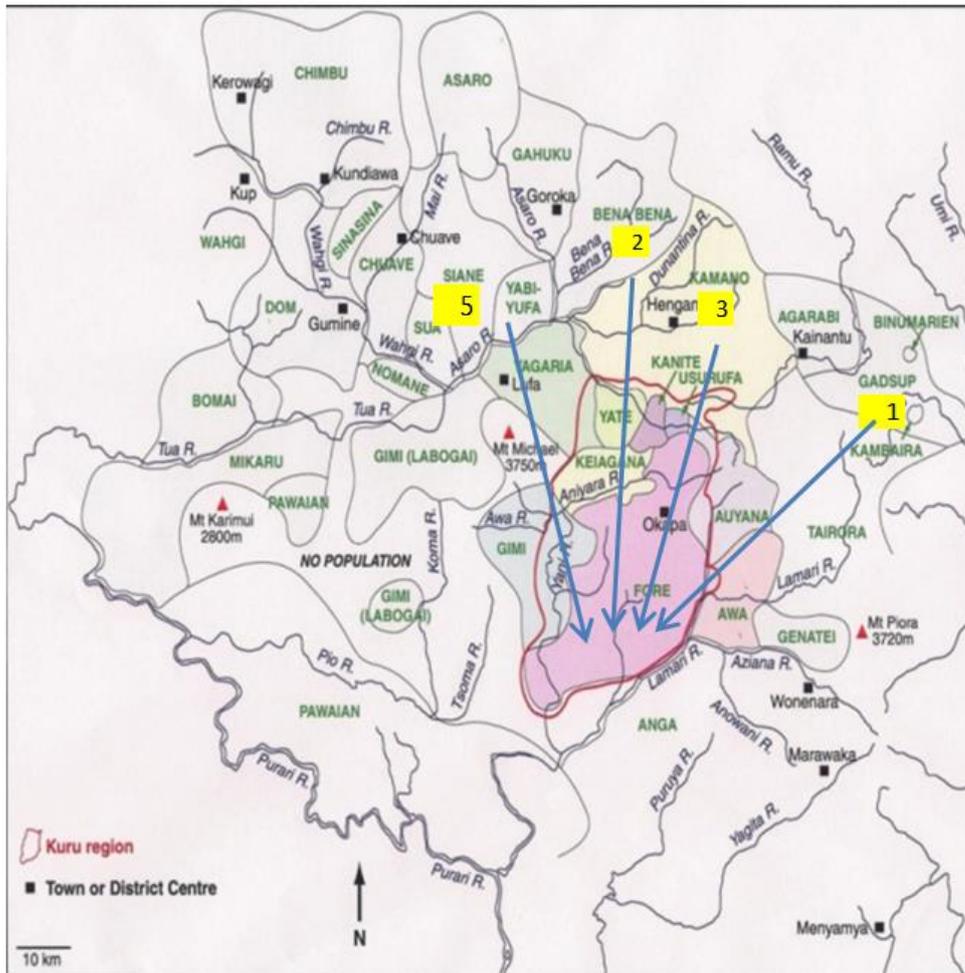


Figure 2.4. Analysis of long distance migrants within the South Fore population after removing individuals with profiles most closely resembling those of neighbouring linguistic groups shows 11 individuals who have ancestry profiles more similar to distant linguistic groups. Two individuals who had ancestry profiles resembling other regions of PNG (Simbu and Morobe footlands) are omitted in this plot. They were born in 1955 and 1966 respectively and may have moved to the region after the initiation of migrant labour exchange schemes in the colonial period.

Analysis to identify migrant individuals in the dataset with aberrant CP copying profiles to the average of their linguistic group reveals genetic evidence for individuals moving across multiple linguistic groups to take up new residences (See appendix 5 for table of results). Individuals that met the criteria for being classified as a migrant had on average a TVD of 0.48 ($n=51$, SD 0.15) from their own linguistic group average copying profile (See appendix 5 of histogram of pairwise genetic distances between individuals and the super individual copying profiles that represent the aggregated ancestry of each linguistic group). Individuals who had CP copying profiles more similar

to other linguistic group CP copying profiles, but failed to meet migrant classification due to their best matching linguistic group CP copying profile being a neighbouring linguistic group had an average TVD of 0.34 (n=178, SD 0.12) from their own linguistic group. These individuals were excluded to remove uncertainty that in fact their aberrant CP copying profiles which resembled neighbouring groups were not in fact due to recent migration over substantial distances but instead possibly due to the small genetic differences between neighbouring linguistic groups. The average TVD of non-migrants from their own linguistic group average in the dataset was 0.16 (n=1,080, SD 0.11). Analysis between different linguistic groups and their proportion of migrants does not appear to show any clear trends with linguistic groups from different points of EHPNG having migrants. Some of these individuals appear to have greater affinity to populations outside of EHPNG and five individuals had CP copying profiles most closely resembling groups from coastal regions. The Gahuku linguistic group contains four individuals with CP painting profiles most closely resembling the Labogai. These individuals may represent a subgroup of individuals who are the result of a migration of an entire village in the recent past. The absence of Labogai individuals with Gahuku resembling CP copying profiles appears to point to the unidirectional nature of the movement and also provides further support that the appearance of migrants is genuine and not due to artefactual similarities between the copying profiles of the linguistic groups.

2.3.6 Admixture Analysis Using fastGLOBETROTTER

No evidence of admixture was detected within all PNG Highland populations when outside populations were used as potential admixing sources using fastGLOBETROTTER. Evidence of admixture within the South Fore linguistic group was detected with an admixture episode estimated at 9.8 generations ago (SD 0.40, 20 bootstrap samplings). Figure 5 shows the breakdown of the contribution of sources to the sides of this 2-way admixture reveal the sides best represented by Western and Eastern EHPNG groups respectively. This was achieved by analysis of principal components of contributing sources to the estimated 2-sided admixture episode.

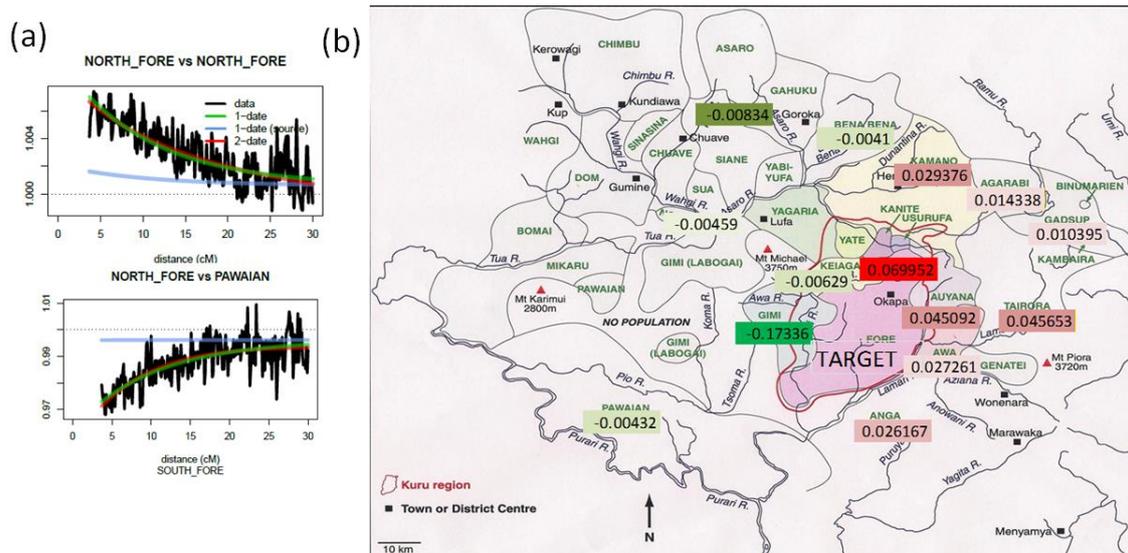


Figure 2.5. (a) Two admixture curves showing the inferred Pawaian and North fore haplotype blocks detected in the South Fore are spaced and separated in such a way as to conform with expectation of an exponential decay after an admixture event. (b) Right, breakdown of contributing sources to each side of the inferred admixture event estimated to have occurred 9.8 generations previously. The South Fore are believed to have extended into their current region in relatively recent times in search for new lands to occupy and build their communities. This expansion may have led to interaction and admixture with groups they had not previously interacted with.

2.4 Discussion

Population genetic analysis performed in this chapter has allowed the delineation of the population structure of EHPNG in an unprecedented manner on a fine scale not usually observed in other regions of similar geographic extent. EHPNG is a unique region of the world that remained out of the influence of major continental demographic and cultural forces for tens of thousands of years until the middle decades of the 20th century(105). This analysis has also provided a valuable platform for further understanding of the kuru epidemic and its impact on the affected communities, the focus of the remainder of this thesis.

The genetic properties of EHPNG as a whole reflect expectations for a region in relative isolation from the major population movements of human history and lacking regular contact with external communities. Runs of homozygosity for samples from the region are only matched by the African Batwa population. The Batwa lived at low population densities throughout history and had limited interaction with surrounding communities (166). The high degree of genetic drift observed in EHPNG

may have been exacerbated by historic population bottlenecks caused by famine, environmental catastrophes and disease epidemics(167).

Linguistic group classification appears the best descriptor of population structure in EHPNG as evidenced by PCA, ADMIXTURE, F_{ST} , CP and FS analysis. Individuals have a strong tendency to cluster with one another when from the same linguistic group. This is remarkable given the geographical extent of the region under study. The average F_{ST} values displayed between the 22 linguistic groups are greater than those observed between Spanish and Finnish populations in Europe separated by 2,900km(84).

Broader linguistic relationships are visible and predictive of genetic affinity, as observed in the geographical analysis of relationships between linguistic groups. Large deviations from the predicted genetic distance based on an exponential function with parameters derived from the observed CP genetic data conform to branches, sub-branches and stock-isolates within the Eastern Highlands branch of the TNG language family. These classifications were based on the work of Stephen Wurm(74). The veracity of the claims by Wurm have been held into question by more recent linguists who have questioned the capability of making firm classifications of such distinct and understudied languages with the data available at the time(73). The findings in this analysis provide some genetic support for the classification of languages proposed by Wurm (see appendix 10 for Wurm's groupings). Other studies have shown the affinity between present genetic relationships and historical linguistic roots(168). This classification of linguistic groups is further supported by FS clustering, ADMIXTURE clusters and PCA analysis (See appendices).

The affinity for linguistics to be an excellent descriptor of the population structure in the region should not be interpreted as causative of group separation and population structure in EHPNG. Individuals in EHPNG pre-European contact did not know the geographic extent of their linguistic speaking domain show that it was not a mode of self-identification or carried significant local cultural significance. The existence of marriage exchange across linguistic boundaries and the

observation of bilingualism in border areas reflect that groups of individuals speaking the same language were not restricted to their own linguistic domain. This means that linguistic classification cannot be considered fully deterministic of reproductive processes and biological affinity. Instead, the linguistic group categorisation can be seen as a proxy for collections of individuals who overwhelmingly have similar ancestral profiles due to residing in close proximity in a region with great restriction of movement due to geography and warfare. In other regions with reduced linguistic diversity and greater population mixture, linguistic boundaries are more blurred (in terms of genetic distinction), making linguistic categorisation a more difficult proxy for population structure(165).

This can be seen by the correlations observed between genetic and geographical distance of individuals and linguistic groups observed in EHPNG. Geographical distance “as the crow flies” has limitations, as it does not accurately reflect historical footpaths and the physical effort required to traverse landscapes. This is particularly relevant in EHPNG which has extreme terrain with alternating grassland, forests, river systems, and swamps on land with marked variation in inclination. A measure that encompasses all of these factors would enhance the descriptive power of geography in describing group structure. Generating such a measure was explored as part of this study but proved to be challenging due to the huge amount of mapping information involved. Such an atlas of thoroughfares in the region would have limitations as of the information regarding historical routes of travel has been lost since the vast overhaul of the landscape since European contact and administration.

Another limitation of geography as a descriptor compared to linguistics in this region is that geographical measures between individuals and groups would be based on their spatial arrangement at the time of sampling. Groups in this particular region have moved due to having subsistence methods that have been referred to as ‘proto-agricultural’(103). This involves a mixture of sedentary and migratory behaviour of groups as they aim to exploit new land to cultivate crops

and still retain practices of hunting and gathering. Linguistic identification of groups does not have this restriction as affinity between language speakers is known to correlate with common ancestors who spoke the same languages regardless of their spatial location presently(169).

The greater relevance of linguistic group classification over geographical distance between points in describing population structure is clearly evident with the contributions of individuals from the Anga and Pawaian linguistic groups. Their distinct patterns of genetic ancestry in CP and FS analysis compared to the other groups, extended ROH, PCA separation and elevated pairwise F_{ST} scores point to this. Their distinct CP painting profiles when other PNG data are used as donors reflect different ancestral history for these groups (Figure 2.6). This confirms findings in studies in other fields that reflect the distinct cultural and linguistic differences in these groups. Scholars have speculated that the Anga inhabited a larger territory in EHPNG in the past only in recent centuries having been pushed to the fringes of the region by incoming groups of agriculturalists living at higher population densities(100). The demarcation of their region and that of the neighbouring Fore at the time of European contact was the Lamari River, which perhaps formed a sufficient physical barrier to the continuation of this expansion. The distinctness of these groups is shown in the geographic analysis. The Pawaian are believed to have lived as semi-nomadic peoples on the southern extremes of the region, living at low population densities over a vast area with different subsistence methods to those observed in the rest of EHPNG(109). A simple linear relationship between geographic distance and genetic distance of individuals in EHPNG is strengthened by the removal of Anga individuals in the analysis. Correlation coefficient increases markedly on their removal in a linear regression analysis between TVD and geographical distance. The geogenetic analysis also shows a strong degree of separation for these groups despite the close geographic distance between them and their neighbours. This degree of separation is also exacerbated by the greater degree of genetic drift

experienced by these groups, especially the Anga who lived at greatly reduced population densities(108). The time depth of the separation between the Pawaian, Anga and other EHPNG groups is difficult to estimate. Their separation from other EHPNG in the CP analysis with other PNG Highland groups as donors but uniformity when Oceanic groups are used as donors (Figure 2.6) suggest this divergence has occurred since the peopling of PNG.

Assessing population structure at the level of village of residence of individuals is more challenging than at the linguistic group level. This is reflected in PCA and CP analysis of individuals from multiple villages within the South Fore, North Fore and Gimi linguistic regions. In both analyses the strongest distinction between individuals again appears based on linguistic group classification (Appendix 8). From a biological perspective, distinguishing villages of residence from one another is challenging due to the likelihood that common ancestry between inhabitants from neighbouring villages is likely to be more recent and more difficult to resolve than linguistic groupings. Processes of gene flow due to continuous migration within linguistic groups (see below) and movement of whole villages in recent history is likely to make such distinctions irresolvable(101, 103, 104).

A caveat when discussing this analysis of assessment of population structure based on village of residence is that villages in the region and patterns of residence have undergone drastic, recent change. In recent times people have congregated into larger villages in order to be closer to colonial infrastructure and goods(103). In the past smaller hamlets existed which would aggregate together to form parishes and then clans as the larger political units(59). This means that in many parts of EHPNG the villages of residence as set out in the colonial census units are not completely faithful references to long standing residential patterns and can only act as a proxy for local historic residential patterns.

The discovery of historical admixture in the ancestral histories of modern South Fore individuals reflects the dynamic and changing history of EHPNG region. It also demonstrates the importance of admixture in shaping population structure within a region much smaller than those where admixture has previously been inferred(170). This further illustrates the importance of considering that the history of a region and its inhabitants cannot be recapitulated by the arrangement of present day groupings and is not sufficient to fully understand this historic peopling of a region. Some commentators believe that the modern occupants of this region to have arrived relatively recently as the result of an expansion due to enhanced agricultural methods that that resulted in greater population densities and expansion into territory previously inhabited by other groups using different subsistent methods (See section 1.4.2 for more on the Ipomean revolution). The indicated admixture between a population resembling modern Western EHPNG linguistic groups the North Fore on one side and Eastern EHPNG linguistic groups on the other resulting in the modern South Fore population possibly represents an expansion and contact between groups of individuals that had not previously been in regular contact stimulated by this 'Ipomean revolution'. Methods exist that can provide inferences of the demographic history of populations and possible causal explanations of why present-day population structure exists. PSMC and MSMC methods have inferred changes in population sizes over time and estimates of historical split times between populations. This data requires whole genome sequence data and has been effective so far in describing trends over longer periods of history than would be relevant for the group structure within EHPNG.

Despite the clear population structure on display in the region, there are large numbers of individuals who do not display strong genetic affinity to other individuals living in closest geographical proximity to them. These individuals have distinct CP painting profiles, cluster with individuals from distant linguistic groups and other PNG regions in FS, and appear differently in PCA analysis

In the migration analysis performed, 51 individuals were identified who harboured sufficiently distinctive CP ancestry profiles to be considered likely migrants over considerable distances from within and beyond EHPNG. Drivers for such migration would have changed over time and in particular since European contact and colonial administration. Endemic warfare that existed in the region was both a source of displacement of individuals and communities and also a restriction on the free movement of individuals in the region. At first contact with gold prospectors in the region, European individuals were pursued by curious local people on their travels(105). These individuals would often stop following the travelling European and Australians due to the party entering a new territory that was held by an enemy group. This cycle would repeat itself every few miles, reflecting the dense patchwork of warring territories that existed in pre-contact EHPNG. Prior to colonial administration, in EHPNG there were no roads or other forms of modern transport infrastructure meaning individuals would have had to use handmade paths that traversed the extreme terrain in the region. Ownership of such paths and routes was complex, and use of a particular path or section of path would depend on having some form of acquaintance with the creator and would be out of bounds for individuals who had disputes with the owner(103). These limitations made journeys circuitous, indirect and dangerous in the case of contravening customs of obtaining consent for use of paths in the region.

Often conflict would result in whole communities becoming displaced and relocating in distant areas away from the original site of conflict and displacement. This is apparent with the repetition of village names in different linguistic groups (e.g. Yagusa, Kalu), reflecting this process of reestablishment of communities in different areas(100). This was enabled by the relative lack of land pressure in the region particular in the southern margins which were only recently inhabited by the South Fore and other groups. Several examples of individuals who experienced such long distance migrations in the region were recounted in an interview with the previous head of PNGIMR who lived in the region for several decades beginning in the mid-20th century. "I recall a man who told me he arrived in the South Fore region all of the way from the Tairora in the 1920's. He had originally

left that region due to a conflict there and made his way to the South Fore acting as a trade intermediary”.

After the establishment of colonial rule in EHPNG, travel over longer distances including to coastal areas of PNG were possible for the first time. Migrant labour schemes were established, roads were constructed and warfare was abolished. Monetisation of the economy facilitated movement and trade within the region and beyond. The presence of some individuals with CP ancestry profiles most closely resembling coastal or distant Highland communities most likely reflect this ‘opening up’ process that occurred in EHPNG and other PNG Highland areas subsequent to colonial control.

Anthropological studies of processes of migration support findings in this study of infrequent and rare migration of individuals over longer distances in EHPNG(101, 104). Movement of individuals between clans within the same linguistic group is believed to have been a much more common occurrence. Documented processes that drove shorter migrations included exogamous marriage, non-kin adoption and an age-mate system(51). Processes of short distance migration within and between linguistic groups would likely have played a role in shaping patterns of genetic diversity in the region as well, but at a finer scale undetectable in this study. Although less common, the observed long distance migrations would have been aided by observed practices of accepting migrants from distant linguistic regions. Land pressure and associated disputes were reduced in EHPNG compared to other places due to the fertility of land in the region, and also the necessity of groups to accept newcomers into their communities as a valuable source of warriors to aide in combat against their local enemies. The distinct shallow genealogies present in the region is believed to have aided in acceptance of migrating groups and enabled their descendants to be regarded as equals by others in the community with the external origin of their ancestors forgotten within one or two generations(101).

EHPNG represent a remarkable set of communities that have undergone a period of great transition in the last century. Prior to the middle decades of the 20th century this region had remained without

direct contact with coastal and external peoples for millennia. The findings from this research reveal a distinct population structure that has been forged over millennia affected by the distinctive terrain, cultural and linguistic diversity and dynamic processes of migration and admixture.

Future epidemiological studies of kuru face challenges posed by the population structure revealed in this study. The large degree of group diversification between kuru-affected and non-kuru affected communities make finding kuru related genetic variants challenging. Methods based on detecting aberrant patterns of linkage disequilibrium are made more difficult by the extensive genetic drift on display in the region.

2.5 Summary of findings

- Population structure is best described by linguistic group classifications.
- Presence of two exceptional population 'isolates within an isolate'.
- Detection of historical admixture event in the ancestries of individuals from the South Fore linguistic group reveals the role of admixture in defining present populations.
- Development of a model to understand processes of migration of individuals within the EHPNG region.

2.6 References

1. Anderson W. The collectors of lost souls : turning Kuru scientists into whitemen. Baltimore, Md.: Johns Hopkins University Press; 2008.
2. Connolly B, Anderson R. First contact. New York, N.Y., U.S.A.: Penguin Books; 1988.
3. Diamond J, Ordunio D, OverDrive I. Guns, Germs, and Steel. 2011.
4. Lindenbaum S. Cannibalism, kuru and anthropology. *Folia neuropathologica*. 2009;47(2):138-44.
5. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
6. Wurm SA. New Guinea area languages and language study. 1, 1. Canberra: Dep. of Linguistics, Research School of Pacific Studies, Australian National Univ.; 1977.
7. Ross M. Pronouns as a preliminary diagnostic for grouping Papuan languages 2005. 15-66 p.
8. Feil DK. The evolution of highland Papua New Guinea societies. Cambridge: Cambridge University Press; 2010.
9. Bergstrom A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science (New York, NY)*. 2017;357(6356):1160-3.
10. Wiesenfeld SL, Gajdusek C. Genetic structure and heterozygosity in the kuru region, eastern highlands of New Guinea. *American journal of physical anthropology*. 1976;45(2):177-90.
11. Gajdusek DC, Alpers M. Genetic studies in relation of kuru. I. Cultural, historical, and demographic background. *American journal of human genetics*. 1972;24:Suppl:1-38.
12. Lindenbaum S. Kuru sorcery : disease and danger in the New Guinea highlands. 2016.
13. Thornton IWB, Cook S, Edwards JS, Harrison RD, Schipper C, Shanahan M, et al. Colonization of an island volcano, Long Island, Papua New Guinea, and an emergent island, Motmot, in its caldera lake. VII. Overview and discussion. *Journal of Biogeography*. 2001;28(11-12):1389-408.
14. Mandeville E. Agnation, Affinity and Migration Among the Kamano of the New Guinea Highlands 1979. 105 p.
15. Alpers MP. Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3707-13.
16. B. Watson J. Society as Organized Flow: The Tairora Case 1970. 107-24 p.
17. Mathews JD, Glasse R, Lindenbaum S. Kuru and cannibalism. *Lancet (London, England)*. 1968;2(7565):449-52.
18. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas of Human Admixture History. *Science (New York, NY)*. 2014;343(6172):747.
19. Halder I, Yang B-Z, Kranzler HR, Stein MB, Shriver MD, Gelernter J. Measurement of admixture proportions and description of admixture structure in different U.S. populations. *Hum Mutat*. 2009;30(9):1299-309.
20. Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. *Elife*. 2016;5:e15266.
21. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. *Science (New York, NY)*. 2014;343(6172):747-51.
22. Haddon AC. Smoking and Tobacco Pipes in New Guinea. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 1946;232(586):1-278.
23. Bourke R. Sweet potato in Papua New Guinea 2002.
24. Mandeville E. Kamano Adoption 1981. 229 p.
25. Sorenson ER, Claessen HJM, du Torr BM, Griffith J, Hockings P, Jablonko A, et al. Socio-Ecological Change Among the Fore of New Guinea [and Comments and Replies]. *Current Anthropology*. 1972;13(3/4):349-83.

26. Whitfield JT, Pako WH, Collinge J, Alpers MP. Cultural factors that affected the spatial and temporal epidemiology of kuru. *Royal Society open science*. 2017;4(1):160789.
27. The Genomes Project C, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. *Nature*. 2015;526:68.
28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007;81(3):559-75.
29. Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, et al. Complex Patterns of Admixture across the Indonesian Archipelago. *Molecular biology and evolution*. 2017;34(10):2439-52.
30. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nature Methods*. 2011;9:179.
31. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS genetics*. 2012;8(1):e1002453.
32. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*. 2009;19(9):1655-64.
33. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. 2015;519(7543):309-14.
34. Perry GH, Foll M, Grenier J-C, Patin E, Nédélec Y, Pacis A, et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(35):E3596-E603.
35. Jewett EM, Steinrücken M, Song YS. The Effects of Population Size Histories on Estimates of Selection Coefficients from Time-Series Genetic Data. *Molecular biology and evolution*. 2016;33(11):3002-27.
36. Hunley K. Reassessment of global gene-language coevolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(7):1919-20.
37. Need AC, Kasperavičiūtė D, Cirulli ET, Goldstein DB. A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. *Genome Biology*. 2009;10(1):R7.
38. Bhatia K, Prasad ML, Barnish G, Koki G. Antigen and haplotype frequencies at three human leucocyte antigen loci (HLA-A, -B, -C) in the Pawaia of Papua New Guinea. *American journal of physical anthropology*. 1988;75(3):329-40.
39. Lopez S, Thomas MG, van Dorp L, Ansari-Pour N, Stewart S, Jones AL, et al. The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection. *American journal of human genetics*. 2017;101(3):353-68.

Appendices

Appendix 1 – Data Used in Analysis

PNG merge

POPULATION	N
CHINA	10
COLOMBIA	10
COLOMBIA (Colla)	10
DENISOVA	1
GREAT BRITAIN	10
INDIA	10
SPAIN	10
JAPAN	10
KHOESAN	10
VIETNAM	10
KENYA	10
MEXICO	10
PAKISTAN	10
SRI LANKA	10
YORUBA	10
NEANDERTHAL	1
EHPNG AGARABI	17
EHPNG ANGA	22
EHPNG ASARO	18
EHPNG AUYANA	22
EHPNG AWA	11
EHPNG BENABENA	24
EHPNG GADSUP	16
EHPNG GAHUKU	16
EHPNG GIMI	28
EHPNG KAMANO	24
EHPNG KANITE	3
EHPNG KEIAGANA	27
EHPNG LABOGAI	19
EHPNG NORTH FORE	126
EHPNG PAWAIAN	16
EHPNG SIANE	24
EHPNG SOUTH FORE	363
EHPNG TAIRORA	38
EHPNG YABIYUFA	22
EHPNG YAGARIA	87
EHPNG YATE	20
PNG HIGH EHIGH	26
PNG HIGH ENGA	10
PNG HIGH MADANG	45
PNG HIGH SIMBU	33

PNG HIGH SOUTHHIGH	85
PNG HIGH WESTHIGH	34
PNG LOW CENTRAL	39
PNG LOW EAST BRITAIN	4
PNG LOW GULF	27
PNG LOW MADANG	28
PNG LOW MILBAY	6
PNG LOW MOROBEFOOT	13
NEW IRELAND	3
PNG LOWLANDS OTHER	12
PNG SEPIK	11
Ust Ishim	1
Total	1462

Appendix 2 - Oceania Dataset

POPULATION	N
AUSTRALIA	2
BORNEO DUSUN	32
BORNEO LEBBO	4
BORNEO MURUT	8
BOUG	2
BURMA	20
CHINA	15
HAWAII	1
INDIA	15
INDON BATAK	3
PHIL BAJO	4
PHIL IGOROT	10
PHIL LUZON	8
PHIL VIZYN	6
PNG HIGH KOIN	3
PNG HIGH KOSIP	3
EHPNG AGARABI	18
EHPNG ANGA	21
EHPNG ASARO	18
EHPNG AUYANA	22
EHPNG AWA	11
EHPNG BENABENA	24
EHPNG GADSUP	16
EHPNG GAHUKU	17
EHPNG GIMI	22
EHPNG KAMANO	23
EHPNG KEIAGANA	27
EHPNG LABOGAI	19

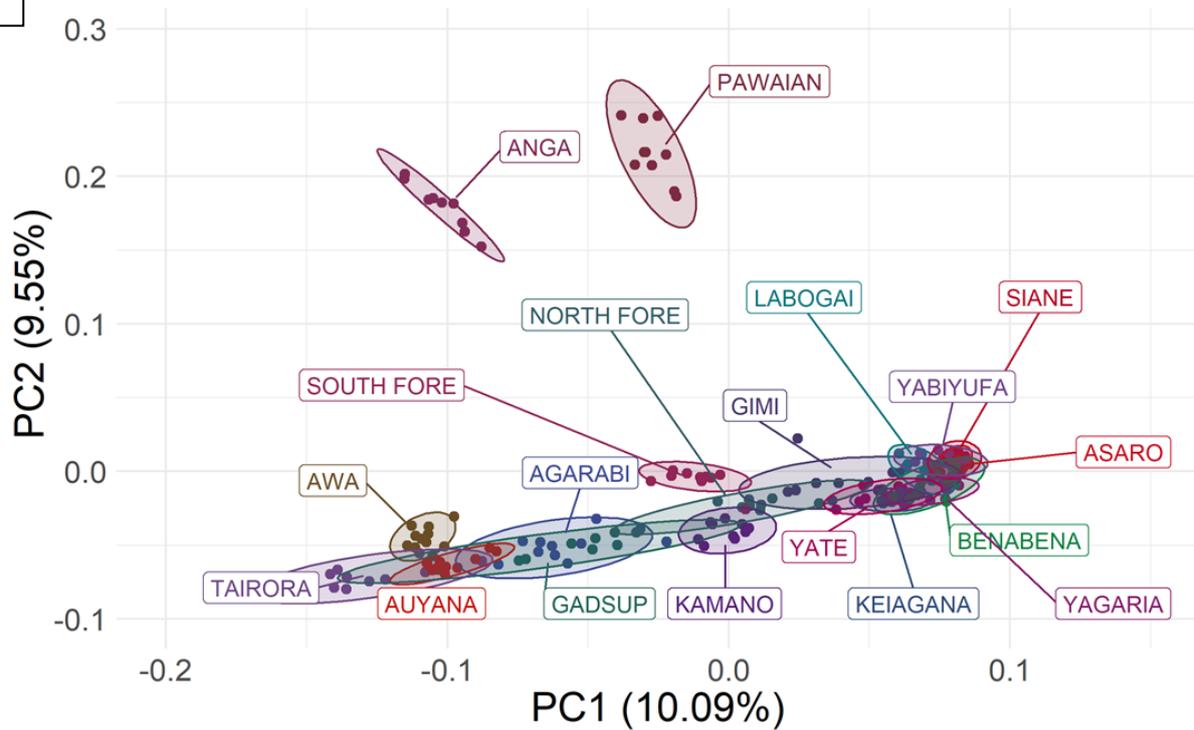
EHPNG NORTHFORE	52
EHPNG PAWAIAN	18
EHPNG SIANE	25
EHPNG SOUTHFORE	174
EHPNG TAIRORA	39
EHPNG YABIYUFA	22
EHPNG YAGARIA	77
EHPNG YATE	19
PNG HIGH EHIGH	26
PNG HIGH ENGA	10
PNG HIGH MADANG	45
PNG HIGH SIMBU	33
PNG HIGH SOUTHHIGH	85
PNG HIGH WESTHIGH	34
PNG LOW CENTRAL	39
PNG LOW EB	4
PNG LOW GULF	27
PNG LOW MADANG	28
PNG LOW MILBAY	6
PNG LOW MOROBEFOOT	13
NEW IRELAND	3
PNG LOWLAND OTHER	12
PNG LOWLAND SEPIK	11
SOUTHERN HIGHLAND	15
VIETNAM	17
Total	1208

This data represents samples remaining after sample merger, quality control procedures and highly related individuals being removed.

Analysis	Section	Figure	Number of Individuals	Name of Primary Dataset
FST	2.3.1	App 9	943	PNG merge
Admixture	2.3.2	App 7	200	PNG merge
PCA	2.3.3	App 3	180	PNG merge
ROH	2.3.4	App 4	943	PNG merge
ChromoPainter FS	2.3.5	2.1	1462	PNG merge
EHPNG MDS with Oceania Samples	2.3.5	2.2	1016	Oceania Merge
Geogenetic Analysis	2.3.5	2.3	1462	PNG merge
Migration Analysis	2.3.5	2.4	1462	PNG merge
EHPNG Admixture Analysis	2.3.5	2.5	943	PNG merge
Outgroup EHPNG Painting Science	2.4	2.6a	1293	PNG merge
Outgroup EHPNG Painting Oceania	2.4	2.6b	1016	Oceania Merge

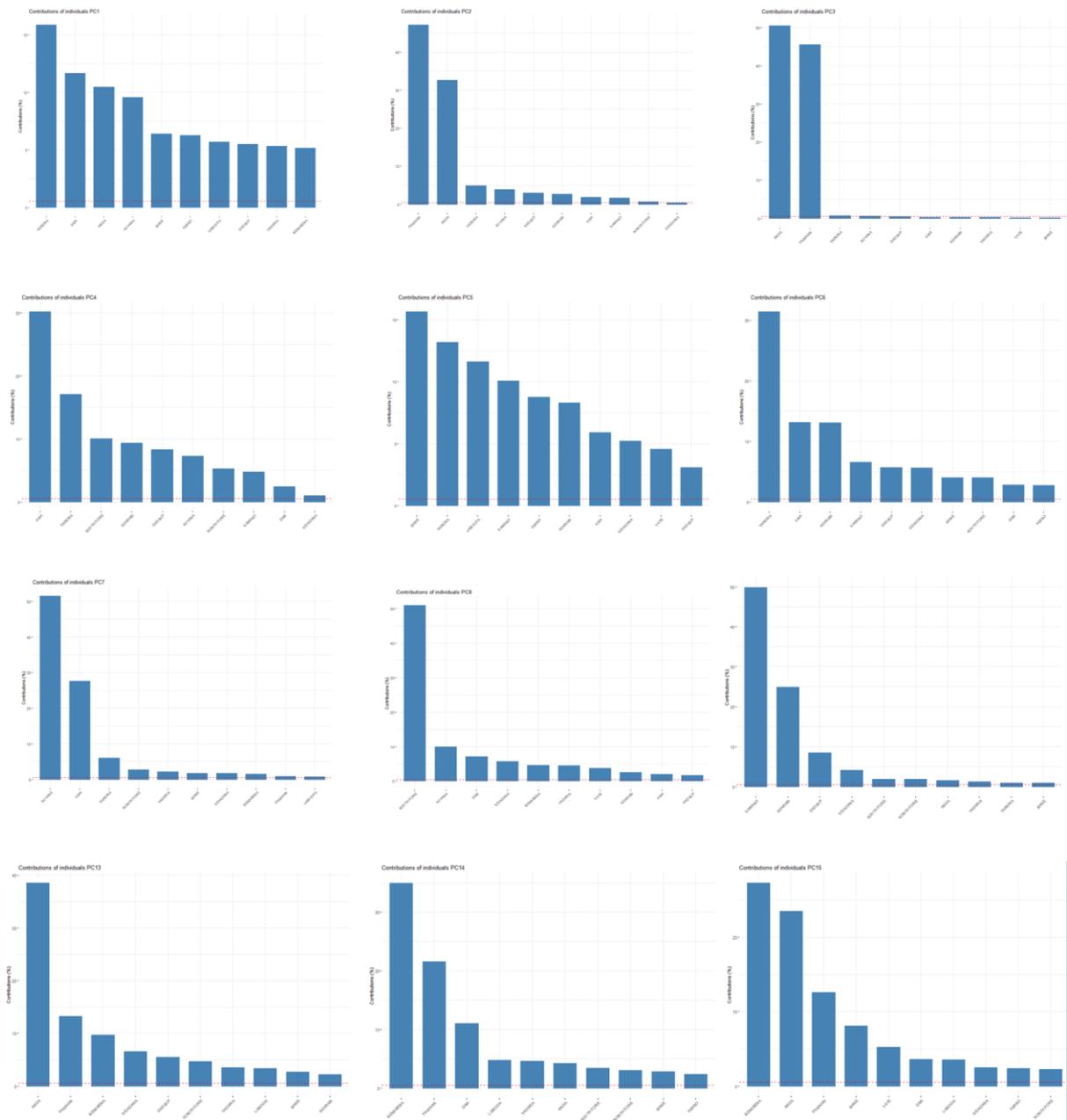
Appendix 3 – Principal Components Analysis of EHPNG

3a



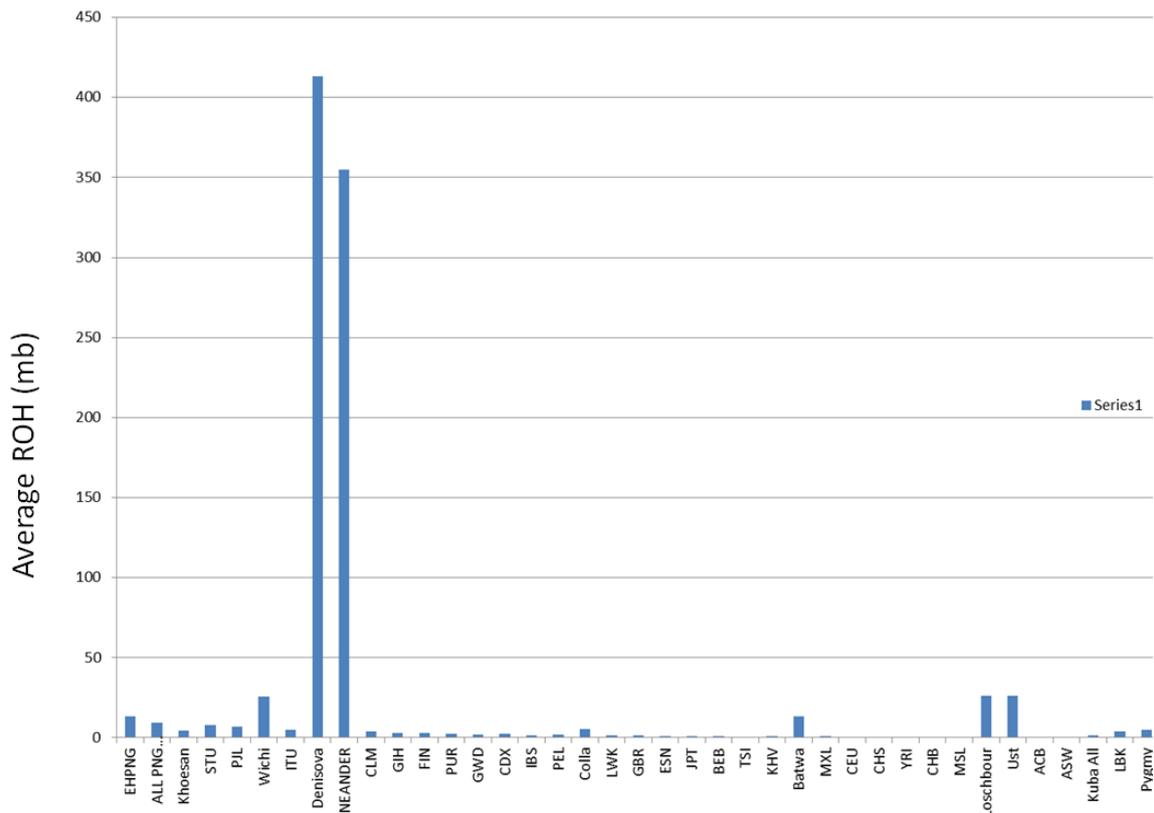
3a. Principal Components Analysis of 18 linguistic groups in EHPNG. Ten individuals were taken from each linguistic group to control for sample size bias and to permit analysis of each linguistic group contribution to each eigenvector. Principal component 1 appears to show an east to west cline and principal component 2 appears to separate the Anga and Pawaian linguistic groups that are described as ‘isolates within an isolate’. See Appendix 3b for group’s contribution to each of the first 12 eigenvector.

3b

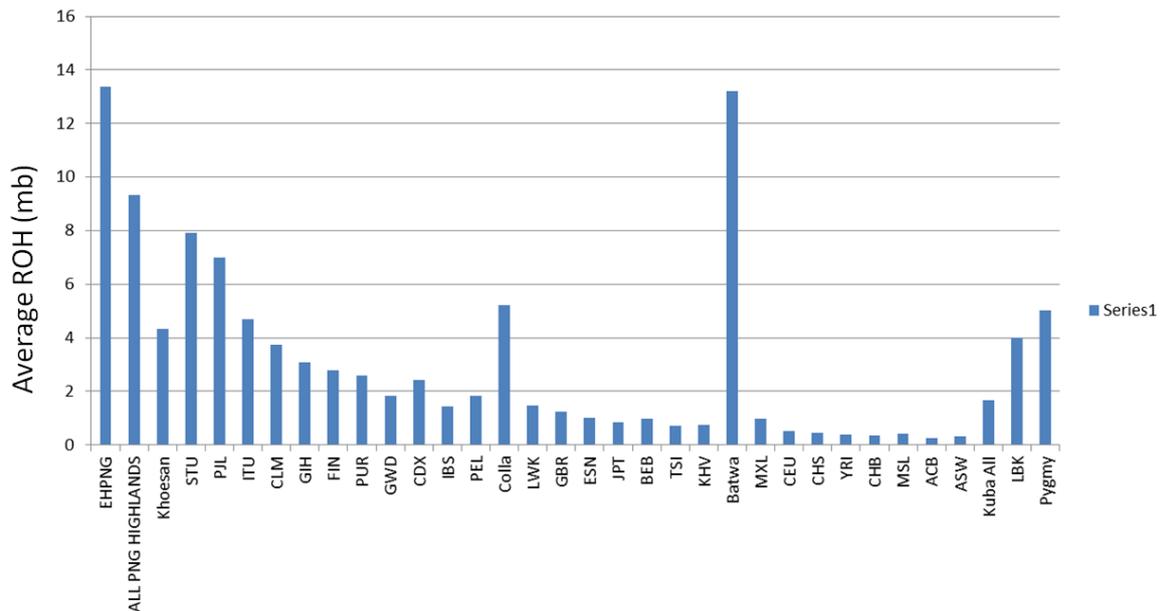


Bar charts showing linguistic group contributions to each of the first 12 principal components. For many of the principal components a single or small group of linguistic groups contribute greatest to the eigenvector. This further reflects the importance of linguistic group membership on describing population structure in EHPNG.

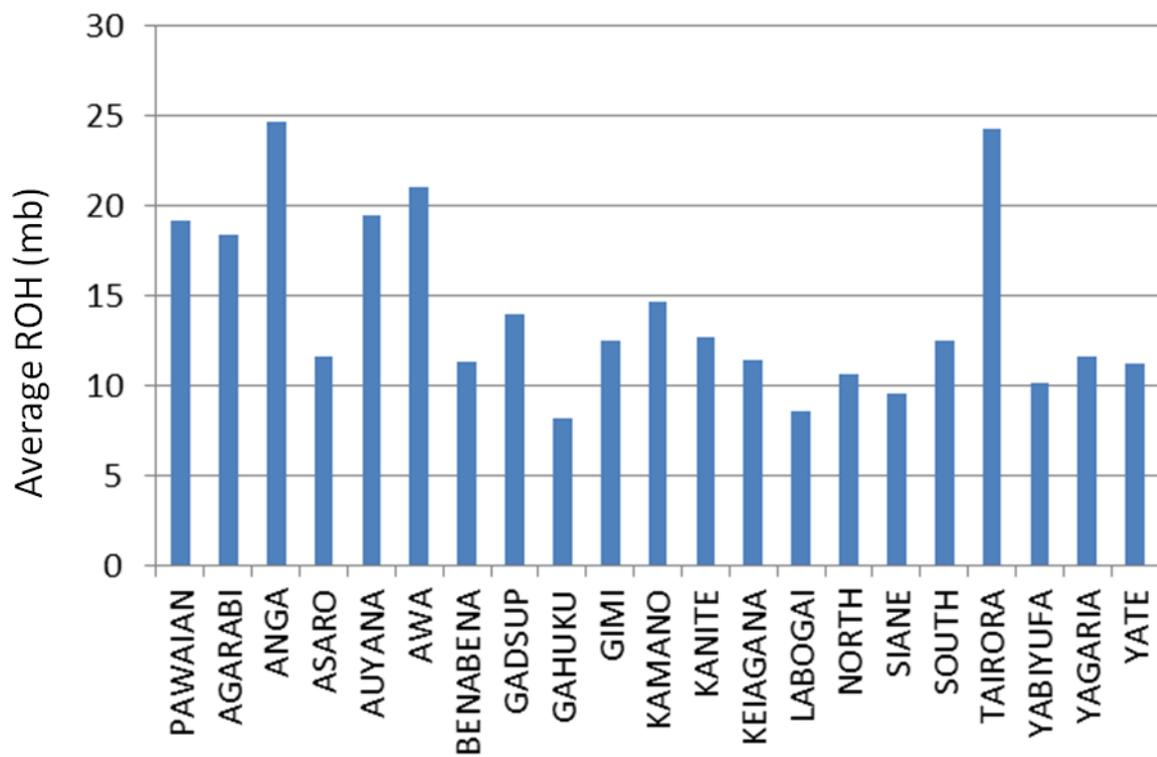
Appendix 4 – Analysis of ROH



Bar chart showing the average length of material in megabases classified as long runs of homozygosity by plink 1.9. Minimum length was set at 1mb.

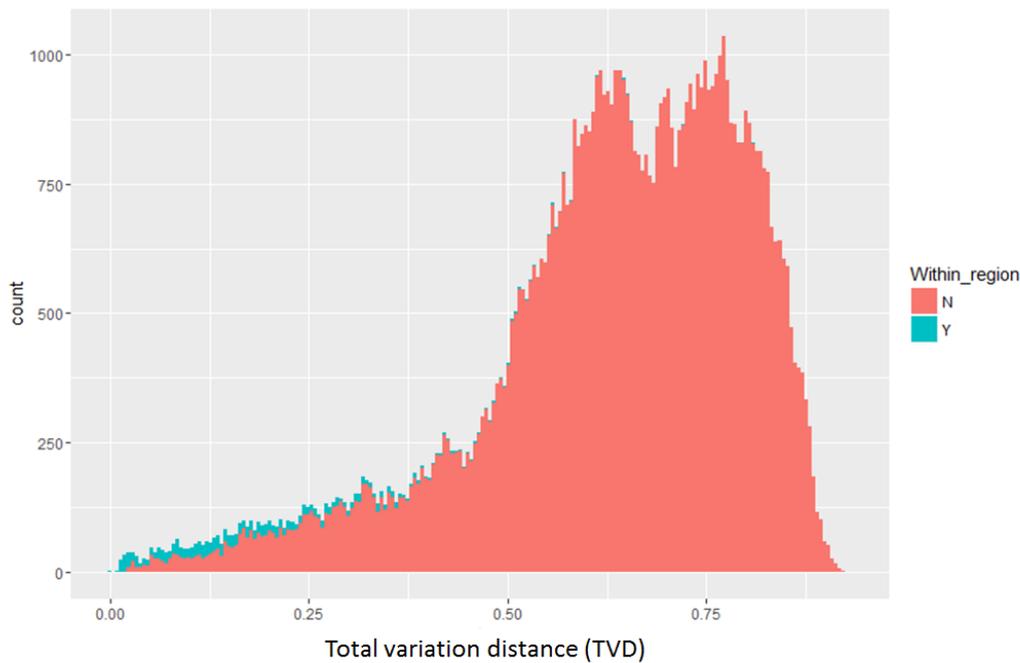


Same plot as above with Neanderthal and Denisovan populations removed. EHPNG population has highest average ROH figure similar to Batwa African population.



Barplot of average homozygosity of 21 linguistic groups within EHPNG.

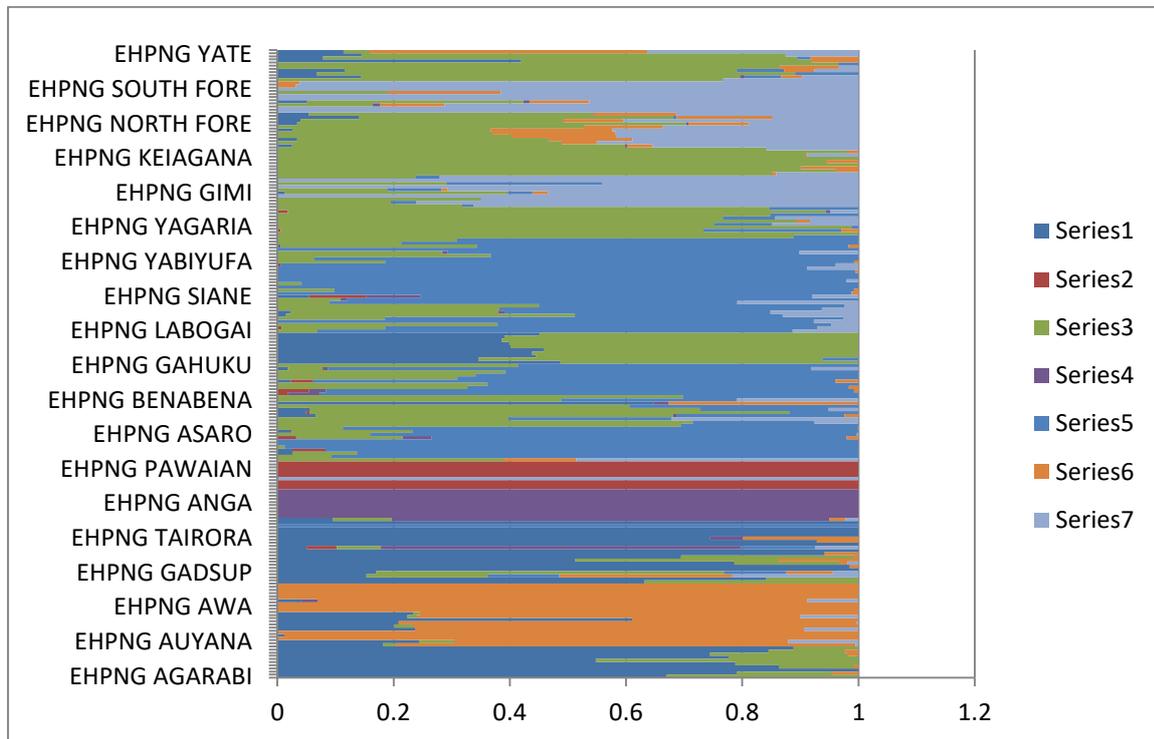
Appendix 5 Migration Analysis



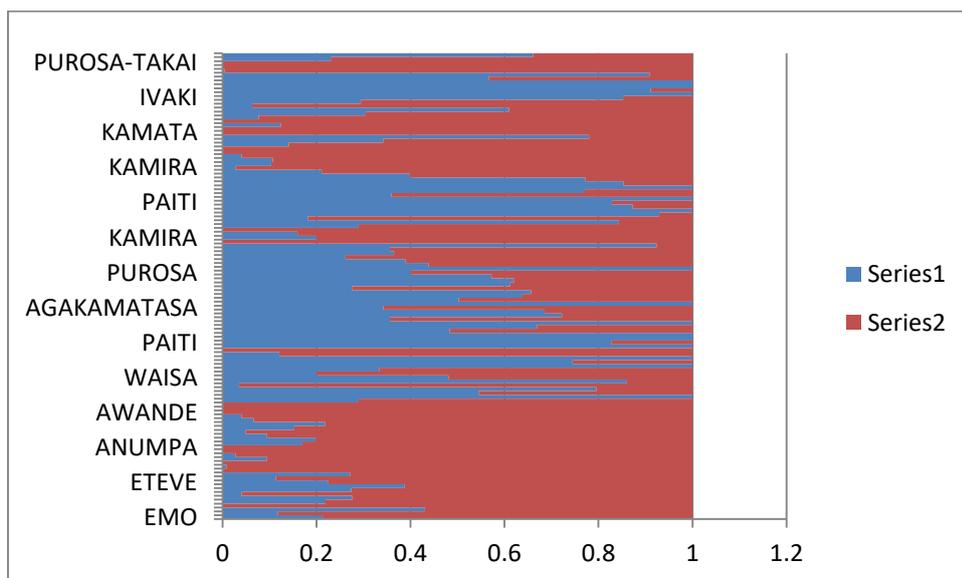
GROUP	N	MIGRANTS	PROPORTION
BENA BENA	24	6	0.25
GAHUKU	16	4	0.25
PAWAIAN	16	3	0.19
ASARO	18	3	0.17
GADSUP	16	2	0.13
TAIRORA	38	4	0.11
GIMI	28	2	0.071
SOUTH FORE	363	19	0.052
ANGA	22	1	0.045
AUYANA	22	1	0.045
YABIYUFA	22	1	0.045
SIANE	24	1	0.042
NORTH FORE	126	4	0.032
AGARABI	17	0	0
AWA	11	0	0
KAMANO	24	0	0
KANITE	3	0	0
KEIAGANA	27	0	0
LABOGAI	19	0	0
YAGARIA	87	0	0
YATE	20	0	0

Histogram of pairwise genetic distances between individuals and each of the 21 linguistic group super-individuals (top). The super individuals were the CP aggregate copying profiles for each of the 21 linguistic groups. Typically an individual would be expected to have a CP copying profile most similar to one composed of individuals from this own linguistic group. When an individual had a difference of 0.25 TVD from his own group super individual copying profile and had a copying profile more closely resembling a linguistic group separated by an intervening linguistic group that individual was classified as a migrant. The overwhelming majority of pairwise TVD values between individuals and the super individual of their own group is less than 0.25 (blue blocks). Results summarised in table (below).

Appendix 7 Admixture Analysis

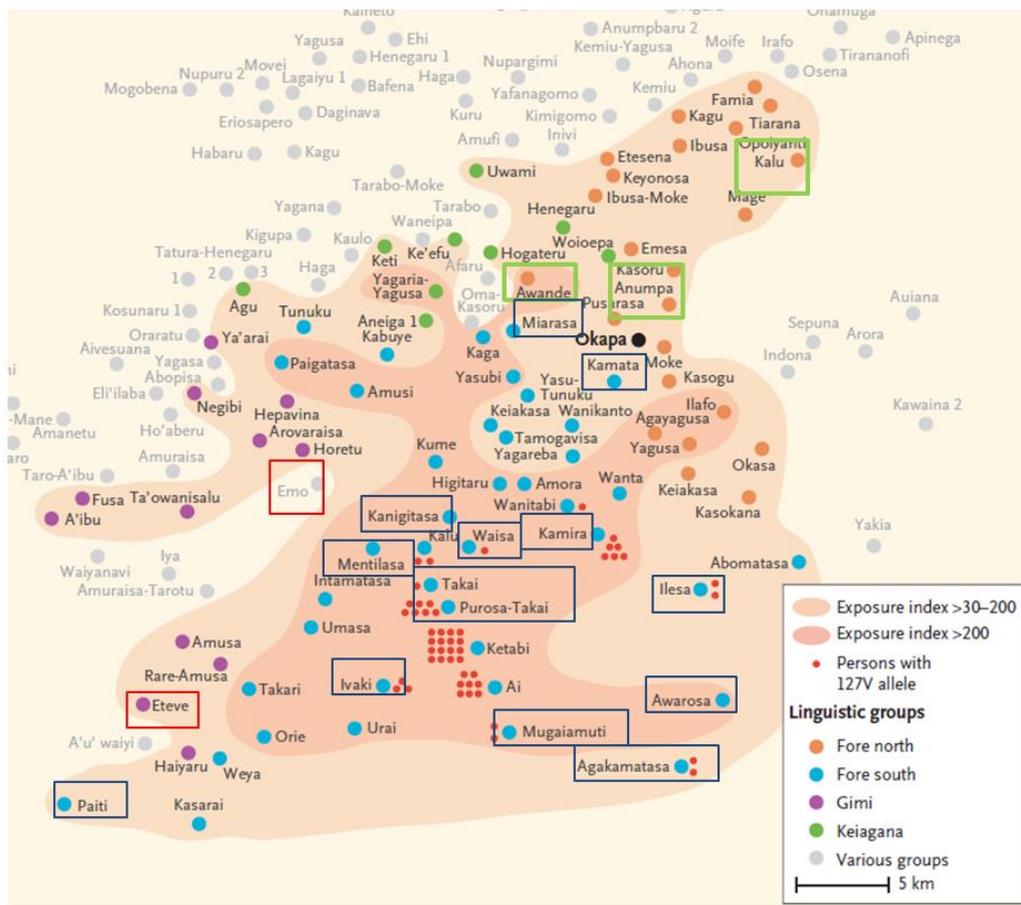
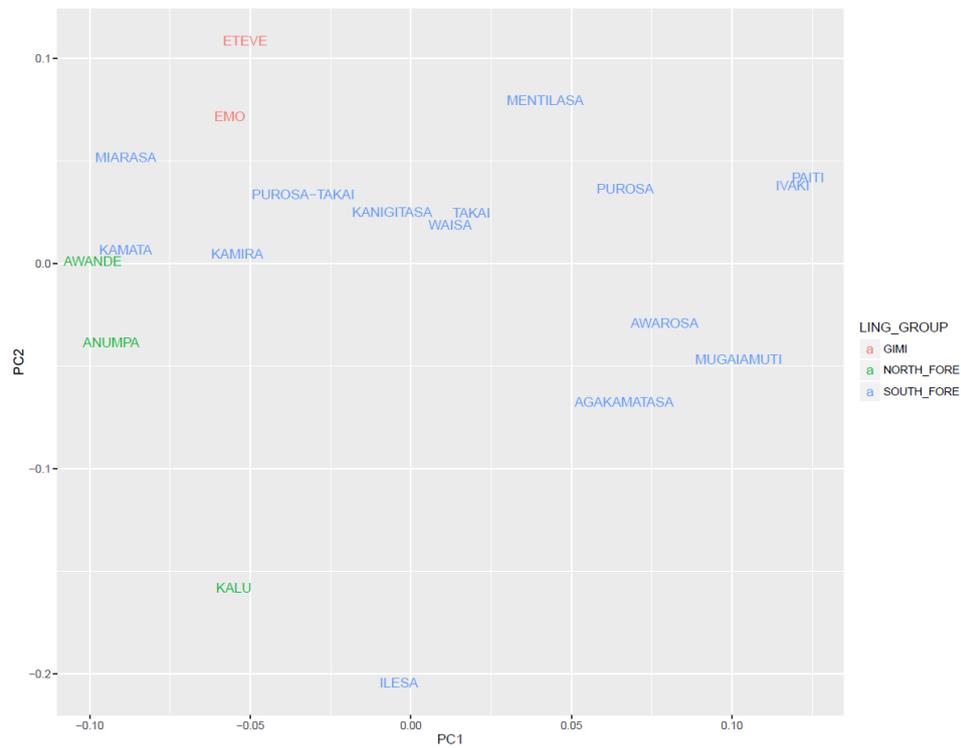


Admixture analysis of only 200 EHPNG individuals from 20 linguistic groups. The admixture analysis does appear to describe overall trends in the population structure when $K=7$. Distinct bands for Pawaian and Anga groups. The orange colour separates the eastern Awa and Auyana and the light blue represents a component from southern groups.



Admixture analysis of 200 individuals from 20 villages from the South Fore, North Fore and Gimi linguistic groups ($K=2$). No clear trends are apparent with analysis at this fine level.

Appendix 8 – Village PCA

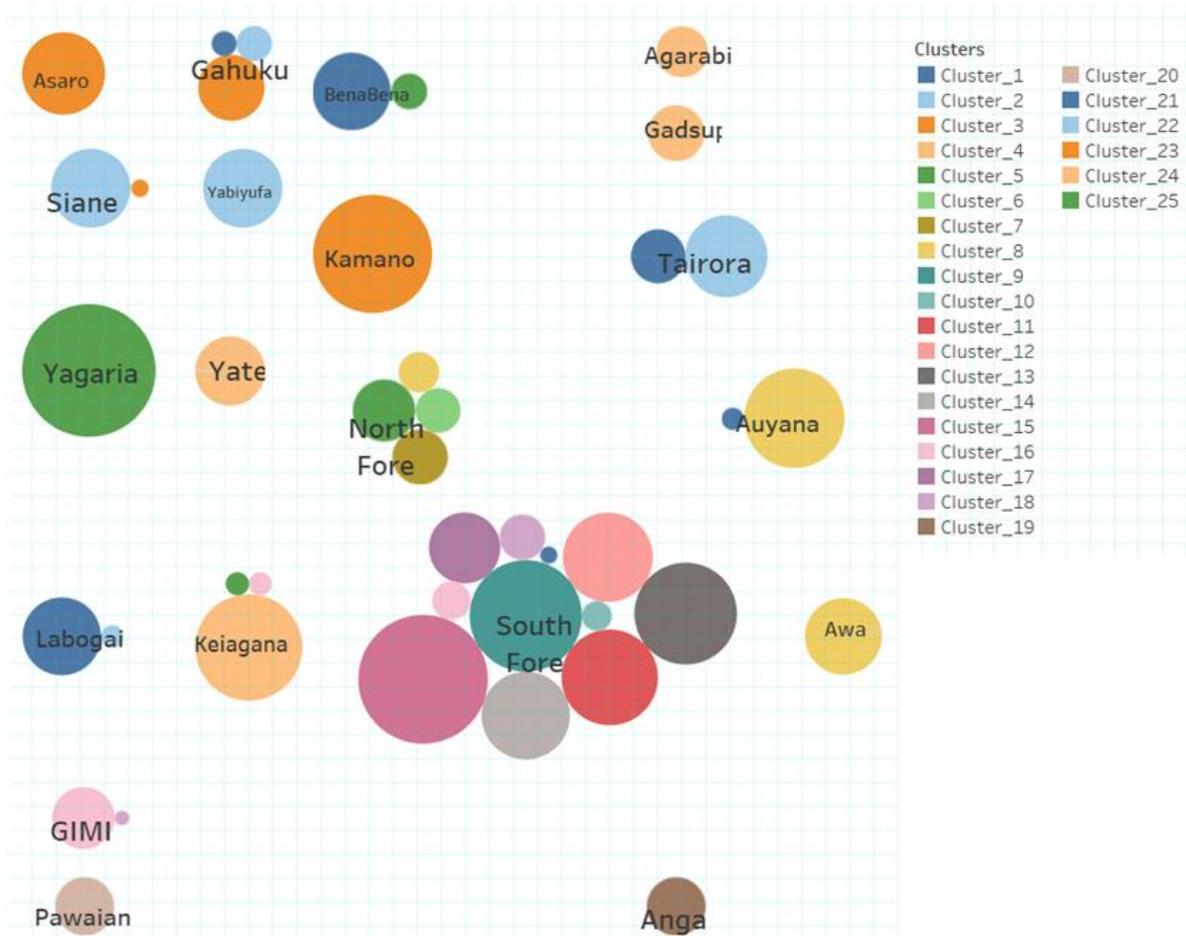


PCA plot (above) of 20 villages within the South Fore, North Fore and Gimi linguistic groups. The map (below) shows the geographical spacing of villages in this region from a previous publication.

Appendix 10 – Wurm Classification of EHPNG languages

TABLE 1		
LANGUAGES* OF THE EASTERN HIGHLANDS† OF NEW GUINEA		
Family and Subfamily	Language and Dialect	Population 1969
Languages of the East New Guinea Highlands Stock		
Eastern (Gadsup-Auyana-Awa-Tairora):		
Gadsup-Oyana	Gadsup Akuna Tompena Agarabi Oyana Uteia Oiyana Ontenu	9,100 10,500 1,300
Auyana-Usurufa	Auyana Asempa Kawaina Kosena Usurufa	5,200 1,300
Awa	Awa Mobuta Elakia Tauna	1,200
Tairora-Binumarien	Tairora Pinata-Kobonbira- Oraura Abiera Batainabura Baira Owenia-Waisara Kambaira Binumarien	12,200 350 c. 150 c. 150
Waffa‡
East-central (Gende-Siane-Gahuku-Kamano-Fore):		
Gende-Biyom‡
Siane	Siane Yabiyufa	c. 16,000 c. 5,000
Gahuku-Bena Bena	Gahuku Asaro Bena Bena	c. 11,500 c. 12,000 c. 12,300
Kamano-Yagaria-Keiagana	Kamano Yagaria Keiagana Kanite Yate	41,800 19,100 11,000 3,200 2,500
Fore-Gimi	Fore Gimi Genatei	15,100 20,100 500
Central (Hagen-Wahgi-Jimi-Chimbu):§		
Wahgi	Wahgi	c. 10,000#
Chimbu-Chuave	Chimbu (Kuman) Nagane Dom	c. 62,000 c. 500 c. 15,000

Appendix 11 – FS clustering reduced from 105 clusters to 25 using ‘greedy’ approach



Plot showing FS cluster membership of 20 EHPNG linguistic group when the original FS tree containing 105 clusters is collapsed to 25 clusters using a greedy based tree branch merging approach. The size of the circles represent the amount of individuals.

Chapter 3 - Impact of kuru on affected communities in EHPNG

3.1 Introduction

While analysis in the previous chapter centred on defining the population genetic structure in EHPNG, work in this chapter was focused on assessing the genetic impact of kuru on affected communities. Specific hypotheses that were tested in this chapter were;

- Mortality due to kuru resulted in an observable reduction in genetic diversity over the course of the epidemic in the most affected areas.
- Analysis of *PRNP* codon 129 frequencies can provide support to existing hypotheses of past kuru-like epidemics occurring in EHPNG and elsewhere.
- A signal of positive selection caused by carriers of the *PRNP* 127V variant will be detectable in a genome-wide scan of the South Fore population.

More information is presented below that provides further context to the challenges and particular complications of addressing these research questions.

Genetic analysis has previously been used in studies to analyse the impact of various phenomena on patterns of genetic diversity and diversification in human populations(122, 171). The inference of an extensive bottleneck as the result of an out of Africa migration was the first such instance to be observed, and had a profound impact on understanding of human history(172). Understanding of differential patterns of linkage disequilibrium between different populations has helped inform fine mapping of disease related variants (173) and the development of tools that have explained the emergence of novel phenotypes in populations including lactose tolerance (174). Impacts of other events and factors on patterns of human genetic diversity over time have been studied and methodologies developed to investigate the legacies of historical epidemics like smallpox and ‘the Black Death’(175), geographical separation(171), admixture(156), wars and invasions(176). Enquiries into the impact of past events using genetic data have been greatly aided by the emerging field of

ancient DNA which has allowed sampling of genetic diversity at distinct spatial and temporal points in the past that greatly aid inferences into the impact of historical processes on observed genetic diversity and differentiation(92, 93).

The kuru epidemic resulted in close to 3,000 deaths and predominantly affected adult women and children of both sexes (only 2% of cases were adult males at the peak of the epidemic). This was a result of mortuary feast participation rights, where adult men would not consume infectious material(50, 151). During the height of the epidemic observers noted many villages largely devoid of adult women(59). The epidemic led to an escalation in conflicts in the area motivated by sorcery accusations, believed by local communities at the time to be the cause of the epidemic(54). Groups not affected by kuru were known to fear the Fore due to their reputation as sorcerers. Such a devastating epidemic is likely to have had an impact on the established cultural and demographic processes in the region. Whether an acute epidemic like kuru which lasted only a few generations can leave a detectable genetic signature on the affected population was an important research question to investigate.

The kuru epidemic which accounted for 200 deaths per year at the start of surveillance of the epidemic in 1957 was the leading cause of death and may have led to a reduction in genetic diversity in the most affected areas (49). The impact of kuru on genetic diversity subsequently will have reduced due to the steady decline in incidence after the prohibition of mortuary feasts. The increase in mean age of onset will also have resulted in a reduced impact on genetic diversity as more individuals who later died of kuru would have had children compared to individuals earlier in the epidemic who had an earlier mean age of onset. Attempts to understand the incidence and location of kuru prior to the start of official surveillance in 1957 have been based on interviews with local residents who lived in the region at the time. Although likely to be complex and affected by many multiple factors, it has been suggested that kuru incidence prior to 1957 had increased in a symmetrical manner to the decline in cases noted in the 1069's and 70's(48). This means that the

period of relatively high incidence of kuru and when it was exerting most influence on genetic diversity in the region may have been restricted to less than 40 years.

Detecting any change in population genetic parameters of kuru affected populations would require disentanglement of changes caused as a result of kuru from other events that may have also caused population demographic change. Distinguishing between the impacts of some of these events is challenging. The impact of European contact and colonial administration, for example, largely overlapped with the course of the kuru epidemic but may have had contrasting effects on levels of genetic diversity in the region(54, 105). Other historical changes in the region further back in time will have made an impact on the genetic properties of the region. Demographic and accompanying genetic change may have occurred over time due to climactic changes, increasing in sea levels, volcanic eruptions and the introduction of alien species into this environment.

Kuru has previously been shown to have impacted variation in the known disease related locus for prion diseases, *PRNP*(30, 177). Individuals who appear to have displayed resistance against developing kuru despite extensive exposure tend to be enriched for heterozygous genotypes at *PRNP* codon 129(69). It has previously been hypothesised that elevated minor allele frequency at *PRNP* codon 129 in EHPNG as a whole and other populations with known historical practices of anthropophagy are evidence of historical kuru like epidemics(178, 179). Other investigators have argued that such fluctuations in the minor allele frequency of this locus could be completely explained by drift(180). An updated analysis of this locus from presently available data from multiple global populations would add further light to these arguments. Additionally, a novel variant found only in kuru affected areas has revealed an additional response at the *PRNP* locus(30).

The discovery of a novel polymorphism at *PRNP* codon 127, encoding the amino acid valine instead of the usual glycine, was shown to be a powerful and unique example of a natural selection response of a population to an epidemic(30). The basis of the discovery of this variant was to perform in depth analysis of variants in the *PRNP* locus, a gene which has variants linked with predisposition to prion

diseases in other populations. Investigating whether this variant would have been discovered without this a priori knowledge of this locus in a genome-wide analysis of the populations affected by kuru would illuminate understanding of what conditions permit perceptible changes in genetic variation after exposure to selection pressures over a relatively short period of time. This understanding would also help tailor approaches to detecting kuru related variants outside of *PRNP* and separating them from candidate variants that have fluctuated in frequency largely due to other forces including genetic drift.

The extent of sampling in the *PNG merge* data set in a relatively small geographical area affected by a deadly and acute epidemic allows examination of the genetic legacies of such events. The abundance of scholarship investigating the impact of long term selection pressures over hundreds of generations including pathogen resistance and climactic conditions has been well established(181). It is likely in human history that shorter term selection pressures similar to kuru in duration and severity would have also played a role in genetic variation in and between populations that is observed today.

3.2 Materials and methods

3.2.1 Assessment of Impact of kuru on EHPNG

In previous studies the effect of present genetic diversity on the spread of disease have been considered, with isolated populations being more prone to disease spread due to a monoculture effect(182). This may well be a factor worth investigating with populations like those in EHPNG going forward due to the extensive genetic drift, and runs of homozygosity (ROH) detailed in the previous chapter. In this chapter a different approach was taken given opportunities that the available dataset provided. With 334 samples from individuals from the most affected South Fore linguistic group with accompanying dates of birth it was hoped any changes in genetic diversity over the period of dates of births which overlapped greatly with the majority of the kuru epidemic would permit a novel insight into time-dependent changes in genetic diversity.

Kuru surveillance revealed approximately 150 deaths per year at the height of the epidemic in the South Fore linguistic group, a population with a census size of ~8,000. One may expect an observable drop in genetic diversity as a result of this level of incidence. This may have been exacerbated with the epidemic occurring in the remote margins of EHPNG, in a region largely isolated from the rest of the world. Accounts of the fearsome reputation of the Fore as sorcerers and the general belief during the epidemic that sorcery was the root cause of the epidemic may also have compounded any reduction in diversity with individuals from outside the region reluctant to enter the region to live, intermarry and have children as the epidemic intensified. Kuru surveillance commenced in 1957 at the height of the epidemic. Ethnographic interviews and investigation have led investigators to believe that the first case of kuru occurred close to the turn of the century in the village of Uwami (See figure 1.3). Knowledge regarding incidence of kuru prior to surveillance in the decades prior has been limited to interviews with elderly people who were alive at the time. An analysis of genetic diversity was hoped to provide additional information regarding incidence of kuru in the decades prior to surveillance.

Dates of birth for individuals in the dataset born prior the full establishment of colonial rule and then independence with recordings of births and deaths were obtained through estimates after interviews with participants. These estimates were informed by making references to well-known events in the region including the arrival of westerners in the 1940's and the subsequent construction of infrastructure including roads that permitted vehicle access to remote regions. These estimates cannot be considered fully accurate, in order to ameliorate any inaccuracies estimates of genetic diversity through time-points in this analysis have been based on a sliding window approach which will hopefully minimise the effect of falsely classifying individuals.

CP analysis was performed using 334 South Fore donor individuals after controlling for relatedness ($PIHAT < 0.1875$ in plink) from the South Fore region with years of births ranging from 1920 to 1988. Restricting testing for genetic diversity changes due to kuru to the South Fore linguistic group was

decided due to the robustness of linguistic group membership as a descriptor of population structure in the previous chapter. Additionally it has been shown through ethnographic and epidemiological investigations that the kuru epidemic moved through the region throughout the epidemic with different linguistic groups having different incidence profiles(48). It was hoped that restriction to the South Fore would reduce the impact of this heterogeneity and permit the observation of the impact of an epidemic on a specific population over several decades. To gauge diversity within the South Fore region all other EHPNG linguistic groups were used to create a panel of CP donors comprising 485 individuals from 18 EHPNG linguistic groups. The resulting CP matrix was scaled and then pairwise genetic distances between individuals were calculated between the 334 South Fore recipients using the Total Variation Distance measure (discussed in the previous chapter). Measures for level of genetic diversity for each year were then created based on the average pairwise TVD measures for individuals born within 5 years of each year.

The pairwise nature of the 161,900 TVD observations and the sliding window approach taken in the analysis meant that any linear trend in genetic diversity over the time period could not be considered as independent and parametric assumptions of the linear modelling of genetic diversity over time were violated. In order to gain some confidence over a trend in diversity observed over the period dates of birth were randomly imputed for individuals in the analysis 5,000 times. The observed coefficient of the regression analysis was compared to the distribution of coefficients in the 5,000 analyses when years of birth were randomly assigned to derive an empirical p-value.

Simpler categorical analyses of measures of genetic diversity were also performed. Individuals born prior to the believed peak of the kuru epidemic (before 1940) and after (1980) were compared for average measures of Identity by descent (IBD) using the `--genome` function in plink 1.9. ROH were also compared using plink1.9 (`--homozyg`) for the same individuals to check consistency and provide assurances over any interpretations made subsequently. Sources for data used in analyses in this chapter can be found at Appendix 1.

3.2.2 Analysis of *PRNP* codon 129

In a previous study 1,004 samples from the EHPNG dataset used in this thesis had *PRNP* gene sequencing performed(30), from this *PRNP* codon 127 and 129 genotypes were ascertained for individuals from 15 linguistic groups in EHPNG. In this study we have estimated an additional 392 samples including individuals from an additional 6 linguistic groups in the region. This has been done through imputation of the entire dataset using phase 3 data from the 1000 genomes project as an imputation panel using the LDAK tool(183).

Separate imputations were performed for data genotyped in the Illumina Omni Express and Illumina 670 platforms. This resulted in 1,003 individuals and 5,738,247 variants in the imputed Illumina Omni Express data and 488 individuals and 5,644,637 variants from the Illumina 670 platform. After assessment of info scores it was decided to remove variants with an info score below 0.9 and not to include variants with an allele frequency below 0.01 in the merged data (Figure 3.1). It was decided to not use the imputed data in analysis of linguistic groups outside of the kuru region where imputation accuracy would likely be further compromised due to fewer samples being available and the genetic divergence of these groups shown in the previous chapter. A further check was performed on samples from 82 individuals who were genotyped on both platforms with assessment of mismatches of data. For the 82 samples the imputation accuracy was >98.5%. The merged data resulted in an intersection of 1,892,020 variants and 1,417 individuals in the *Imputed Merger*.

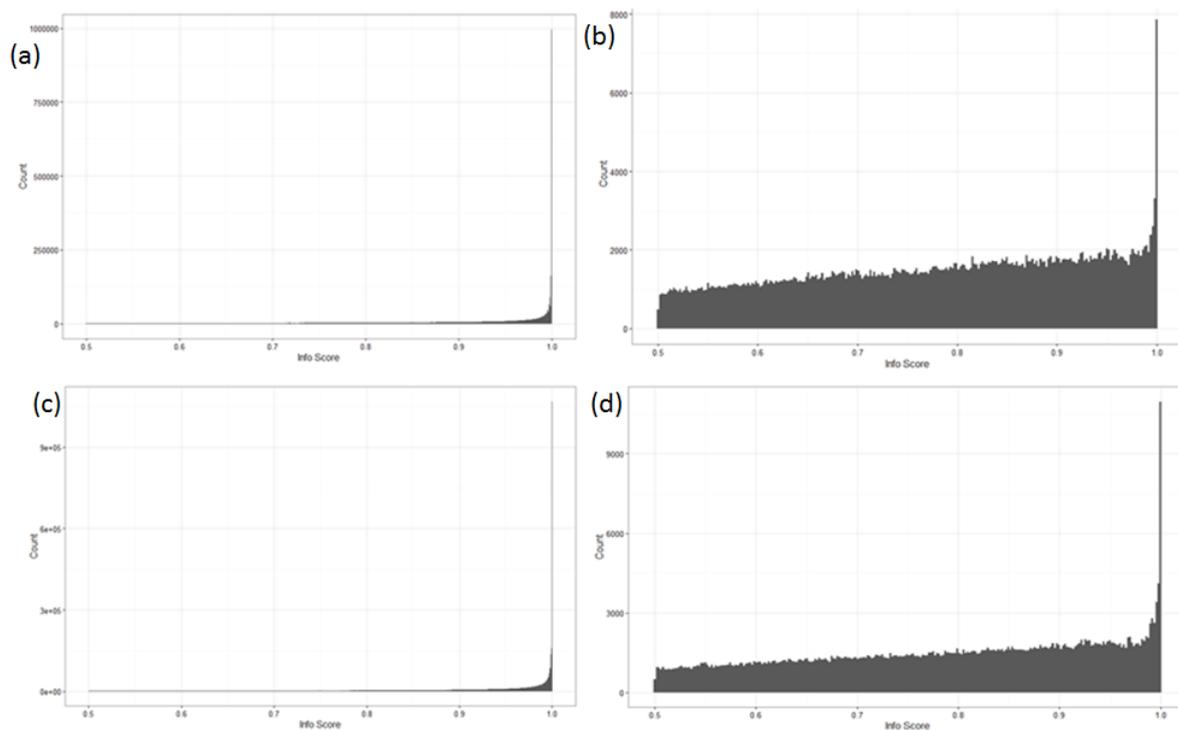


Figure 3.1. (a) Histogram of info scores of all variants after imputation of 488 individuals from the Illumina 670 genotyping array. (b) Results from the same imputation as (a) but filtered to show only variants with minor allele frequency < 1%. (c) Identical breakdown for 1,003 individuals imputed from the Illumina Omni Express genotyping platform

The info score obtained for *PRNP* codon 129 (rs1799990) was 0.94. Further clarification of the accuracy of this imputation was obtained by comparison of the previously sequenced 1,093 individuals with their imputed data. There were five mismatches in total meaning accurate genotypes were obtained for 99.54% of the dataset assuming laboratory sequencing was 100% accurate. As part of the analysis of genetic impacts of kuru during the epidemic, *PRNP* codon 129 valine frequencies were plotted per decade and the proportion of observed heterozygous genotypes as this locus were compared to expectation as per Hardy-Weinberg equilibrium via a Chi squared test. This was done for the most heavily affected South Fore linguistic group.

PRNP codon 129 genotypes were obtained from other global populations through publicly available sequencing projects. These include GnomAD exome sequencing project which contains data from over 125,000 individuals globally. These are grouped into major continental groupings (See appendix 2). *PRNP* genotypes were obtained for ancient DNA samples from a database of 425 individuals from

the Eurasian landmass after request from colleagues working on European ancient DNA(184). Date estimates were available for 346 of these samples (36,730- 375 years old). After excluding single variant calls 126 samples were available (see appendix 1).

3.2.3 Understanding of *PRNP* 127V haplotype in kuru affected communities

Previous publication found the *PRNP* 127 allele to be carried by a minority of individuals (prediction of allele frequency was at < 5%) in the kuru affected region and shown to be under positive selection, providing protection against all PD's(30, 71). The allele was felt to be of recent origin due to the extended identity of microsatellite markers over an extended region and its geographical restriction to the Purosa valley within the South Fore linguistic group (there has since been the identification of the *PRNP* 12V in a single European individual although this is felt to be independent of the allele in Papua New Guinea). This variant is an example of natural selection in modern human populations and a better understanding of its structure in terms of linkage disequilibrium patterns and haplotype distribution will be informative of how natural selection progresses in populations and also inform analysis in attempts to find further variants that may have been under selection pressure during the kuru epidemic (see following chapter).

A series of tests were performed with the tool Relate to gauge if the region of extended IBD shared between carriers of *PRNP* 127V due to its recent origin would result in a signal in a mixed population. The relate program has been used to find variants with phenotypic effects for traits including Breast cancer in European populations and Type 2 diabetes in the Greenlandic Inuit(185). The first test was to test this in the entire South Fore population and then if unable to detect a signal to find at what frequency was required for a signal to emerge. This was hoped to provide an expectation of the power of the Relate tool to find additional variants with similar effect and evolutionary history as *PRNP* 127V in selection analyses. The *PRNP* 127 allele is carried by 44 individuals in the dataset (all heterozygous carriers). This means that the maximum allele frequency permitted in these analyses is 50%. Quality control procedures were performed in plink 1.9 to give individuals lacking in IBD2

relatedness which is known to affect IBD1 estimations and data was LD pruned with a maximum R^2 of 0.4 permitted between variants in the analyses.

A first step was performed to assess pairwise IBD per each individuals in the analysis using the relateHMM program that produces a matrix of pairwise IBD. Testing of elevated IBD between *PRNP* 127V carriers was then carried out using the HMMtest function in which 10,000 permutations were performed with the imputation of the test population identifiers (*PRNP* 127V carriers and non-carriers) with a control population.

Analysis was then performed using a tool under development called 'Chromomatcher'. This program also has been designed to find areas of long matching between phased haploids of individuals as a result of recent selection, though while simultaneously identifying clusters of haploids who share long matches. At each SNP, Chromomatcher first finds the centimorgan length of matching Identical-by-state (IBS) spanning the SNP for each pair of haploids. Next, at this SNP haploids are merged (clustered) together in a hierarchical manner under a greedy approach. In particular the two haploids with the longest match are merged, replacing them with a new merged "haploid" whose matching to each other haploid is set to the average of that of the two merged haploids. These merging steps continue until the maximum matching length between any two "haploids" falls below a pre-defined cM threshold. Any haploids that merged with each other during this procedure are considered clusters at this SNP. Haploids from clusters containing $\leq XX$ individuals, and haploids that did not merge at the SNP, are put into a "null" cluster. Each cluster containing $> XX$ individuals is scored as the average pairwise matching at the SNP among all haploids within that cluster divided by the average pairwise matching at the SNP among all haploids in the null cluster (ignoring matches of length 0, i.e. where the allele differed at the SNP). This approach is somewhat similar to that in the commonly-used iHS and XP-EHH scores, though it separates haploids into groups rather than using pre-defined labels.

The matrix of matching formed by this process is then decomposed through a clustering process to find subpopulations of haplotypes with extremely high matching compared to the population averages. This was hoped to be beneficial in application to recent selection as variants under selection may not in all cases have reached elevated frequencies or fixation due to the episode of selection being ongoing or not having lasted long enough to have resulted in high frequencies or fixation. The clustering technique is also ideal in scenarios where unknown natural subpopulations may exist. This may be due to selection pressure acting in a geographical subsection of a population due to a variants localisation to a particular area.

Ascertainment of long haplotypes through a simple counting of SNP identity between pairs of individuals in a population can lead to the appearance of false positive signals of selection. In populations long regions of matching between individuals will naturally arise due to sharing stretches of DNA due to sharing a recent common ancestor. This is likely to be more prevalent in populations such as those in EHPNG where extended genetic drift is present, consanguineous marriage and a lack of immigration and population mixing compared to continental populations. Such signals driven by these 'demographically only' derived haplotypes will be similar to those also influenced by natural selection. It is hoped these sources of false positive signals could be filtered through the use of accompanying phenotypic data. In the case of our analysis of kuru, an excess of individuals in a high scoring cluster who died of kuru resulted in the elimination of that haplotype as a possible site of recent selection due to the kuru (see chapter four for more in depth discussion).

It has also been documented that certain areas of the genome are more prone to recombination events than others and referred to as 'recombination hotspots'. This could result in an enrichment of long haplotypes in the analysis from certain regions of the genome and may mask genuine signals of selection. There is also inconsistency in any analysis using genome wide genotyping arrays in terms of the density of variants in genomic regions. Mismatches in actual DNA sequences may be missed in one region and found in another due to a greater number of SNP's.

To control for these sources of error, inconsistencies and false positive signals, high scoring clusters that were originally based on average matching of individuals within that cluster would instead be scored by a mean ratio. This ratio is comprised of the average matching of individuals within a cluster divided by the average matching of all other individuals in the analysis. Localised factors including SNP density and recombination rate should be reduced by comparing to other individuals outside of the cluster who should on average be affected by such factors similarly.

To further focus the analysis on recently selected novel variants Chromomatcher parameters were adjusted to restrict scoring to clusters containing a minimum amount of individuals and minimum average matching distance. In the analysis performed of the South Fore population minimum average matching length was set at the top two percent of matching lengths per SNP and the minimum amount of individual haplotypes was set to 10. After quality control procedures were performed in plink (--genome PIHAT < 0.1875 of the South Fore population) this resulted in 192 individuals and 384 haplotypes. This would allow any high scoring haplotypes present at 2.6% frequency in the population to be scored.

3.3 Results

3.3.1 Analysis of genome-wide signatures of population response to kuru

Analysis of ROH in the pre and post-peak kuru Fore population shows no difference in average number long segments (paired t-test, $p = 0.20$) between any of the time points used in the analysis.

Analysis of pairwise TVD between individuals by birth decade using a sliding window approach (Figure 3.2) shows a decline in genetic diversity as measured by TVD (lm gradient coefficient = 0.00076). This trend could be explained by the ongoing kuru epidemic during the period negatively impacting genetic diversity

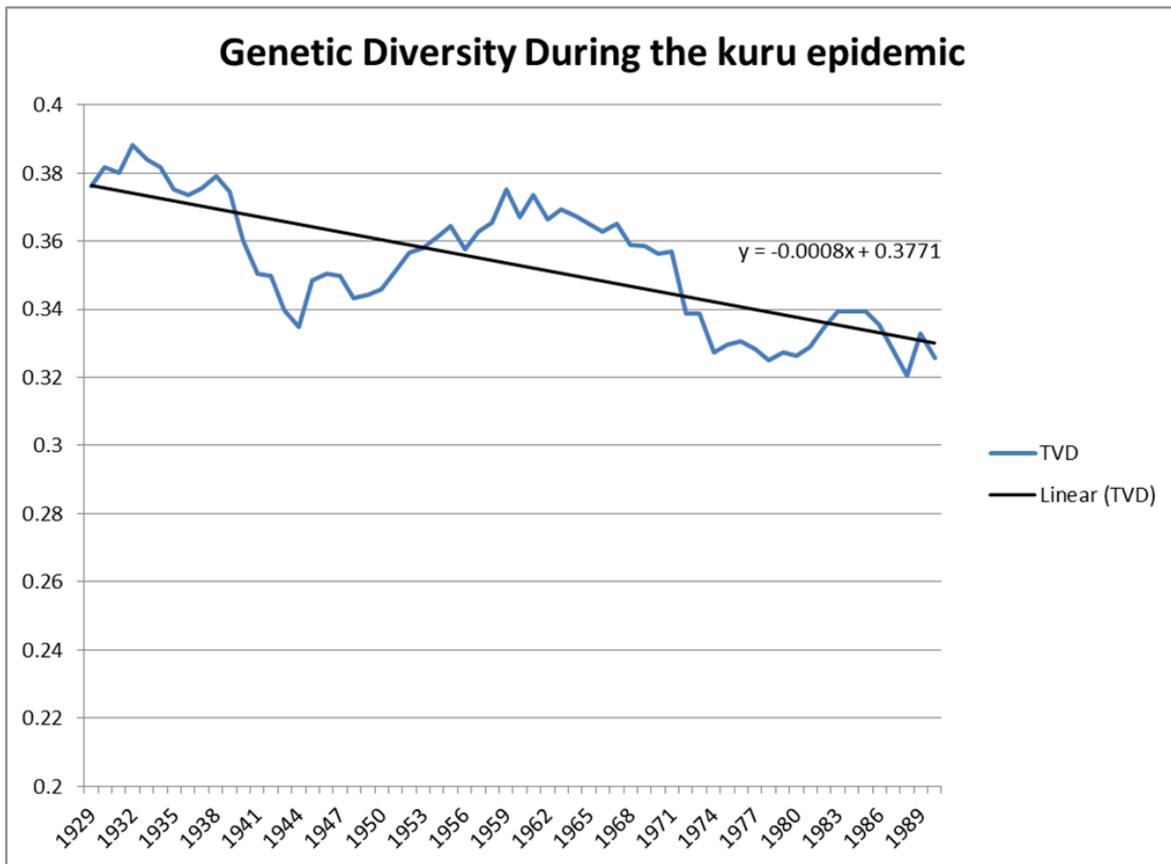


Figure 3.1. Analysis of genetic diversity measured during 62 years of the 20th century in the South Fore linguistic group. A sliding window approach was used with measurements for each year in the analysis being comprised of pairwise TVD for all individuals born within five years of that date. Analysis with a simple linear model shows a reduction in genetic diversity across the period.

Results from permutation of years of birth randomly to individuals in the analysis show that the reduction of diversity over the period (gradient = -0.00076) results in a statistically significant empirical p-value ($p = 0.035$)(Figure 3.3 shows the full distribution of linear model gradient values for all 4,000 permutations).

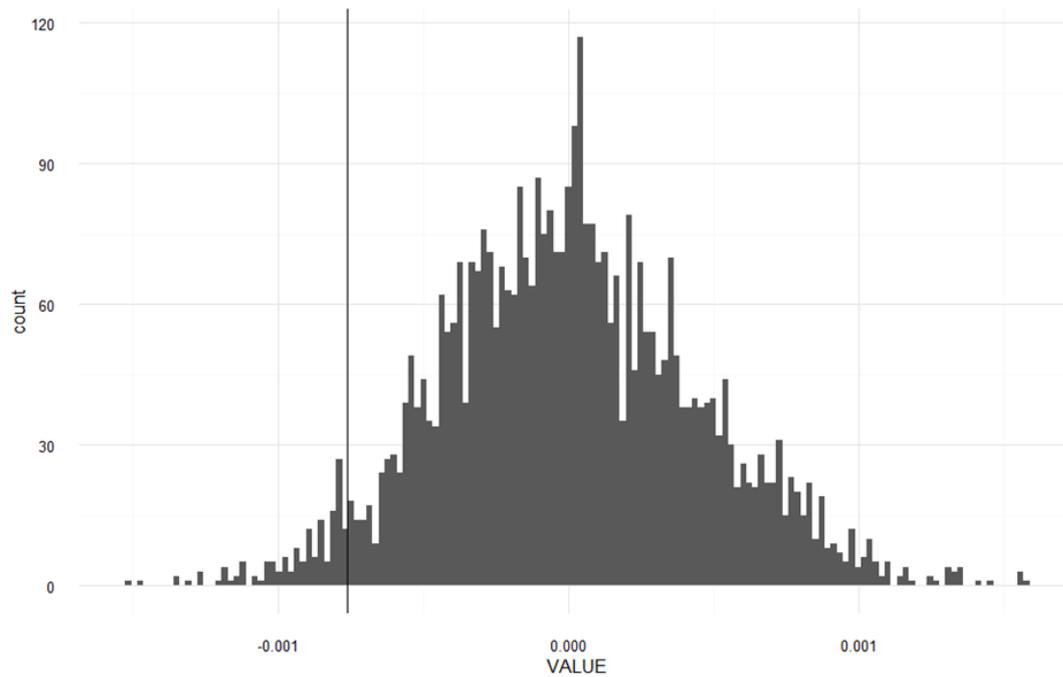
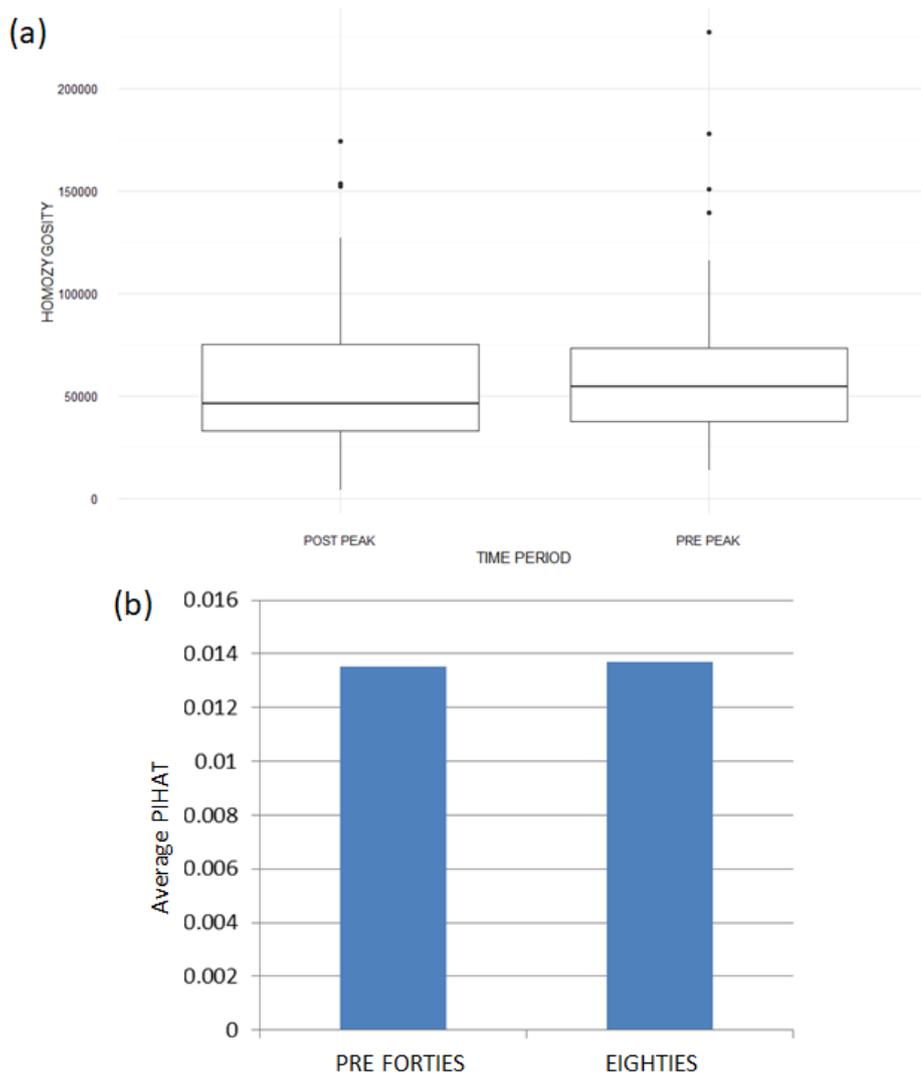


Figure 3.2. Histogram of gradient values for 4,000 repeated sliding window analyses of pairwise TVD as a function of time in the South Fore population through a sliding window approach. The vertical line shows the value of the negative gradient (reflecting a loss of genetic diversity over the period) . An empirical p-value for this observed gradient in this distribution of simulated p-values is 0.0348.

Analysis of homozygosity and pairwise identity by state of individuals before and after the peak of the kuru epidemic (Figure 3.4) do not reveal any overall changes in these parameters which gauge genetic diversity within populations ($p > 0.05$ two sample t-test).



Boxplot (a) of average lengths of homozygosity between individuals in two cohorts representing years of birth before the peak of kuru (Pre 1940s) and after the peak of the epidemic (1980's). Homozygosity is the kilobase amount of homozygosity per individuals contained within runs of homozygosity which were set as a minimum of 1 megabase. Bar plot (b) showing the average amount of relatedness in the same year of birth cohorts. This was measured through pairwise PIHAT values between individuals which is analogous to Identity by State. No statistically significant differences were found between the two cohorts in either analysis.

3.3.2 Analysis of *PRNP* codon 129 frequencies

Appendix 2 shows allele frequencies of *PRNP* codon 129 taken from the GnomAD exome sequencing data base. *PRNP* codon 129 minor allele frequencies range from 0.02 (East Asian) to 0.41 (American Latino). Figure 6 shows genotypes obtained from 126 Eurasian ancient DNA samples, with an overall average minor allele frequency of 0.33. They have been plotted on a map in Figure 3.5. Read depth

for each of the samples varies greatly giving uncertainty in some of the genotype calls. The *PRNP* codon 129 derived allele is found in the Neanderthal genome via the Neanderthal genome browser(186).

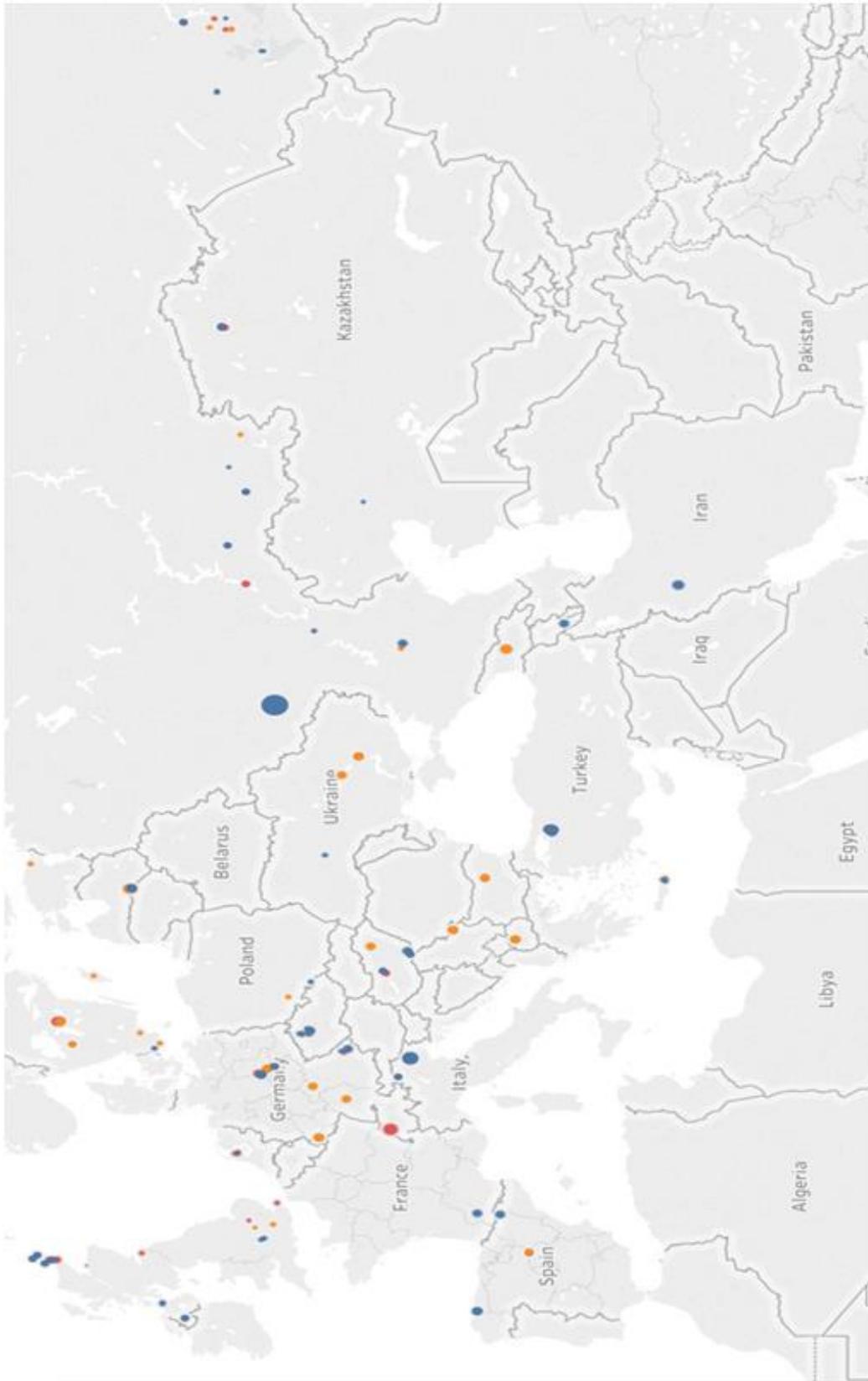


Figure 3.4. PRNP codon 129 guanine frequencies per ancient DNA samples in Europe. Colours represent genotype for homozygous ancestral (blue), homozygous derived (red) and heterozygous (orange). Size of the points reflects the age of the sample. See appendix for precise dates and locations.

PRNP codon 129 average minor allele frequencies in EHPNG inferred from imputation of 1,496 samples is 0.56. In the 11 kuru affected linguistic groups the average 129V frequency is 0.55 and in the non-affected linguistic groups it is 0.57. In the most heavily kuru affected South Fore linguistic group valine frequency is 0.54 (Figure 3.6).

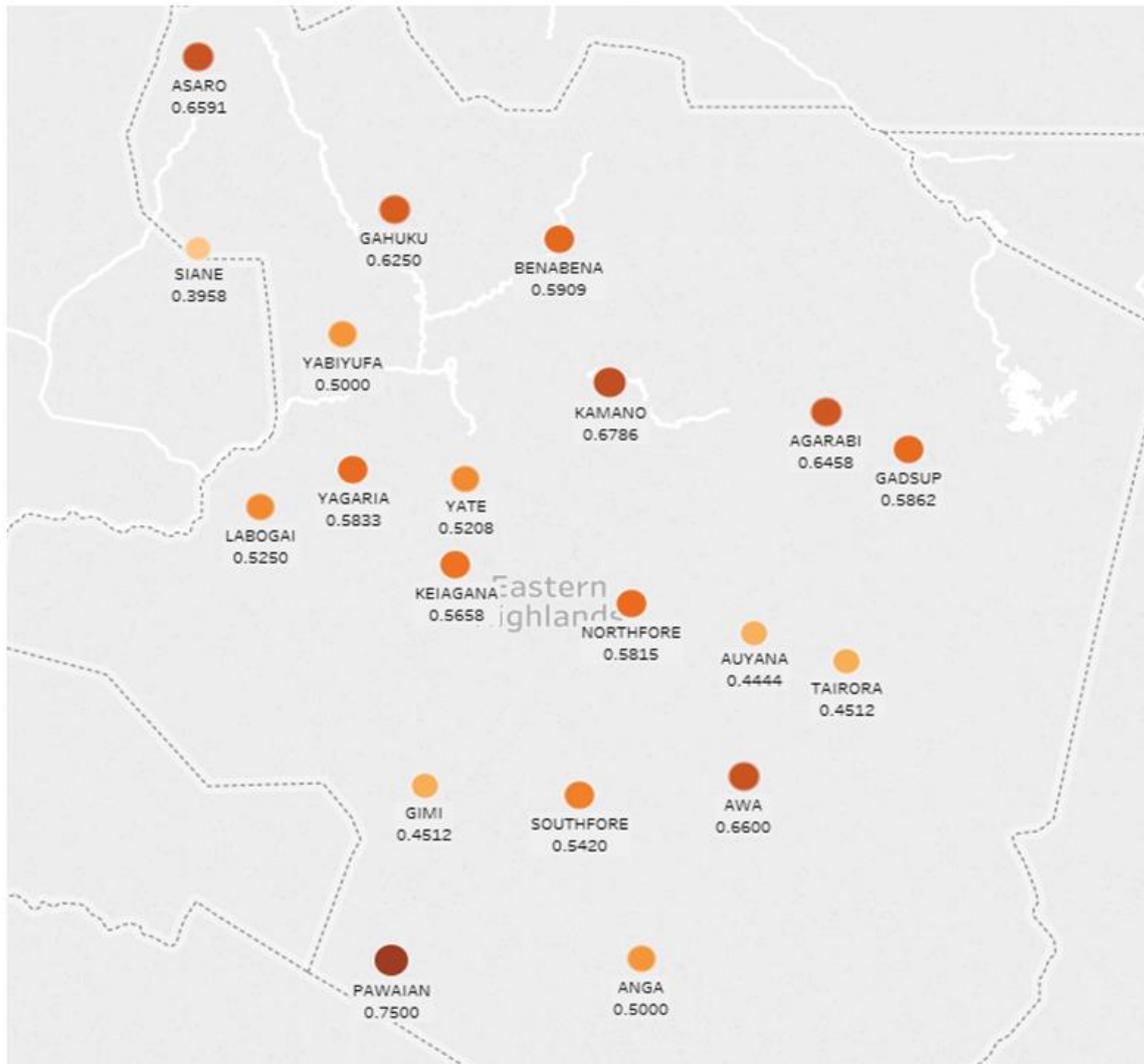


Figure 3.5. *PRNP* codon 129 guanine frequencies per EHPNG linguistic groups based on imputation of EHPNG data. There does not appear to be any pattern in frequencies based on kuru exposure. The overall frequency in the region (0.55) is the highest observed in any global region. The extremely high value makes inferences in relation to balancing selection at this locus difficult.

Proportion of *PRNP* codon 129 heterozygote genotypes over time in the South Fore show enrichment in individuals born in the 1940's in excess of what would be predicted from Hardy-Weinberg Equilibrium expectations ($p = 0.000029$, Chi squared test). When stratified by sex this

difference is apparent in females born in both the 1930's and 1940's ($p = 0.05$ and 0.000018 , Chi squared test).

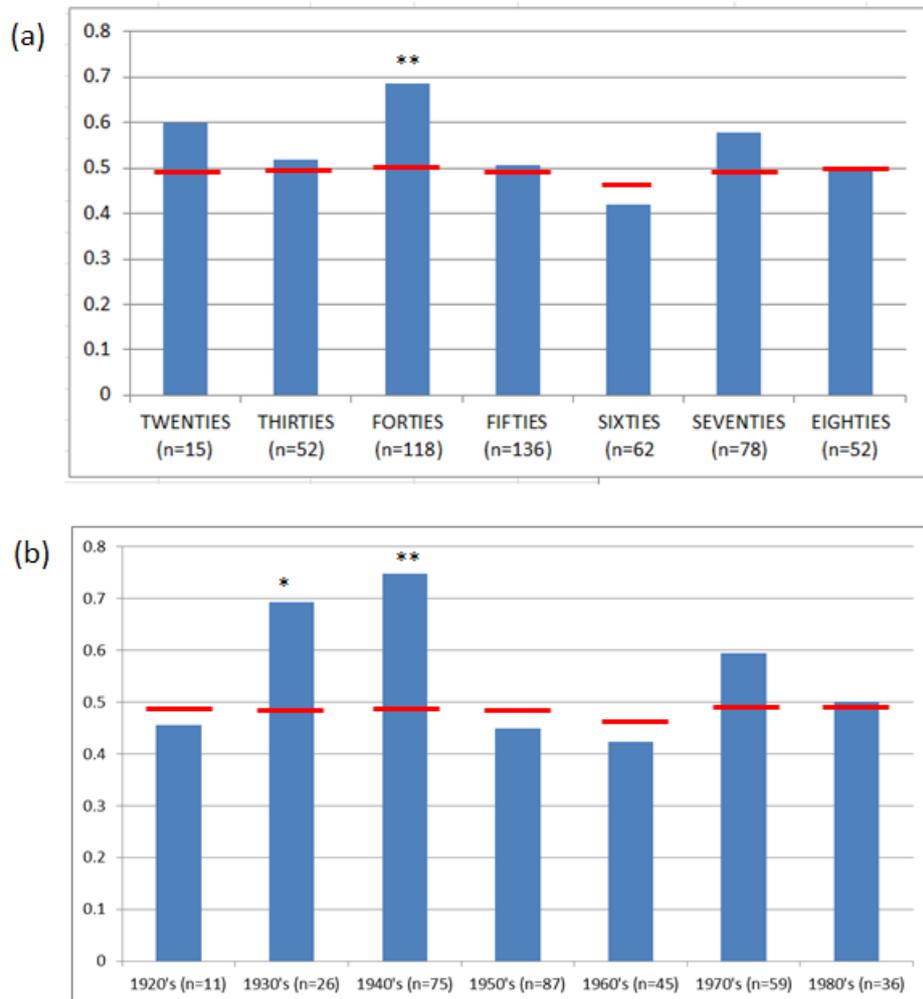


Figure 3.6. Bar plots showing the proportion of *PRNP* codon 129 heterozygote genotypes in the South Fore linguistic group in different decade bins. The bar plot of both sexes (a) shows a distinct deviation from Hardy-Weinberg equilibrium in the 1940's ($p < 0.0001$) and the plot for women only (b) shows a deviation in both the 1930's ($p < 0.05$) and 1940's ($p < 0.0001$). Red lines are proportion of expected heterozygote genotypes.

3.3.3 *PRNP* 127V variant does not provide striking signature of selection in genome-wide scan of kuru-affected population

Analysis of 192 individuals from the South Fore population using the Relate program did not reveal any genome wide significant regions of shared IBD throughout the genome including the *PRNP* locus, that may have been driven by *PRNP* 127V carriers.

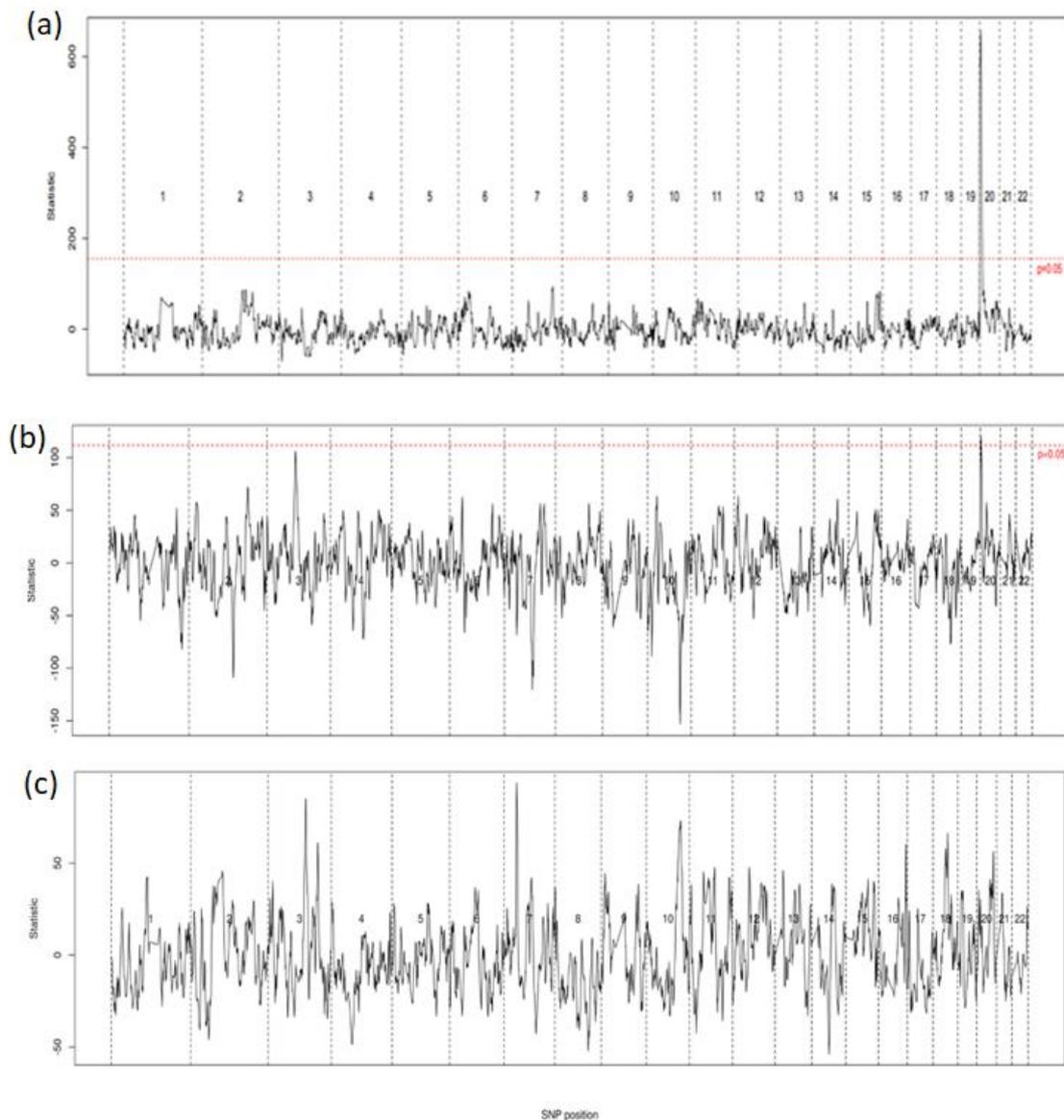


Figure 3.7. (a) Relate Analysis of 33 *PRNP* 127V carriers when being used as a test population, frequency =50%. Similar analysis (b) when test population comprises 33 *PRNP* 127V carriers with 103 other South Fore individuals, frequency =16.5%. (c) Analysis when 33 *PRNP* 17 carriers are in a test population with 140 other South Fore individuals, frequency =9.5%.

A Subsequent analysis of 33 127V carriers as a test population show an elevated signal at the *PRNP* locus. This score is reduced when more non-carriers are added until the point where the signal is below genome-wide significance (Figure 3.8).

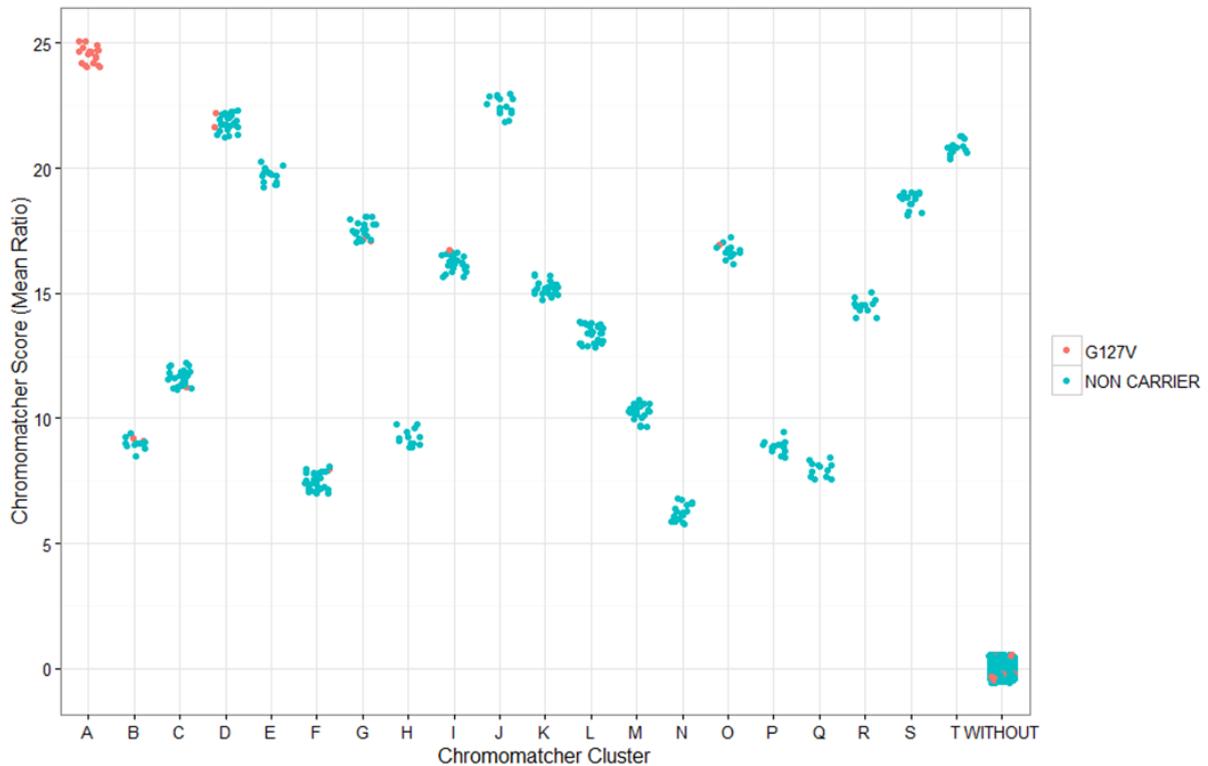


Figure 3.8. Plot showing the distribution of South Fore individuals when clustered by Chromomatcher for shared long regions of matching at rs6052787 the variant closest to *PRNP*. When default parameters of Chromomatcher are set the highest scoring cluster at this locus is comprised completely of *PRNP* 127V carriers. Not all *PRNP* 127V alleles are captured in this cluster and the remainder are dispersed amongst other clusters or not assigned to a cluster. Individuals belonging to the same cluster have identical mean ratio scores, noise has been added to permit visualisation of cluster placement of each individual.

The clustering profile for SNP rs6052787 shows the ability of Chromomatcher to cluster *PRNP* 127V carriers in an analysis with 340 non carrier haplotypes (Figure 3.9). This clustering is not completely faithful with only 18 of the total 44 being found in cluster A, the highest scoring cluster at this locus. Some other carriers are dispersed among other clusters and 16 individuals not forming scoring clusters.

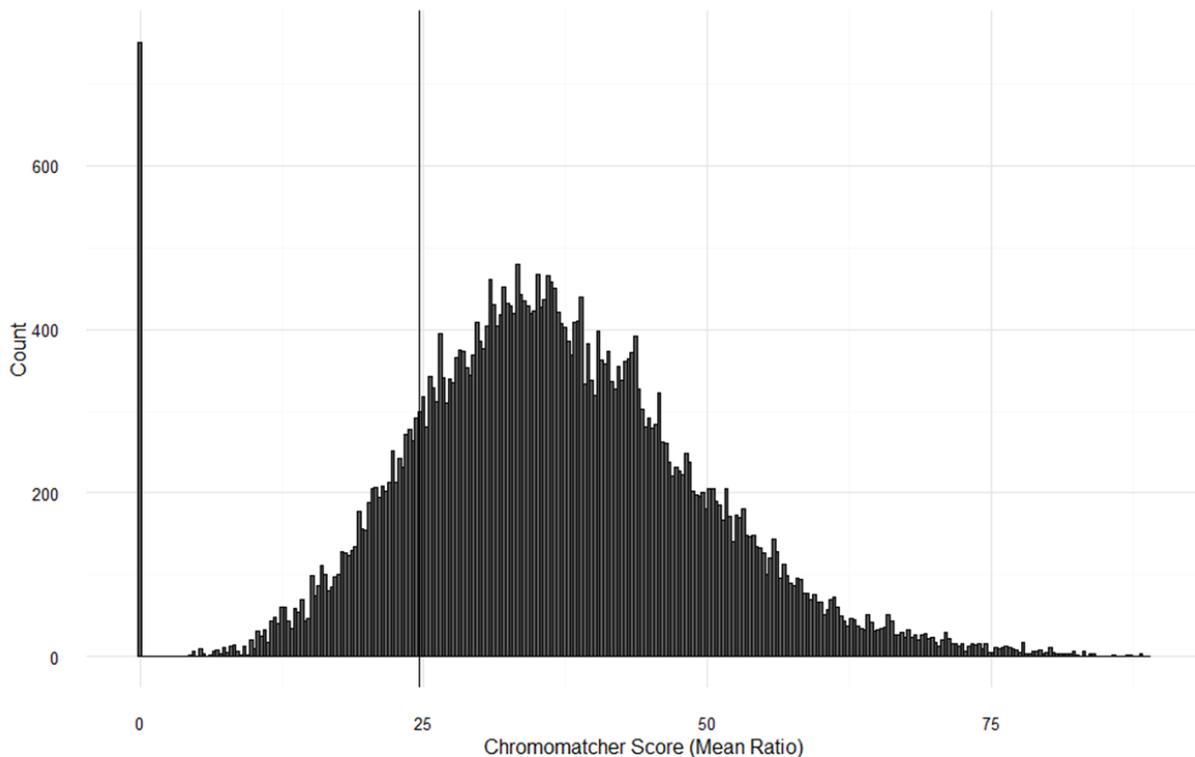


Figure 3.9. Histogram of Chromomatcher mean ratio squared scores genome-wide in an analysis of 383 individuals in the South Fore linguistic group. The vertical line represents the highest scoring cluster at rs6052787, the variant closest to the *PRNP* gene.

This high scoring cluster (mean ratio 24.7) does not rank highly genome wide (Figure 10). This ultimately shows that in an analysis of the South Fore population using the Chromomatcher tool would not have uncovered the *PRNP* 127V variant through this approach.

3.4 Discussion

3.4.1 Change in diversity

CP analysis of data from individuals from the South Fore linguistic group shows a trend in reduction of genetic diversity over the 61 year period that data is available for. A categorical analysis of relatedness and levels of homozygosity does not find any significant change in genetic diversity before and after the peak of the epidemic.

The lack of observed change in relatedness and homozygosity summary statistics may be due to the lack of sensitivity in these measures compared to the sliding window pairwise TVD approach using CP. The heterozygosity measure and relatedness measure were based on applications of plink 1.9 to

genotyped array data. CP successfully described subtle differences in population structure at the linguistic group level in the previous chapter. This haplotype based approach was chosen as a means to limit the effects of ascertainment bias of the genotype arrays which resulted in a very low number of polymorphic variants when EHPNG populations were genotyped. Using phased haplotype data can help mitigate this reduction in resolution by using differences in linkage disequilibrium to construct patterns of haplotype sharing that carry large amounts of genealogical information.

An overall trend in genetic diversity is explained by the hypothesised effects of the kuru epidemic which at its peak resulted in the deaths annually of 150 people in the South Fore, a group with a standing population of approximately 8,000(49, 100). This conceptually may have resulted in a bottleneck effect with many genetic variants in the population being lost due to the levels of mortality as a result of kuru.

New genetic diversity entering the region in the form of migrants from other linguistic groups may not have been sufficient to counteract the effects of kuru mortality and inwards migration may have been impeded through widespread fears throughout EHPNG of the Fore due to their widespread reputation as sorcerers.

Findings of changes in genetic diversity derived in this study have ultimate limitations due to the sample set the analysis was based upon and the near-impossibility of replicating such a finding independently. It will be extremely difficult to obtain a data source with individuals from the kuru affected region born at different time periods of the 20th century.

An assumption in this analysis was to view the South Fore as a population that experienced the kuru epidemic consistently throughout the linguistic region at any single time. This was motivated by findings in the previous chapter that showed the robustness of linguistic groups as population descriptors and further decomposition of groups at the clan and village level resulting in a loss of resolution. Ethnographic research has revealed that the kuru epidemic had a dynamic course,

arriving at certain parts of the South Fore region many years later than others(48, 51). This difference would have resulted in differences in exposure and incidence of kuru which have been previously noted (48). A fine grained analysis at the level of a single village would not likely provide useful results due to the drop in sample size and the lack of statistical power to observe significant changes in diversity measures.

An assumption that levels of genetic diversity were stable and constant prior to the arrival of kuru in the region would not reflect the reality of a population likely to be under multiple demographic pressures. As discussed in the previous chapter scholars have noted the likelihood that the entire highlands region was undergoing a dramatic change demographically due to the arrival of the sweet potato in recent centuries (See section 1.4.2 for more information)(103, 112). There are likely multiple other factors that have not been documented by scholars, compounded by an absence of written history in the region prior to European contact(54, 105). Making inferences of kuru stimulated changes in genetic diversity in populations undergoing multiple other contemporaneous changes is challenging and limits are placed on the scope of any interpretations. Investigating possible changes and pulses in kuru incidence prior to kuru incidence is limited ultimately due to the lack of samples and the absence of possible replicability to support any findings.

With a more fine-grained analysis using WGS data accompanied with DNA extraction from blood samples from a non-kuru-affected group, claims of changes of genetic diversity due to kuru could be more strongly argued. Limitations would still remain as using a non-kuru affected group would not guarantee elimination of other factors effecting change exclusively in the South Fore. As noted in the previous chapter groups in EHPNG are extremely heterogeneous not only in terms of their genetic properties but also linguistically, demographically and culturally.

Analysis of sex-biased changes in genetic diversity would allow testing of a hypotheses if the observed sex-bias in incidence resulted in a detectable genetic signature. This could be done by a similar analysis with X chromosome data only. An expectation of a more drastic reduction in genetic

diversity would be expected, although the reduced amount of genetic data on the X chromosome compared to the autosomes may reduce sensitivity to detect such changes.

3.4.2 *PRNP* codon 129 minor allele frequency study

Analyses of allele frequencies of *PRNP* codon 129 show it to have been highly polymorphic through time and location. It has previously been hypothesised that differences in minor allele frequency in different global populations could be the result of historical kuru-like epidemics in EHPNG and elsewhere resulting in balancing selection at this locus(178). This theory has been challenged with it being stated that the suggestion of historical balancing selection at this locus could be explained by ascertainment bias alone(187).

This is shown in the analysis of *PRNP* codon 129 frequencies in time cohorts. The statistically significant enrichment of heterozygote genotypes in individuals born in the 1940's reflect the selective advantage of heterozygosity at this locus. This is a replication of a finding of a previous study that found Hardy-Weinberg disequilibrium in women born before 1950 in the kuru affected region.(30)

Within EHPNG the updated analysis of *PRNP* codon 129 minor allele frequencies do not show the kuru region in having particularly high valine frequencies in comparison to other non-kuru-affected linguistic groups in the region.

The elevated minor allele frequency in PNG had previously been used as evidence to support a claim of historical kuru-like epidemics within EHPNG. The consistency of cultural practices and cosmology in EHPNG between groups inside and outside of the kuru affected region based on the belief in 'five souls' and veneration of the physical form may have historically permitted the practices of mortuary consumption in the past and outbreaks of kuru(151, 177). The likelihood of past kuru –like epidemics would have been impacted by other factors including population size. This is a result of the understanding that the index case of kuru was a case of sCJD(45), and the likelihood of sCJD occurring in a population is best predicted by population size with incidence of sCJD consistent

across populations(39). Estimations of historical population size based on WGS data from PNG Highland individuals showed the highland populations to have undergone consistent expansion in population size over the last ten millennia, coincidental with the development of agriculture in the region(84, 188). Scholarship from other fields has pointed to a more recent ‘population explosion’ resulting from the Ipomean revolution in the last four hundred years which was discussed above. These factors point to favourable cultural and demographic parameters for past kuru-like epidemics. The genetic analysis of current *PRNP* codon 129 allele frequencies does not provide additional genetic evidence to support these hypotheses. The highest derived allele frequencies were found in the Pawaian linguistic group. This linguistic groups is situated outside of the sphere of cultural similarity in EHPNG, and as noted in the previous chapter can be considered as isolate within an isolate(110). These groups also exhibit large amounts of genetic drift, due to having historically lived at low population densities. Variation in allele frequencies at *PRNP* codon 129 are best supported by being primarily driven by genetic drift and other demographic effects than historical balancing selection.

Evidence from ancient Eurasian ancient DNA samples ranging from 37kya to 1kya reflect a polymorphism in *PRNP* codon 129 present temporally as well as spatially. The possibility of mortuary feasts existing in prehistoric Europe has been debated by scholars in the field of archaeology for much time. The finding of mass burials of individuals in Magdalenian Europe (30kya -12kya) with signs of post mortem removal of flesh has offered support to the idea of mortuary feasts being practiced(189). The practice of mortuary feasts is a more efficient way of spreading prions through a population compared to exocannibalism with greater participation, practice and availability of infectious material. Finding genetic evidence to support such hypotheses is difficult. Making inferences based on allele frequencies of modern day Europeans would be fraught with difficulty due to large proportions of ancestry of modern Europeans being represented by populations who migrated into Europe after this period and admixed with the previous populations. This model of population change has been characterised with the arrival of Anatolian farmers into the region and

their admixture and replacement of the previous hunter gather population and then a subsequent migration from the Steppe region of Eastern Europe causing a major change in demography(190, 191). This is a simplistic representation of the population history of Europe over many millennia but still reflects the complexity in making inferences of past selection events based on modern allele frequencies. The use of ancient DNA from individuals that existed closer to the hypothesised periods of selection would provide more accurate inferences. Obtaining a sample set of sufficient size, quality and of belonging to the correct temporal period to support such hypotheses remains challenging. The ancient DNA evidence reported here does not permit any conclusion other than the *PRNP* codon 129 polymorphism appears to be present across multiple areas of pre historical Europe and across a range of historical periods.

3.4.3 Analysis of *PRNP* 127V

Use of haplotype based techniques to ascertain further the properties of the *PRNP* 127 variant within the affected populations has provided a better understanding of how recent novel variants that have been positively selected are characterised in terms of haplotype backgrounds within populations like those in EHPNG. Most known examples of positively selected variants have been found to have been at high frequency as a result of long periods of positive selection (192, 193). The *PRNP* 127V example has provided the opportunity to study a variant under positive selection for only a short period of time and at low allele frequency in the overall population under study.

This understanding has also informed approaches taken in the following chapter to uncover other variants that were under selective pressure during the kuru epidemic. It is clear that attempting to find recent novel variants under selection during the kuru epidemic would not uncover variants equivalent to *PRNP* 127V. The *PRNP* 127V variant was itself discovered through a focused analysis of the *PRNP* gene due to its previous association with other forms of prion disease(30, 39). The *PRNP* 127 variant would not have been discovered with analysis using the Chromomatcher or Relate programmes without the prior understanding of the importance of the *PRNP* gene to prion disease

mechanisms. The haplotype that was successfully found within the South Fore population as a whole did not rank highly when scored for average mean matching of haplotype carriers compared to non-carriers in Chromomatcher.

It was expected given the recent selection pressure and the likely recent origin of the variant that this would result in a particularly long shared haplotype between carriers of the variant compared to non-carriers. Although the average shared region of matching of *PRNP* 127V carriers in the highest scoring cluster is over 4MB this does not rank highly or appear exceptional within this population where long shared haplotypes are abundant due to the population dynamics discussed in the previous chapter.

The Chromomatcher tool could be further optimised to be better placed to find recently selected variants during the kuru epidemic. Challenges of excessive false positive signals in Chromomatcher due to long haplotype sharing caused by recent ancestry are accentuated when applied to EHPNG due to the historical genetic drift, isolation and practice of consanguineous marriages(100, 102). Finding variants that are specifically a result of kuru selection pressure may also be complicated by the presence of variants and shared long haplotypes being present as the result of multiple other selection pressures that populations in the region would have encountered historically. Improved clustering of *PRNP* 127 carriers could be obtained through experimentation with Chromomatcher clustering and scoring parameters and provide an improved basis for finding other variants on the precedent of an evolving loci in the affected population. Changing parameters to test hypotheses of variants with different population distributions and residing on different haplotype backgrounds could expand the investigation into finding other variants and genes that play a role in kuru.

3.5 Summary of findings

- The kuru epidemic appears to have resulted in a reduction in genetic diversity over the course of the epidemic.

- New analysis of *PRNP* codon 129 minor allele frequencies in primary and publicly available data does not provide additional support to hypotheses of past kuru-like epidemics.
- Use of a new Chromomatcher tool successfully clusters carriers of the *PRNP* 127V variant due to similarities in the haplotype in which the variant resides. This haplotype score does not provide a strong signal of selection when the whole South Fore population is tested. In order to find similar variants to *PRNP* 127V Chromomatcher may have to be used in conjunction with other methods.

3.6 References

1. Bradburd GS, Ralph PL, Coop GM. A Spatial Framework for Understanding Population Structure and Admixture. *PLoS genetics*. 2016;12(1):e1005703.
2. Pedersen CT, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, et al. The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics*. 2017;205(2):787-801.
3. López S, van Dorp L, Hellenthal G. Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary bioinformatics online*. 2016;11(Suppl 2):57-68.
4. Musunuru K, Kathiresan S. HapMap and mapping genes for cardiovascular disease. *Circ Cardiovasc Genet*. 2008;1(1):66-71.
5. Ranciaro A, Campbell Michael C, Hirbo Jibril B, Ko W-Y, Froment A, Anagnostou P, et al. Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics*. 2014;94(4):496-510.
6. Karupiah G, Panchanathan V, Sakala IG, Chaudhri G. Genetic resistance to smallpox: lessons from mousepox. *Novartis Foundation symposium*. 2007;281:129-36; discussion 36-40, 208-9.
7. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. *Science (New York, NY)*. 2014;343(6172):747-51.
8. Tabajara de Oliveira Martins D, Rodrigues E, Casu L, Benitez G, Leonti M. The historical development of pharmacopoeias and the inclusion of exotic herbal drugs with a focus on Europe and Brazil. *Journal of ethnopharmacology*. 2019:111891.
9. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature*. 2016;538(7626):510-3.
10. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010;468:1053.
11. Collinge J, Whitfield J, McKintosh E, Beck J, Mead S, Thomas DJ, et al. Kuru in the 21st century--an acquired human prion disease with very long incubation periods. *Lancet (London, England)*. 2006;367(9528):2068-74.
12. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
13. Lindenbaum S. Cannibalism, kuru and anthropology. *Folia neuropathologica*. 2009;47(2):138-44.
14. Anderson W. *The collectors of lost souls : turning Kuru scientists into whitemen*. Baltimore, Md.: Johns Hopkins University Press; 2008.
15. Alpers MP. Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3707-13.
16. Whitfield JT, Pako WH, Collinge J, Alpers MP. Cultural factors that affected the spatial and temporal epidemiology of kuru. *Royal Society open science*. 2017;4(1):160789.
17. Connolly B, Anderson R. *First contact*. New York, N.Y., U.S.A.: Penguin Books; 1988.
18. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A Novel Protective Prion Protein Variant that Colocalizes with Kuru Exposure. *New England Journal of Medicine*. 2009;361(21):2056-65.
19. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Beck J, et al. Genetic susceptibility, evolution and the kuru epidemic. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3741-6.
20. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A novel protective prion protein variant that colocalizes with kuru exposure. *The New England journal of medicine*. 2009;361(21):2056-65.

21. Mead S, Stumpf MP, Whitfield J, Beck JA, Poulter M, Campbell T, et al. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science (New York, NY)*. 2003;300(5619):640-3.
22. Hardy J, Scholz S, Evans W, Goldfarb L, Singleton A. Prion genotypes in Central America suggest selection for the V129 allele. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2006;141b(1):33-5.
23. Kreitman M, Di Rienzo A. Balancing claims for balancing selection. *Trends in genetics : TIG*. 2004;20(7):300-4.
24. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*. 2014;15:379.
25. King KC, Lively CM. Does genetic diversity limit disease spread in natural host populations? *Heredity (Edinb)*. 2012;109(4):199-203.
26. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature genetics*. 2019;51(2):277-84.
27. Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, et al. Early Neolithic genomes from the eastern Fertile Crescent. *Science (New York, NY)*. 2016;353(6298):499-503.
28. Asante EA, Smidak M, Grimshaw A, Houghton R, Tomlinson A, Jeelani A, et al. A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*. 2015;522(7557):478-81.
29. Albrechtsen A, Sand Korneliusen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol*. 2009;33(3):266-74.
30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome research*. 2002;12(6):996-1006.
31. Gajdusek DC, Alpers M. Genetic studies in relation of kuru. I. Cultural, historical, and demographic background. *American journal of human genetics*. 1972;24:Suppl:1-38.
32. Lindenbaum S. Kuru sorcery : disease and danger in the New Guinea highlands. 2016.
33. Sorenson ER, Claessen HJM, du Torr BM, Griffith J, Hockings P, Jablonko A, et al. Socio-Ecological Change Among the Fore of New Guinea [and Comments and Replies]. *Current Anthropology*. 1972;13(3/4):349-83.
34. Watson JB. THE SIGNIFICANCE OF A RECENT ECOLOGICAL CHANGE IN THE CENTRAL HIGHLANDS OF NEW GUINEA. *The Journal of the Polynesian Society*. 1965;74(4):438-50.
35. Wadsworth JDF, Joiner S, Linehan JM, Asante EA, Brandner S, Collinge J. Review. The origin of the prion agent of kuru: molecular and biological strain typing. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3747-53.
36. Lloyd SE, Mead S, Collinge J. Genetics of prion diseases. *Current opinion in genetics & development*. 2013;23(3):345-51.
37. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538(7624):207-14.
38. Bergstrom A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science (New York, NY)*. 2017;357(6356):1160-3.
39. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
40. Rougier H, Crevecoeur I, Beauval C, Posth C, Flas D, Wißing C, et al. Neandertal cannibalism and Neandertal bones used as tools in Northern Europe. *Scientific reports*. 2016;6:29005-.
41. Reich D. Who we are and how we got here : ancient DNA and the new science of the human past 2019.

42. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szecsenyi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. *Nature*. 2018;555(7695):197-203.
43. Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, et al. Lactase haplotype diversity in the Old World. *American journal of human genetics*. 2001;68(1):160-72.
44. Zhong M, Zhang Y, Lange K, Fan R. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. *Statistics and its interface*. 2011;4(1):51-63.
45. Lindenbaum S. First impressions of the Fore. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3648-9.

Appendices

Appendix 1 – Sources of Data Used in Analyses

Analysis	Section	Figure	Number of Individuals	Name of Primary Dataset
Analysis of genome-wide signatures of population response to kuru - Chromopainter	3.3.1	3.1	236	PNG merge
Analysis of genome-wide signatures of population response to kuru - Plink Heterozygosity	3.3.1	3.2	88	PNG merge
Ancient DNA Analysis	3.3.2	3.4	126	Courtest Yoel Diekman
EHPNG - Linguistic group PRNP 129 frequencies	3.3.2	3.5	1496	PNG merge
EHPNG - South Fore PRNP 129 frequency by DOB	3.3.2	3.6	513	PNG merge Imputed
Relate analysis of PRNP 127V	3.3.3	3.7	192	PNG merge
Chromomatcher Analysis of PRNP 127V haplotype	3.3.3	3.8	192	PNG merge

Appendix 1b – Table of ancient DNA samples with PRNP codon 129 frequencies

Sample Name	Longitude	Latitude	GENOTYPE	Estimated Date (years before present)
RISE446	10.18	50.01	AG	4779
I0057	11.04	51.89	AA	7157
RISE563	13.02	48.69	AA	4550
RISE247	18.90	47.33	AA	3696
RISE483	18.80	47.24	GG	3950
RISE373	20.20	46.22	AA	3836
RISE374	20.30	46.32	AG	3816
I1736	35.14	48.30	AG	6248
RISE548	43.70	46.54	AA	4950
RISE394	55.16	52.45	AA	3899
RISE500	85.45	53.46	AA	1950
I7571	0.03	52.34	GG	1448
I5377	-0.27	51.47	AG	1894
I0409	0.60	43.50	AA	5310
I0412	0.50	42.50	AA	5310
I7580	-0.51	52.1	AG	1256
I2443	-1.32	51.8	AA	2284
I2445	-1.42	51.90	AA	2396
I2601	1.36	51.33	GG	1955
Iceman	10.92	46.75	AA	3484
I0046	11.04	51.89	AA	5211
I0100	11.14	51.99	GG	5202
I0054	11.53	51.7	AG	7159
I0118	11.63	51.45	AA	2471
I0172	11.7	51.41	AA	3360
I0103	11.77	51.52	AG	2578
I9030	12.31	46.25	AA	14180
I6591	12.85	48.94	AA	2500
I5834	12.75	48.84	GG	2500
RISE97	13.06	55.56	AA	2025
RISE174	13.1	55.56	AG	427
Gokhem2	13.41	58.17	AG	3050
RISE98	13.45	55.38	AG	2275
I7213	14.07	50.41	AG	2500

I7286	14.17	50.51	AA	2500
I7214	14.17	50.41	AA	2500
I7249	14.16	50.19	AA	2500
RISE94	14.23	56.03	AG	2621
RISE577	14.31	50.16	AA	2200
I7200	14.37	50.05	AG	2200
I7195	14.37	50.05	GG	2200
I4893	14.45	50.12	AA	4449
I4891	14.56	50.22	AA	2281
I4888	14.16	50.12	AA	2190
I0014	15.05	58.54	AG	5877
I0011	15.15	58.64	GG	5721
I0012	15.15	58.54	GG	5714
I0017	15.05	58.64	GG	5721
RISE150	17.00	50.92	AG	1885
I6582	18.10	50.09	AA	2343
Ajvide58	18.55	57.50	AG	2900
I0408	-2.33	41.25	AG	3893
I0406	-2.43	41.25	GG	3900
I2568	-2.44	55.97	GG	2286
I2631	-2.57	59.23	AA	3097
I2651	-2.87	59.35	AA	3360
I2933	-2.92	58.74	AA	3010
I2630	-3.02	58.84	AA	2581
I2932	-2.92	58.74	AA	2570
I3085	-2.92	58.84	GG	3339
I2979	-2.92	58.64	GG	3333
I0449	20.43	46.36	AA	5000
I1500	20.80	47.81	AG	5210
I0676	21.35	41.90	AG	5979
I4875	22.05	44.53	AG	6615
I4432	25.13	56.28	AG	6075
I4596	25.13	56.38	AG	6061
I4439	25.23	56.28	AA	5763
I4628	25.23	56.38	GG	5302
I0070	25.83	35.08	AG	2400
I0073	25.93	35.08	AG	2400
I9005	25.73	35.08	AA	2400
I0071	25.83	35.18	AA	2400
I2510	25.88	43.16	AG	2906
I2509	25.98	43.16	AG	4452
I5769	25.89	43.16	GG	500
RISE00	27.03	59.41	AG	2575
ILK001	27.69	49.56	AA	2899
I1584	29.57	40.30	AA	3943
I1100	29.67	40.30	AA	6500

I0744	29.47	40.30	AA	6402
I1580	29.57	40.20	AA	6387
I0745	29.57	40.40	AA	6374
I1583	29.67	40.40	GG	6426
I7554	-3.25	59.00	AA	3366
I2861	-3.39	57.72	AA	977
I5883	33.76	48.91	AG	5208
Kostenki14	39.05	51.40	AA	36730
KK1	43.29	42.28	AG	7940
RISE552	43.33	46.60	AG	2849
I0374	44.67	49.97	AA	2800
I1407	45.20	39.73	AA	4350
I1290	48.12	34.45	AA	8179
I1955	48.22	34.45	AA	1430
I0231	48.24	52.42	GG	2921
I4073	5.10	52.73	AA	2195
I4070	5.20	52.73	GG	1880
I5748	5.10	52.83	GG	2579
I0444	51.13	53.10	AA	3335
I0232	54.44	48.10	AA	1850
RISE386	55.16	52.45	AG	2298
RISE523	57.01	53.04	AA	1598
RISE395	59.54	52.64	AG	1960
Rathlin2	-6.19	55.30	AA	2024
Rathlin1	-6.29	55.30	AA	2026
Loschbour	6.29	49.80	AG	6210
I0585	-6.89	43.48	AA	5983
Bichon	6.97	47.02	GG	11820
BOT15	67.65	53.31	AA	3327
BOT14	67.75	53.31	AA	3371
BOT2016	67.65	53.21	GG	3521
Ballynahatty	-7.33	53.46	AA	3343
RISE504	85.45	53.46	AA	721
RISE505	85.55	53.46	AA	1746
RISE502	88.57	51.91	AA	1496
RISE499	88.57	51.81	AA	1400
I0018	9.18	48.80	AG	5310
RISE496	90.19	53.00	AG	1414
RISE495	90.19	53.00	AG	1400
RISE497	90.29	53.00	AG	1400
RISE670	90.21	53.16	GG	2141
RISE683	90.36	53.71	AG	2138
RISE511	90.78	54.58	AA	2909
RISE509	90.68	54.58	GG	2887
RISE664	91.03	53.55	GG	2459
RISE493	91.05	53.16	AA	1531

Appendix 2 – PRNP codon 129 frequencies

GnomAD

Population	Allele Count	Allele Number	Homozygote Number	Frequency
Latino	14597	35432	3025	0.412
European (non-Finnish)	43606	128942	7436	0.3382
African	8415	24884	1397	0.3382
Other	2243	7228	331	0.3103
European (Finnish)	7533	25108	1129	0.3
Ashkenazi Jewish	2852	10366	383	0.2751
South Asian	7797	30616	1009	0.2547
East Asian	478	19946	9	0.02396
Female	40628	129300	6924	0.3142
Male	46893	153222	7795	0.306
Total	87521	282522	14719	0.3098

EHPNG Populations

Linguistic Group	N	G Frequency
AGARABI	24	65%
ANGA	31	50%
ASARO	25	66%
AUYANA	29	44%
AWA	28	66%
BENABENA	27	59%
GADSUP	29	59%
GAHUKU	26	63%
GIMI	42	45%
KAMANO	28	68%
KANITE	3	50%
KEIAGANA	42	57%
LABOGAI	22	53%
NORTH FORE	238	58%
PAWAIAN	32	75%
SIANE	28	40%
SOUTH FORE	571	54%
TAIRORA	83	45%
YABIYUFA	25	50%
YAGARIA	126	58%
YATE	26	52%
TOTAL	1485	56%

EGA PNG Highlands Whole Genome Sequence Data *PRNP* codon 129 genotypes

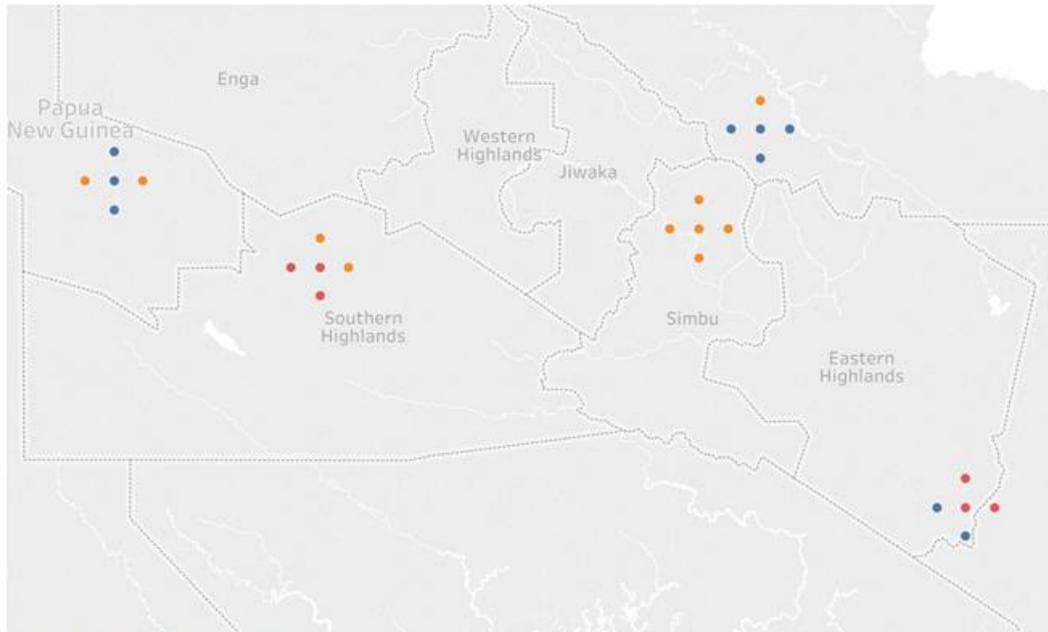


Figure 1. *PRNP* codon 129 guanine frequencies per ancient WGS samples from a previous publication of PNG Highland genomes. red points indicates homozygous derived genotype, orange points homozygous and blue homozygous ancestral allele. Noise has been added to the location of points from the same population to add clarity.

sample_ID	BAM_ID	latitude	longitude	language
BUN01	Papuans6103671	5.67338	145.17883	Gende
BUN02	Papuans6103673	5.67338	145.17883	Gende
BUN03	Papuans6103675	5.67338	145.17883	Gende
BUN04	Papuans6103676	5.67338	145.17883	Gende
BUN05	Papuans6103677	5.67338	145.17883	Gende
KUN01	Papuans6103680	6.01666	144.96666	Kuman
KUN02	Papuans6103681	6.01666	144.96666	Kuman
KUN03	Papuans6103682	6.01666	144.96666	Kuman
KUN04	Papuans6103685	6.01666	144.96666	Kuman
KUN05	Papuans6103689	6.01666	144.96666	Kuman
MEN01	Papuans6103694	6.14777	143.65722	Angal
MEN02	Papuans6103695	-	143.65722	Angal

		6.14777		
MEN03	Papuans6103696	-	143.65722	Angal
MEN04	Papuans6103697	6.14777	143.65722	Angal
MEN05	Papuans6103699	-	143.65722	Angal
TAR01	Papuans6103705	5.85101	142.9462	Huli
TAR02	Papuans6103706	5.85101	142.9462	Huli
TAR03	Papuans6103707	5.85101	142.9462	Huli
TAR04	Papuans6103708	5.85101	142.9462	Huli
TAR05	Papuans6103709	5.85101	142.9462	Huli
MA1	Papuans6103712	6.97436	145.88853	Angan
MA2	Papuans6103714	6.97436	145.88853	Angan
MA3	Papuans6103715	6.97436	145.88853	Angan
MA4	Papuans6103716	6.97436	145.88853	Angan
MA5	Papuans6103717	6.97436	145.88853	Angan

Appendix 3 – Samples in *Impute Merger*

Linguistic Group	N
AGARABI	24
ANGA	31
ASARO	25
AUYANA	29
AWA	28
BENABENA	27
GADSUP	29
GAHUKU	26
GIMI	32
KAMANO	28
KANITE	3
KEIAGANA	40
LABOGAI	22
NORTH FORE	212
PAWAIAN	32
SIANE	28

SOUTH FORE	541
TAIRORA	83
YABIYUFA	25
YAGARIA	126
YATE	26
Total	1417

Chapter 4 – Search for variants under selection pressure during the kuru epidemic

4.1 Introduction

Work performed in this chapter focused on the discovery of variants that played a role in the population response to the kuru epidemic. Such variants may reveal information about the underlying biology of kuru and other prion diseases. The principal hypothesis that was tested in this chapter was;

- Genetic variants exist in kuru-affected communities that exerted a strong influence on kuru resistance. These will have similar properties to the previously discovered *PRNP* 127V variant and may exist outside of the *PRNP* locus.

Analyses were performed in this chapter to ascertain if such variants existed in the affected populations at varying population frequencies. This was done using a variety of computational tools and approaches. This included a GWAS to check for common variants, linkage disequilibrium and haplotype based analyses on focused groups of individuals to attempt to find variants sitting at lower frequencies or within subpopulations, and Whole Genome Sequence (WGS) data analysis to find rare variants. Analyses were designed to take the complications and nuances of the kuru epidemic and population genetics into account (See Chapter 2 for more) in order to minimise the presence of false positive signals and erroneous interpretation of findings. Information is provided below regarding some of these complicating factors that impacted on experimental design.

4.1.1 Complex patterns of incidence and exposure

Once at mortuary feasts, differences existed in terms of the distribution and consumption of infectious material(110). Complex rituals and rights that varied over time and locale within the kuru-affected region would determine these differences(48, 110). At the height of the epidemic, adult males represented only 2% of cases, which reflects consumption patterns at feasts(49). Males are

believed to have only consumed infectious material as young infants prior to initiation (6-10 years)(50). Whether dosage of infectious prion material at mortuary feasts (type of material consumed and amount) and the number of feasts attended played a role in downstream phenotypic presentation is still unknown. Consumption of kuru dead at mortuary feasts was prohibited in villages in the Gimi and North Fore linguistic groups prior to the prohibition and cessation of feasts by colonial authorities(48).

These uncertainties combined with the complex patterns in cessation of mortuary feasts make experimental design challenging. It is difficult to know for certain if individuals born in the 1950's would have definitely attended a mortuary feast. An individual's likely exposure to kuru at a particular time can be estimated by referencing the amount of kuru deaths in that village to the village size. This was the approach taken in a previous publication to develop an exposure index(30). As kuru surveillance with precise recording of kuru cases and regular village census' only began in 1957 several decades after the commencement of kuru, the relevance of these exposure estimates earlier in the epidemic can be questioned.

4.1.2 Aetiology and architecture of kuru

Understanding of the genetic aetiology of kuru can be further understood by examining differences beyond a simple case/control categorisation of individuals. Differences were observed in terms of incubation period between exposure to prions at mortuary feasts and onset of symptoms(47, 50). This incubation can only be roughly estimated with a dietary based acquired prion disease (as is the case with vCJD), as the particular time of exposure cannot be known for certain. This is particularly the case if individuals attended multiple feasts and dates of feasts. Dates and frequency of feasts can only be roughly estimated from interviews conducted in some cases many decades after the cessation of mortuary feasts in EHPNG. Individuals who attended mortuary feasts and would have been highly exposed to infectious prions did not go on to develop prion disease up to five decades

subsequently and present resistant individuals, some of whom have been shown to harbour protective genetic variants in the *PRNP* gene(30).

Many unknowns still remain regarding the kuru disease process that present complications when attempting to design highly powered analyses. The likelihood of developing kuru after attending a mortuary feast is still unknown. It has been established that mortuary feasts did not take place after 1959 due to evidence obtained from ethnographic interviews and supported by the absence of anyone born after this date developing prion disease(49). Prohibition of mortuary practices was implemented with the gradual extension of colonial rule in the region during the 1950's(54). This would have applied differently to different areas within the kuru-affected region and also there would not have necessarily been an immediate cessation of feasts after official prohibition, especially in more remote villages where colonial rule was less firmly felt(51).

4.1.3 Complex population structure

Analysis of EHPNG and the kuru affected region in earlier chapters of this thesis from population genetics analysis show a highly structured set of communities, experiencing high levels of relatedness and genetic drift in addition to extreme genetic heterogeneity between the different linguistic group communities. When designing experiments, regular thresholds for relatedness of participants may have to be reconsidered to ameliorate the impact of false positive findings. The geographical isolation, population differentiation and genetic drift observed make it likely that EHPNG populations have experienced population bottlenecks in the past. It has been documented before for the tendency for historic population bottlenecks followed by population recovery to generate linkage disequilibrium patterns that appear similar to those generated by genomic regions under positive selection pressure(194).

This tendency is clearly apparent in our analysis of patterns of haplotype sharing in carriers of *PRNP* 127V in the previous chapter. The *PRNP* 127V variant lies on a long haplotype background compared to typical values at this site in the same population, but this locus is not an outlier in genome-wide

and population-wide analyses. The low frequency of the variant in the South Fore linguistic group overall reflects the short duration of the epidemic and the variants recent origin. Approaches to find new variants must not discount the possibility of variants that were at a higher frequency due to drift and deeper historical origin. The genetic action and role of any new variant may be completely different to those previously discovered, particularly if residing outside of the *PRNP* locus. Strategies to find variants of effect sizes similar to *PRNP* 127V will benefit from recognition of the variants low population frequency overall. Rare variants with a recent origin may not be widely distributed. Strategies to focus attention on informed phenotypic sub-populations could therefore be valuable. The distinct ancestral history of the region as a whole will result in novel variants existing in EHPNG that do not exist in European and other more regularly studied populations.

An advantage of the extensive genetic drift experienced by EHPNG populations means that many variants that may have been at too low a frequency to be detected by previous association studies of prion disease may exist at higher population frequencies in the kuru region and more readily detectable. Additionally the prolonged separation of these groups for millennia since the time of most recent common ancestry with previously studied population also raised the possibility of novel variants existing that may have powerful effects on kuru.

4.1.4 Recent and short epidemic

Ethnographic work suggests that first kuru case occurred at the end of the 19th century in the village of Uwami(51, 177), and exposure to kuru prions at mortuary feasts ceased by 1960. Assuming these dates represent the full duration of kuru exposure, selection pressure persisted in the region for a period of 4-6 generations. This is assuming that there were no previous kuru-like epidemics in recent centuries. Most selective sweeps require dozens if not hundreds of generations for a variant at low frequency to reach high frequency and fixation. Selection pressure in the kuru affected region during the epidemic does appear to have been high with many villages in the most affected areas being

devoid of women(195). Given the recent and short nature of the selection pressure in question, tools that had highest power to detect periods of recent selection were carefully chosen.

4.2 Materials and methods

4.2.1 Phenotypic information in the dataset

Individuals were placed into kuru-related phenotypic categories informed by work from a previous publication(30). This measure took into account their village of residence, their year of birth and whether they developed kuru. Year of birth contributed to estimations of individual’s exposure to kuru. Individuals born after 1960 were assumed to be unexposed. There was a gradual cessation of the practice of mortuary feasts in the region, particularly in the 1950’s with the onset of colonial administrative control (see Introduction of this chapter for more information on mortuary rights, exposure estimations and assumptions). The sex of individuals also contributes to exposure estimates with the clear sex bias in cases of kuru due to participation rights. Individuals who died of kuru can also be further stratified based on their age at death which can act as a proxy for incubation period, although the precise date of exposure and consumption of infectious prions is unknown. This led to the categorisation of individuals accordingly for the purposes of investigations carried out in this chapter from individuals from the *PNG merge* data.

Category	Exposure	N
Young Kuru Cases	High	47
Older Kuru Cases	High	43
Elderly Women Highly Exposed	High	141
Women Born in 1950s	Medium	85
Men Born Pre 1950s	Medium	117
Men Born in 1950's	Low	78
Unexposed Kuru Affected Region Born After 1960	Zero	353
EHPNG Individuals from Kuru Unaffected Region	Zero	561
Total		1425

Table 4.1. Data from *PNG merge* used in analysis in this chapter. This collection represents individuals after quality control measures have been performed (See Chapter 2 for details).

Individuals from within the kuru affected region can be categorised as having resided in villages with high, medium and low kuru exposure. These categorisations were used in order to design experiments where individuals could be grouped and excluded based on their expected shared genetic loads for kuru risk variants. This was designed to lead to more highly powered studies and a reduction in false positive findings.

4.2.2 GWAS

Performing a GWAS analysis on kuru has to be regarded in a different context to GWAS on common traits, and separate from the common disease common variant hypothesis (see section 1.6). Kuru as an acquired prion disease epidemic took place in a discrete location during a handful of generations. The genetic architecture in the affected population may be different due to the different selection pressures experienced as a result of kuru compared to that of common diseases studied in GWAS.

Practical limitations of any association analysis also exist when performing a GWAS with kuru as the trait of interest. The number of recorded deaths during the epidemic was < 3,000, meaning a theoretical upper limit of participants with kuru and any associated variants is relatively low compared to conventional GWAS which are now recruiting hundreds of thousands of participants.

The time period of the epidemic, its location and the challenge of building ties with affected communities has presented enormous logistical and cultural challenges in recruiting individuals who died from kuru. In this dataset 151 individuals with kuru were recruited and DNA samples obtained. The majority of these individuals were recruited after the peak of the kuru epidemic and individuals in the *PNG merge* data set appear enriched for individuals with extended incubation periods. In this dataset 25% of individual are over 40 years of age compared to 10.5% in the peak years of the epidemic(49). This prolonged incubation period may be due to the presence of genetic factors that contribute to kuru incubation and resistance. It may be worth considering kuru risk as a quantitative trait that is best viewed in terms of multiple factors that inhibit the disease process developing over time instead of completely. In some cases this resistance may persist longer than a lifetime. The

quality of DNA observed from many of these samples particularly samples from the National Institute of Health archive derived from human serum are of such a low quality as to be unusable in any GWAS. This has left 90 samples of individuals with kuru available for analysis. The age and condition of these samples although passing quality control measures in terms of genotyping rates and missing data may contain errors as a result of the amplification techniques required in some cases to obtain DNA.

The high amount of relatedness and the extent of linkage disequilibrium observed in the population will reduce the power to find variants and also present challenges for mapping any associated variant to a possible to a causal variant(125). The genetic heterogeneity observed between kuru affected linguistic groups and other non-affected groups within the region provide challenges in obtaining a control group of individuals lacking exposure to kuru. Inclusion of individuals from non-kuru affected linguistic groups as a control would cause downstream problems in terms of population stratification, and any associated variants that segregate between the kuru affected region and non-kuru-affected regions will largely be due to differential ancestry than solely down to differential exposure to selection pressure from kuru. Exposure to kuru cannot be clearly defined due to a lack of documentation regarding participation of feasts and the period of time since exposure to kuru prions ceased. It is also challenging to gauge the predicted genetic load of individuals born after the exposure to kuru as their ancestries will contain both resistant individuals, unexposed individuals and individuals who died of kuru after having children.

The overwhelming problem for a GWAS of kuru is that it is acutely underpowered to find variants that are typically uncovered in conventional GWAS for common traits. It has to be accepted that a GWAS with numbers of samples available of the order in the *PNG merge* data set will ultimately be limited to finding variants of very strong effect sizes. There are a few precedents for this, including in prion disease with the discovery of strong genome-wide association of *PRNP* codon 129 genotype with vCJD status, in a study involving only 119 cases(72). The strategy deployed in attempting to find

a similarly associated variant was based on maximising available power in the dataset and designing the study in such a way to take into account the nuances in the epidemiology of kuru disease.

In order to overcome these challenges, two independent GWAS analyses were performed and their results combined in a meta-analysis. Combination of the results from these two separate analyses in a meta-analysis could help protect against false positives by ensuring that variants with opposite directions of effect in each of the analysis resulting in a null finding in the meta-analysis.

The first analysis was a case/control analysis with young kuru cases and kuru resistant individuals. This was achieved by categorising individuals who died of kuru below the age of 25 as cases and elderly women who lived during the period of highest exposure as controls. Selecting an age threshold of 25 was based on the expected lack of kuru incubation factors compared to older individuals. Elderly women as controls were further filtered to contain only those lacking protective variants at the *PRNP* locus (*PRNP* 127V carriers and *PRNP* codon 129 heterozygotes). After quality control procedures performed in plink 1.9 (--maf 0.005 --hwe 1e-5 --geno 0.01) 122,622 variants and 147 individuals (31 case and 116 control individuals)(160).

The second analysis used individuals born after the kuru exposure period (1960) as cases as they may contain kuru risk variants and all exposed individuals as a control group comprised of resistant individuals. No individuals from the first analysis were used in the second analysis to ensure independence of findings prior to combination in a meta-analysis. This analysis after the same quality control procedures contained 122,622 variants and 515 individuals (241 case and 274 control individuals).

Concerns about the lack of SNP density in these analyses and its inability to resolve subtle patterns of linkage disequilibrium in the analysis led to a parallel analysis being performed using the *Imputed* merger data set (see section 3.2.2). The GWAS analysis was performed in plink 1.9 using the --assoc

and `--adjust` functions, `--logistic` function was used when population covariates were required to reduce the impact of genomic inflation due to population stratification.

4.2.3 XP-EHH and Relate Analysis

In the previous chapter it was clear that the Relate tool was incapable of detecting *PRNP* 127 at lower population frequencies (Section 3.3.3). A similar lack of sensitivity is observed with XP-EHH, another tools used to detect recent positive selection in humans(196). If a signal of positive selection is driven by a subpopulation of the test population then it is not immediately feasible to observe which individuals are driving this signal. This is disadvantageous for studying kuru which has both great population and epidemiological heterogeneity between individuals.

To improve the likelihood of finding a positively selected variant using the Relate and XP-EHH tools, individuals from only the South Fore linguistic group (highest incidence and most densely sampled group) and distinct epidemiological groups were used (young kuru cases and kuru resistant individuals). If there are common variants in the population with a strong effect size on kuru then it was hoped that the GWAS performed would be able to detect such variants. It was then decided to focus efforts using these tools at finding variants at lower frequencies in the kuru-affected region as a whole.

XP-EHH analysis was performed within the Selscan application(197). After controlling for excessive relatedness (plink 1.9. `--genome PIHAT < 0.1875`), 21 individual elderly women highly exposed to kuru were analysed as a test population with 23 individuals who died of kuru under the age of 25 acting as the reference population.

To compliment the XP-EHH analysis, the same 21 highly exposed elderly women and 23 young kuru cases were analysed in relate (185)to search for long shared genomic regions exclusive to these particular cohorts.

4.2.4 Chromomatcher analysis

Unlike Relate and XP-EHH Chromomatcher as seen in the previous chapter, Chromomatcher has sensitivity to find shared haplotype regions at lower frequencies (section 3.4.3). This is due to the clustering of the Chromomatcher output which is designed to find subpopulations within a wider population. The interface of output which allows users to understand which individuals belong to clustered subgroups also has the benefit of using the phenotypic labels to eliminate false positives driven by non-kuru related phenomena. The ability to stratify output and find subpopulations means that experiments in this analysis did not have to be as focused as in the XP-EHH and Relate analyses. This permits the inclusion from lower exposure and unexposed populations which may allow the increase in frequency of any recently selected variants in an analysis.

Conventional tools designed to detect selected variants have been used regularly in hypothesis generating approaches. Work in kuru research is more highly focused than hypothesis generating research. A clear research question to answer is if there are genetic variants that have evoked a similar population response in affected communities due to providing kuru resistance like *PRNP* 127V. The Chromomatcher tool is ideally suited to exploit the particular circumstances of the kuru epidemic (see sections 4.11-4.1.4 for more details) compared to conventional approaches. Chromomatcher effectively combines assays of long haplotypes at anomaly high frequencies to suggest genomic loci under recent selection, and a clustering technique that will prioritise subpopulations carrying the haplotype within a heterogeneous population.

Chromomatcher has illustrated the properties of the *PRNP* 127V variant in the Fore population showing that it did not reach intermediate frequency in amongst the Fore and was geographically localised to the Purosa valley (section 3.3.3). This replicates findings from a previous study on these populations(30), showing Chromomatcher to be an effective tool in effectively highlighting carriers of alleles like 127V and potentially other variants with similar properties. The *PRNP* 127 haplotype did not result in a high score in the South Fore population as a whole (mean ratio score 25.7,

genome-wide average 34.6, section 3.3.3). This is largely due to the low frequency of the haplotype in the South Fore population (<3% in this analysis).

Chromomatcher ability to cluster groups of individuals will be limited if shared long haplotypes are at exceedingly low frequencies in an analysis. This is particularly the case in populations like those in the kuru affected region with the presence of extensive genetic drift and the practice of consanguineous marriages. Within a larger population many individuals will have long regions of haplotype sharing simply because of recent shared ancestry. It is therefore important to maximise the power of Chromomatcher to find shared regions due to positive selection through clustering, whilst utilising the full dataset and accompanying phenotypic information.

The first analysis performed was on 47 resistant individuals with high exposure to kuru. Individuals with known *PRNP* protective mutations or genotypes were removed from the analysis (*PRNP* 127V carriers and *PRNP* codon 129 heterozygotes). 12 kuru cases were also included in the analysis to help filter false positive high scoring haplotypes, giving a total of 59 individuals. Candidate regions were ruled out if the cluster was comprised of more than 15% of kuru case individuals.

A second analysis was performed on individuals from villages where there was high exposure to kuru. It was hoped that where selection pressure was highest the population response may have resulted in any kuru protective variants rising to a higher frequency more quickly making it more probable that Chromomatcher would have power to detect any such haplotypes. This was defined as villages that had a village exposure index > 250 (number of cases divided by 1958 census population multiplied by 1,000, as per a previous study (30)). This analysis contained 282 individuals which included 159 exposed individuals, 96 unexposed individuals and 27 kuru case individuals. If a candidate cluster contained more than 10% kuru case individuals this region was ruled out.

To gauge the lower limit of sensitivity of Chromomatcher to find *PRNP* 127V in an analysis, two analyses were performed on a set of 59 individuals (the number of individuals in analysis 1) with 22

and 11 *PRNP* 127V carriers respectively. An understanding of this sensitivity would provide understanding of Chromomatchers ability to find other variants similar to *PRNP* 127V. For each analysis Chromomatcher minimum sharing length and size parameters were altered systematically (minimum sharing 0.5MB, 1MB, 2MB, 3MB, 4MB; minimum cluster size 2%,5%,9%,18% of total individuals). The parameters that gave optimum clustering of the *PRNP* 127V carriers were used in subsequent Chromomatcher analyses.

4.2.5 Analysis of WGS data of kuru-resistant individuals

Analysis of available WGS sequence data was then seen as a possible way for finding high effect size variants that may exist at too low frequencies to be detected by either association analysis or the haplotype selection mentioned above.

30 samples from the *PNG merge* dataset were whole genome sequenced. Variant calling and genome annotation was performed by GATK and output placed into a standard query language (SQL) database. Analysis of this WGS data in order to assess the possibility of rare variants possibly having an effect on kuru analysis of WGS data from 30 EW kuru resistant individuals from the kuru affected region was performed. This group of individuals includes five *PRNP* 127V carriers. Presence of possible disease relevant variants in the group of 30 elderly women were checked to provide additional support for any candidate variants or regions suggested by the GWAS and selection analyses.

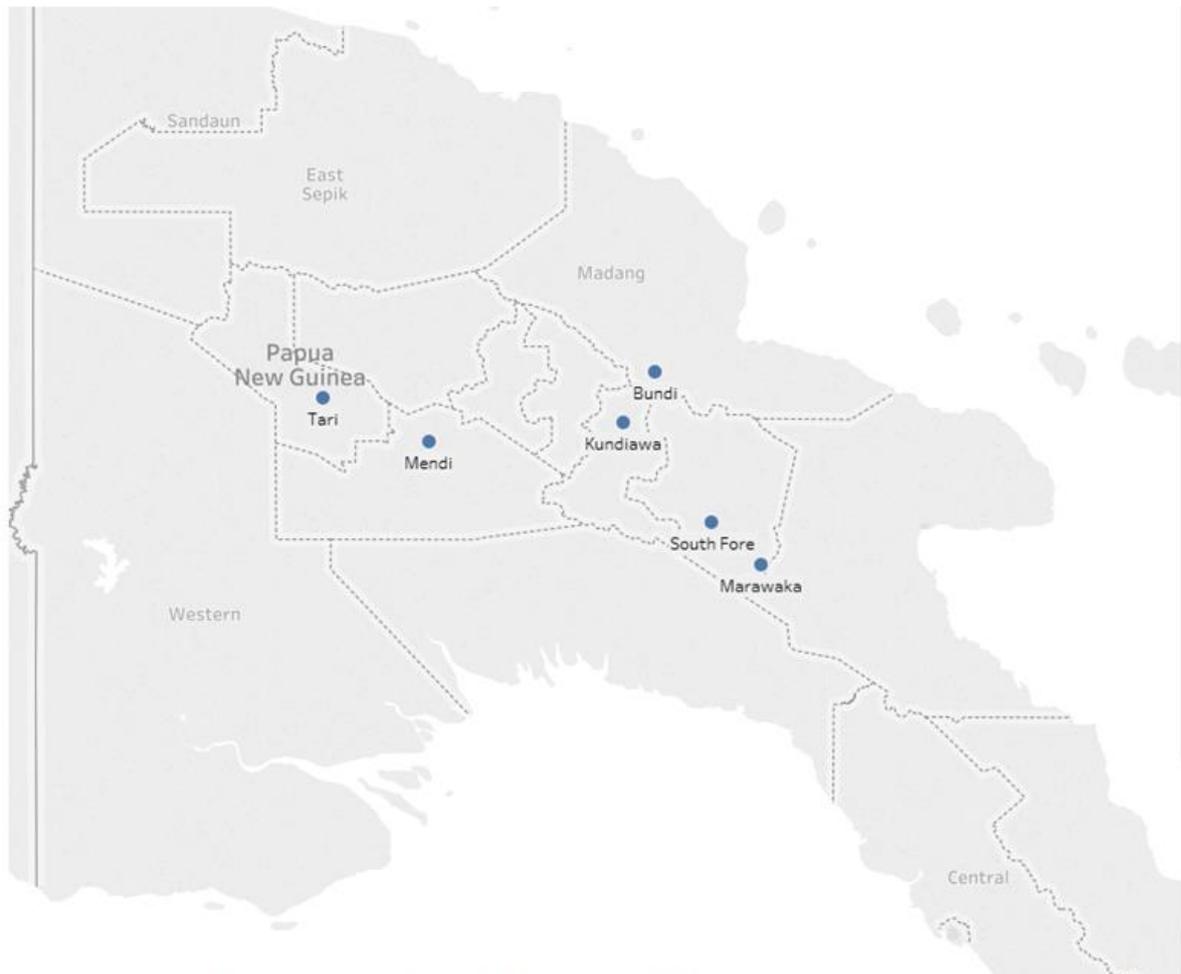


Figure 4.2 Map of Papua New Guinea showing sampling location of WGS data used in this chapter.

The analysis was focused on variants in these individuals that were found to be rare in global populations particularly European populations by referencing if a variant was common in the GnomAD database of >125,000 exome sequences. The reasoning being that a common variant globally was likely to have been tagged via linkage disequilibrium with the > 6,000,000 variants used in the latest sCJD GWAS. This was the case for the *PRNP* codon 129 polymorphism which has been shown to affect multiple prion diseases globally and is polymorphic in nearly all populations (in east Asians is < 3% according to GnomAD). Due to the huge number of variants generated in this analysis in particular due to the isolated nature of the populations it would be expected that there would be enrichment for deleterious variants generally. There was an initial filter of variants having a CADD score in excess of 40. To gauge the possible polymorphic nature of these variants within Papua New

Guinea (there are no Papuan samples in GnomAD) 25 WGS samples from five different PNG Highland groups available from a previous study of Papuan population were analysed. This data came from groups separated by considerable distances (both geographically and genetically) in the context of the PNG Highlands. If a variant was found to be in high frequency in the in multiple other PNG Highland individuals from different populations then it could be considered to be common and would have been expected to have been associated with a variant in the kuru meta-analysis performed earlier.

Population	Number of Genomes	Pairwise Fst with South Fore	Distance from South Fore (km)
Bundi	5	0.0265737	102.29
Kundiawa	5	0.017093	90.21
Mendi	5	0.0208593	217.00
Tari	5	0.0208593	300.05
Marawaka (Anga)	5	0.0346317	44.9

Table 4.3 Whole Genome Sequence Sample Information from samples derived from a previous study of Papuan Highlands populations. Fst values were obtained from analysis performed in Chapter 2 of this thesis.

High scoring windows from Chromomatcher analysis which contained individuals for which WGS data was available were also analysed in the same set of genomes. A lower CADD score threshold of > 20 was utilised as a filter due to the more focused search in this analysis. As with the analysis above variants were filtered in both the GnomAD web resource and available PNG Highlands WGS sequences to check if the variants were common either globally or more locally in the PNG Highlands. Remaining variants were then checked in a similar way for variant and gene function and relevance to neurodegeneration.

An additional analysis was performed where 4 genes recently to have variants in association with sCJD in a recently performed GWAS analysis of European populations were checked in the same 30 WGS samples from the kuru affected region. A lower CADD score threshold of 15 was applied due to the focused targeting of specific loci in this analysis and the lower concerns for false positive findings as a result of the associations found in sCJD. Original sources of data for analyses in this chapter can be found at Appendix 1.

4.3 Results

4.3.1 GWAS

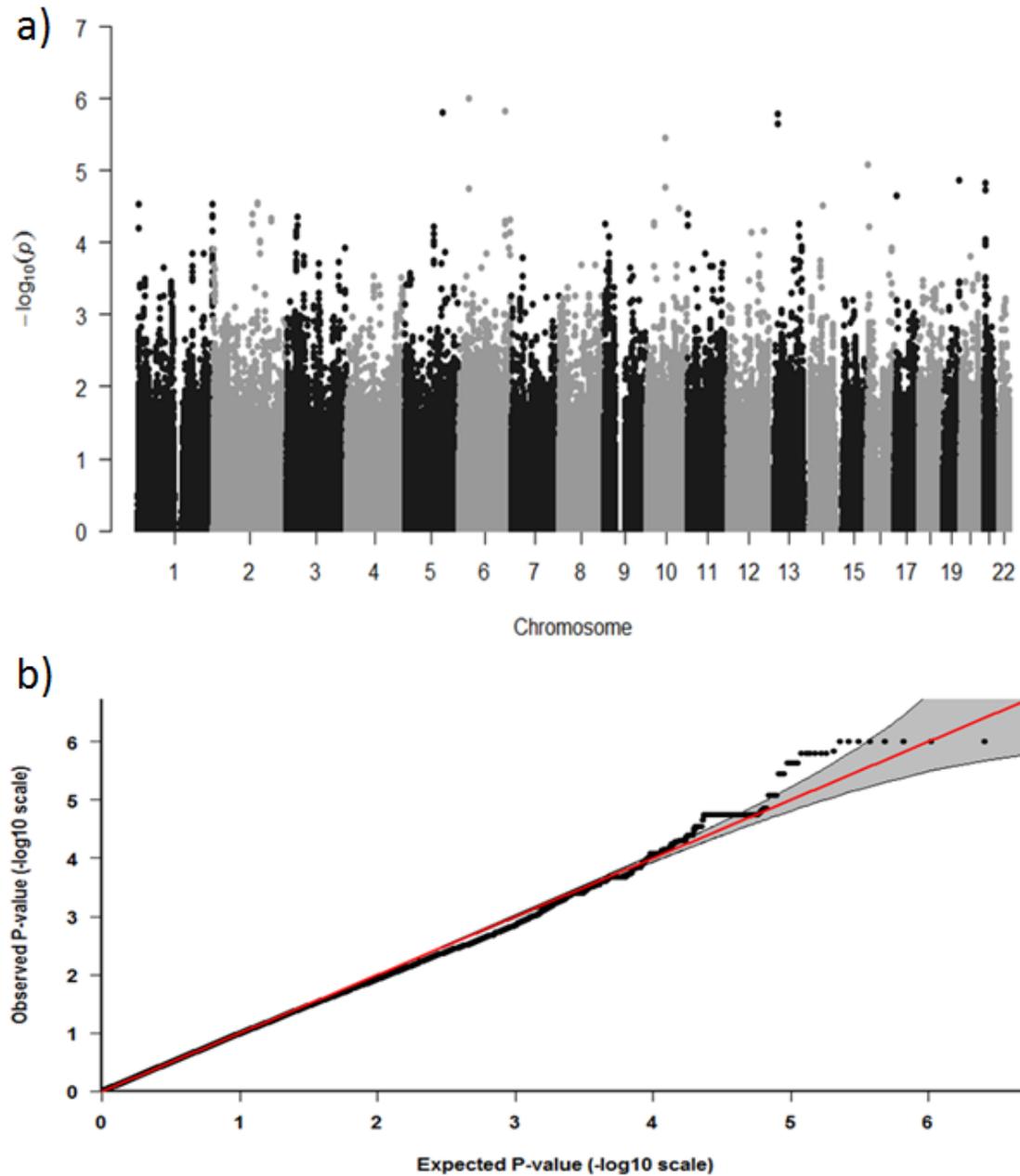


Figure 4.4 a) Manhattan plot of meta-analysis of kuru disease. This is the result of the combination of two supporting association analyses. (b) QQplot showing relationship between observed and expected p-values for this.

For the first GWAS analysis of 147 individuals a lambda value for genomic inflation of 1.04831 was obtained and inspection of the qqplot did not reveal excessive genomic inflation that may give rise to population stratification and downstream false positive findings. The second analysis of 515 individuals gave a lambda value of 1.04 but application of 10 population covariates using the --cluster -mds-plot 10 appears to have removed some extreme outlier SNP's in the distribution of observed p values that may have been due to being artefacts from the imputation process or remaining poorly genotyped SNP's. These SNP's were not in in linkage disequilibrium with surrounding SNP's which would be expected of a variant tagged to a casual variant nearby.

RANK	CHR	BP	SNP	OR	P	Gene	Brain Expressed
1	6	30046292	rs7940639	2.7692	9.81E-07	RNF39	N
2	13	28513719	rs9507970	2.3887	1.59E-06	BDX1	N
3	10	64547920	rs150904	2.603	3.47E-06	ADO	Y
4	16	3806693	rs129965	0.3663	8.20E-06	CREBBP	Y
5	19	53793042	rs8109165	0.3216	1.36E-05	BIRC8	Y
6	21	21174729	rs1518249	1.7379	1.50E-05	-	N
7	17	5750112	rs4985641	0.5891	2.19E-05	LOC33916	N
8	2	146950908	rs6216954	1.7262	2.82E-05	-	N
9	1	4237878	rs1203295	0.2819	2.89E-05	-	N
10	14	63058612	rs1710010	3.2399	3.04E-05	KCNH5	N

Table 4.5 Summary of top scoring variants in meta analysis of kuru. Brain expression was set at >5 RKPM as per the GTEX gene expression tool.

No variants have p-values below the conventionally accepted threshold of $p < 5 \times 10^{-8}$, and only four variants passing below a conventional 'suggestive threshold of $p < 5 \times 10^{-5}$. Given the underpowered nature of the study it is clear from the findings that there is no variant with strong effect sizes that have been detected in this analysis. Of the highest scoring variants only three resided in genes that are expressed in the brain. The gene CREBP has been associated with Huntingdon disease and other brain disorders. For most common traits findings of variants with strong effect sizes like that observed with the association with the APOE variant to Alzheimer's and the association of *PRNP* codon 129 to vCJD are something of a rarity. The degree of association for the highest scoring variants in this study are not of the order of that sign for *PRNP* codon 129 and it appears that none of the findings in this analysis indicate a variant with an effect of similar magnitude.

4.3.2 XP-EHH and Relate analysis

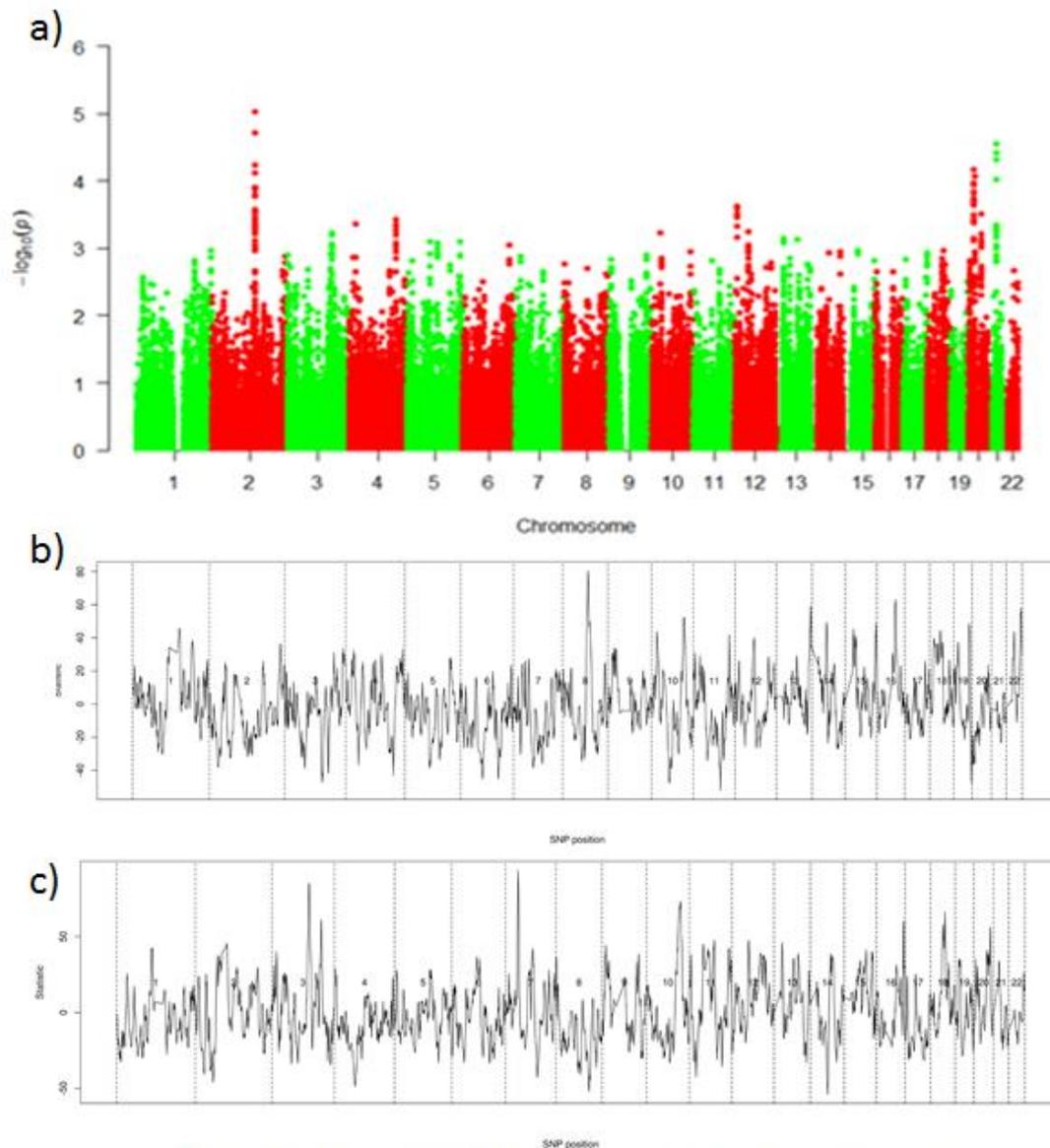


Figure 4.6. (a) XP-EHH analysis of 21 Elderly women from the South Fore linguistic group versus a reference population of 23 individuals who died from kuru below the age of 25. Efforts were made to match the degree of relatedness between individuals with each of the test populations. Relate analysis of the same 21 elderly women (b) and 23 young kuru cases (c). No genomic areas in either Relate analysis reached genome-wide significance after 10,000 permutation reruns with 73 control samples from the same population linguistic group.

The XP-EHH plot above shows regions of extended haplotype homozygosity in the analysis. These are summarised in the table below. Attempts to validate the findings in Relate analysis has not produced any regions of strong identity by descent sharing after performing 10,000 permutations with 72 control individuals from the same South Fore population ($p=0.05$ requiring a score of 135.6).

Of the regions of high scoring in XP-EHH analysis the highest scoring can be confidently ruled of having association with kuru as it is not highly brain expressed and its role is only predicted by computational annotation and should be treated as preliminary. PCSK2 is lowly expressed in the brain (RPKM <5). Brain expression has been used throughout this chapter as a criterion to filter candidate signals of selection. This leaves the possibility of genetic loci expressed in other tissues that may play a more peripheral role in prion disease. The high brain expression accompanied by the primary site of disease being in the nervous system made this assumption worthwhile in order to reduce the amount of false positives in analyses..

CHR	SNP	GENE	Brain Expressed
2	rs4954477	THSD7B	N
21	rs2830165	-	-
20	rs1024175	PCSK2	N
12	rs10774435	NCAPD2	N
4	rs13135284	-	-

Table 4.7 Summary of top scoring variants in XP-EHH analysis.

4.3.3 Chromomatcher analysis

Chromomatcher optimisation analysis shows sensitivity of the tool to detect a strong genome-wide signal for *PRNP* 127V when present at 9% frequency in an analysis. There was considerable weakening of the signal when the frequency reduced from 18% to 9% but still remained highest genome-wide (Figure 4.8). These optimised results were found when minimum length of sharing was set to 2MB and minimum cluster size was 9%. These parameters were then used in subsequent analyses.

Closer inspection of the genomic region around *PRNP* and the variance of Chromomatcher scores revealed a lack of accuracy in the highest scoring variants proximity to the shared novel mutation of haplotype carriers. The analysis with 18% *PRNP* 127V allele frequency reveals an area of high scores extending a region of 4MB in total with the highest scoring variant being over 2MB away from the

PRNP locus. It is a similar problem with in the 9% analysis with a large region extending over 4MB again but this time the highest scoring SNP being <200kb from the *PRNP* locus. Any results in the downstream analysis were judged as haplotype regions based on genomic areas of extended high scoring rather than automatically assuming that the highest scoring SNP was most proximal to any causal variant. This reflects problems of fine mapping variants residing in relatively long haplotypes within a population with already high levels of linkage disequilibrium throughout the genome.

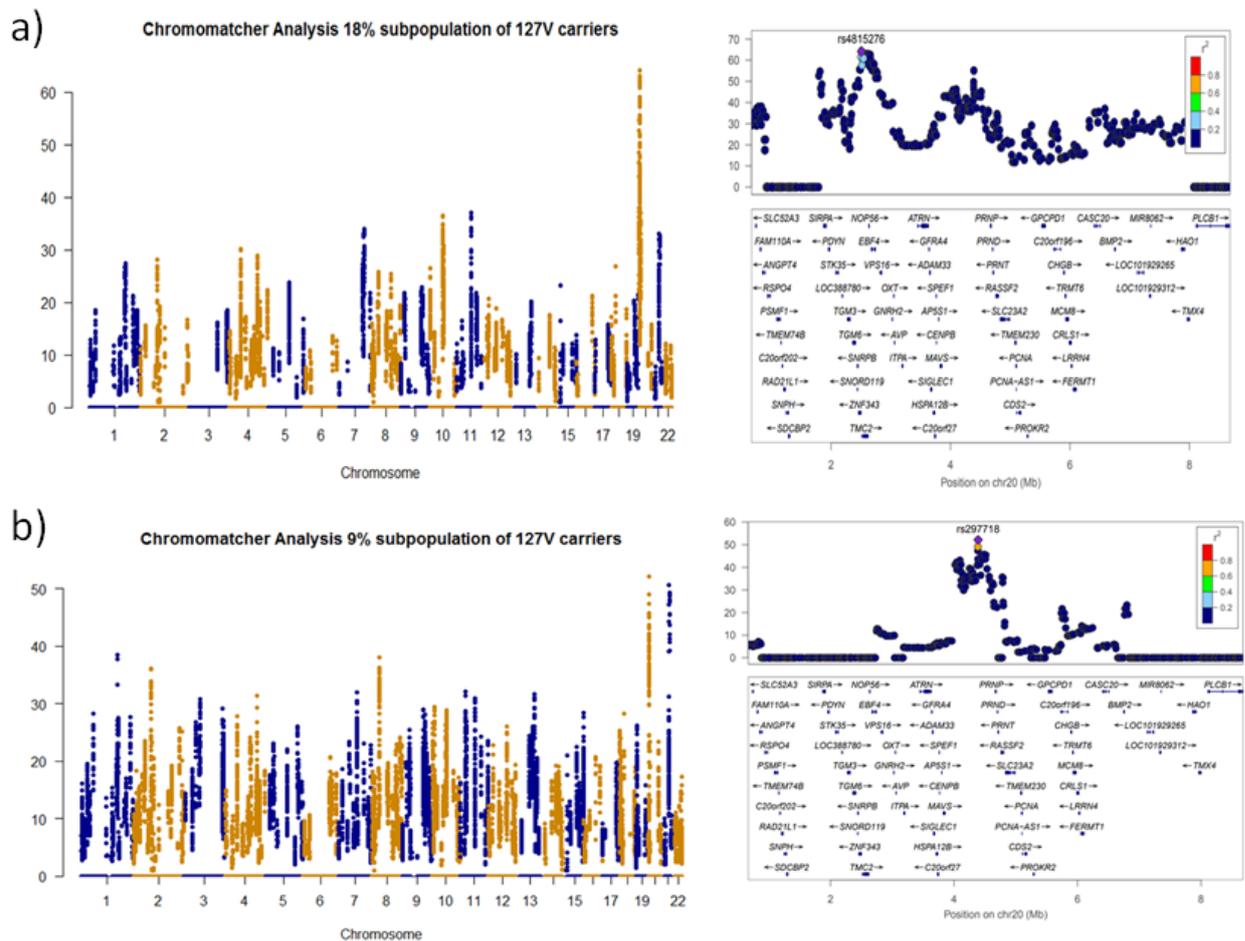
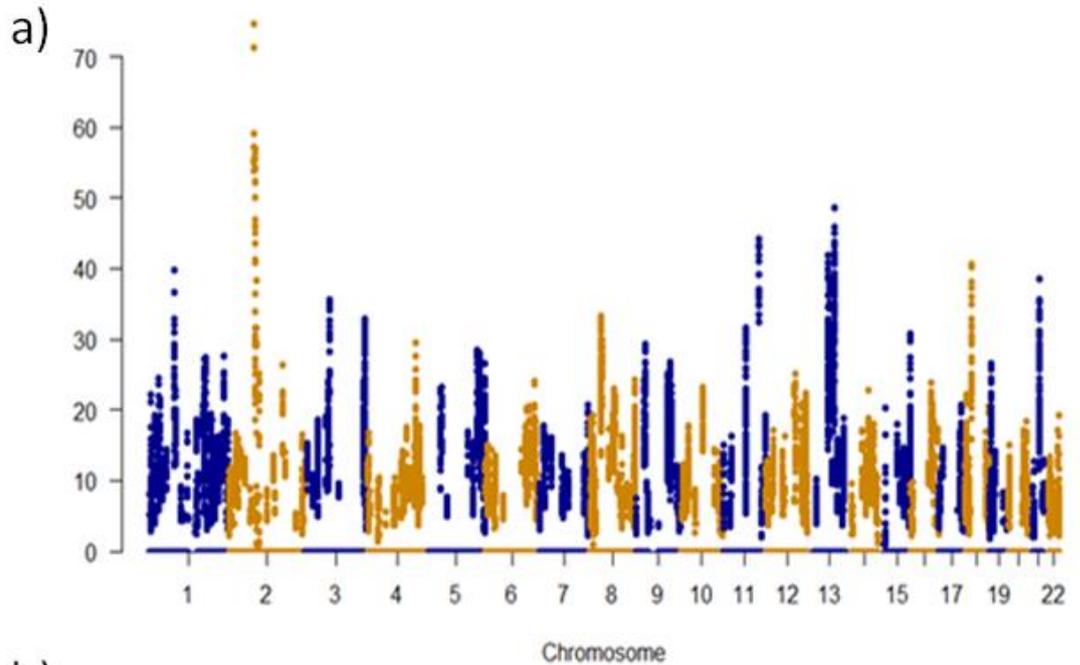


Figure 4.8 Chromomatcher optimisation analysis using *PRNP* 127V carriers to gage sensitivity and optimise parameters. (a) Analysis with 18% frequency of *PRNP* carriers showing high signal on Chromosome 20 of Manhattan plot, and SNAP plot showing extent of high-scoring region on chromosome 20. (b) Similar analysis, this time with 9% frequency of *PRNP* 127V carriers.

Figure 4.9 shows Chromomatcher mean ratio scores for genome wide variants in the Elderly Women analysis. The strategy to eliminate false positive high scoring haplotypes has been based on eliminating signals due to the individuals present having died of kuru disease and the related

genomic regions not containing genes likely to have an effect on kuru (brain expressed genes).

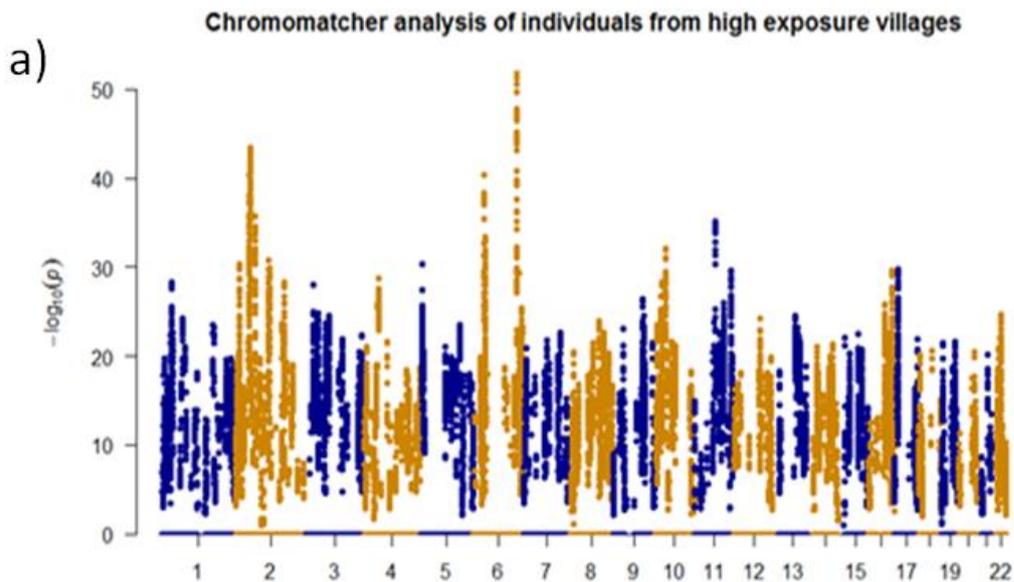
Figures 4.9 and 4.10 show the top scoring regions in both Chromomatcher analyses, the accompanying table highlights signals eliminated for kuru-disease enrichment. If WGS data was available, genes of interest were taken forward to be inspected for specific genetic changes that could be of relevance to kuru in the WGS data analysis (section 4.3.4).



b)

RANK	SNP	SCORE	CHR	BP	n.k	Excessive cases	WGS
1	rs1340093	74.63208	2	79,864,923	11	N	Y
3	rs9546257	48.63579	13	83,460,857	11	Y	Y
4	rs7110543	44.21949	11	113,524,040	11	N	Y
5	rs2649137	41.98406	13	63,179,285	11	Y	Y
6	rs1276322	40.72502	18	20,921,129	12	N	Y
7	rs1952363	39.75918	1	78,743,308	11	Y	Y
8	rs1882766	38.53202	21	36,811,159	12	N	Y
9	rs2593875	35.64453	3	76,749,882	11	N	Y
10	rs2040115	33.36751	8	36,908,104	11	Y	Y
11	rs4965881	30.18666	15	102,024,136	11	N	Y

Figure 4.9 (a) Chromomatcher analysis of 59 individuals comprised of 47 elderly women (kuru-resistant) and 12 individuals who died of kuru. (b) Table of highest scoring variants (right) highlighting the number of individuals in the cluster and whether or not the cluster is enriched for kuru cases.



b)

RANK	CHR	RS	SCORE	BP	n.k	Excessive	WGS
1	6	rs9386147	51.9	146,516,537	41	N	Y
2	2	rs1112523	43.4	49,643,063	41	N	N
3	6	rs7757037	40.4	35,548,236	42	N	Y
4	11	rs948192	35.2	70,439,364	41	N	Y
5	2	rs961328	34.6	67,555,446	42	Y	Y
6	10	rs2505171	32.0	37,323,786	42	N	N
7	5	rs4702484	30.4	7,649,860	52	N	Y
8	17	rs2721833	29.8	12,297,769	45	Y	N
9	16	rs8060741	27.8	79,972,948	46	Y	Y
10	1	rs904257	27.7	31,022,237	41	N	N

Figure 4.10 (a) Chromomatcher analysis of 282 individuals who reside in highly exposed kuru regions. This cohort includes individuals from various exposure and case/control groups. (b) Table of highest scoring variants (right) highlighting the number of individuals in the cluster and whether or not the cluster is enriched for kuru cases.

Analysis	CHR	Range (GRCh38/hg38)	Number of WGS Individuals
Elderly Women Analysis	2	79637547-86880838	4
Elderly Women Analysis	13	80021081-83874717	1
Elderly Women Analysis	11	113283153-113707705	5
Elderly Women Analysis	13	61540588-66157095	5
Elderly Women Analysis	18	22717662-24356807	3
Elderly Women Analysis	1	78255304-80711590	4
Elderly Women Analysis	21	34901629-36797928	2
Elderly Women Analysis	3	76502002-76881613	7
Elderly Women Analysis	8	34339819-38611785	5
Elderly Women Analysis	15	100518480-101556042	6
High Exposure Analysis	6	145027495-146694394	1
High Exposure Analysis	6	33898032-35717928	1
High Exposure Analysis	11	69362029-70606640	1
High Exposure Analysis	5	6832117-8554609	1

Table 4.11 Summary of Chromomatcher high scoring regions taken forward for WGS analysis

4.3.4 WGS analysis

Chr	BP (GRCh38/hg38)	RS	CADD Score	Gene	gNOMAD Frequency	Found in PNG Highland Genomes	Gene Exon Count
5	75669297	rs34358	40	POC5	0.5972	-	11
15	55430684	rs57809907	42	DYX1C1	0.1116	-	8
11	47622719	rs796096347	44	MTCH2	0.007613	N	13
2	152560914	.	42	FMNL2		Y	26
6	26199844	.	40	HIST1H4E	-	N	1
4	121333008	rs1298832736	40	QRFPR	0.0001443	N	6
6	152293740	rs1483590277	56	SYNE1	0.0001259	N	146
6	152293722	.	56	SYNE1		N	146
17	39875380	rs773666085	43	ZPBP2	0.000003983	N	8
17	75757478	rs371616348	42	GALK1	-	N	8
6	56530087	rs1418808051	47	DST	0.0000513	N	95
20	2749751	rs1319487171	44	EBF4	0.00007029	N	16
6	56530096	rs1186380101	43	DST	0.00005005	N	95
1	156768222	rs375857609	40	PRCC	-	N	7
4	176177130	rs1298832736	49	WDR17	0.0001443	N	30
9	98835336	rs751621695	42	GALNT12	0.00001193	N	10
13	44379676	.	42	SERP2	-	N	3
9	93292735	rs1176303443	40	WNK2	0.0008991	N	30

Table 4.12 Summary of High CADD score variants (CADD > 40) in analysis of 30 Whole Genome Sequences of Elderly Women

After applying a CADD score filter of 40 to the analysis of WGS data from 30 kuru resistant individuals from EHPNG, 18 high scoring variants remained. Two of these variants were at elevated frequencies in the GnomAD exome database (prior frequency threshold for elimination was set at >1%) and an additional variant was found in multiple genomes in multiple other PNG groups. The

remaining 15 variants were analysed for brain expression and possible roles in neurodegeneration and cellular process related to prion disease.

The variant SERP2 was found to be intermediately expressed in the brain (33.10 RPKM) and involved in protein folding mechanisms and protein glycosylation within the endoplasmic reticulum which are processes that may affect prion protein function.

The gene SYNE1 was shown in a study of individuals with rare mutations in the gene to produce a range of phenotypes in addition to cerebral ataxia (ataxia is one of the early symptoms of kuru) including motor neurone disease(198). This gene has an exceedingly large protein coding region (146 exons) and is more likely to contain rare variants by chance compared to the majority of genes in the genome. The two variants were found within this gene separated by 18 base pairs and shared by all five elderly women. The proximity of these predicted deleterious variants partly mitigates concerns the gene is present due to ascertainment bias in the analysis.

Two variants were also uncovered in DST gene. This gene has shown to produce sensory autonomic neuropathy in animal model experiments when certain mutations are present(199). Four individuals carry two mutations that are separated by nine base pairs and should be considered as possible candidates for roles in kuru.

CHR	RS	BP	Mutation	GENE	CADD Score	NOMAD FREQ	EW Frequency (N=30)	Highland Frequency (N=25)
6	rs13205210	34864079	missense variant	UHRF1BP1	25.4	0.1128	0.22	0.22
15	rs1385498929	101305265	missense variant	PCSK6	25.4	0.00004096	0.13	0.16
8	rs12541651	37872850	missense variant	RAB11FIP1	21.2	0.6588	0.23	0.24
21	rs3746866	36245426	missense variant	DOPEY2	23.8	0.1542	0.78	0.42
2	rs4247303	85326957	missense variant	TGOLN2	22.7	0.5602	0.93	0.5
2	rs187605466	85551041	missense variant	GGCX	28.9	0.00002785	0.15	0.04
2	rs767834712	86482074	missense variant	KDM3A	34	0.00002785	0.1	0
18	rs779411388	23749484	missense variant	LAMA3	23.8	0.00008019	0.1	0
6	rs17075385	145805261	missense variant	FBXO30, LOC100507557	22.8	0.008235	0.25	0.2
6	rs117586623	145954847	missense variant	SHPRH,SHPRH	20.9	0.008153	0.25	0.18
5	rs6883562	7835442	missense variant	C5orf49	25.8	0.2064	0.33	0.4

Table 4.13 Analysis of high CADD scoring variants (CADD > 20) in high scoring Chromomatcher regions

Table 5 shows high scoring CADD variants in the 14 high scoring Chromomatcher regions that were not eliminated previously due to enrichment for individuals with kuru. Of these, five variants have a

global allele frequency in GnomAD in excess of 5%. These were ruled out for further analysis in kuru as variants at this frequency would have been expected to have been discovered in previous prion GWAS.

Of the remaining 6 variants that are at extremely low frequencies globally in the GnomAD tool (all < 1%) only the candidate in the gene PCSK6 was felt likely to play a causative role in prion disease. The other genes have tissue expressions or are involved in processes peripheral to already known actions and locations of action of the prion gene.

The variant in PCSK6 was then checked for its presence in all PNG WGS data available. The variant was found in all 5 WGS samples from individuals that comprised the high scoring cluster. The variant was also present in three other elderly women in the analysis who were not part of the cluster

CHR	GENE	BP	RS	CADD Score	Gnomad frequency	EW Frequency (N=30)	PNG Highland Frequency (N=25)
7	PDIA4	149,015,000	rs2290971	24.6	0.017	0.15	0.6
20	PRNP	4,699,600	rs267606980	26.9	0.000040	0.083	-
22	GAL3ST1	30,578,841	rs754908238	27.8	0.000088	0.017	-
22	GAL3ST1	30,557,308	rs2267161	22.8	0.30	0.067	-
22	GAL3ST1	30,555,239	rs112976399	22.7	0.0070	0.033	0.2

Table 4.14. Summary of high CADD score variants (CADD > 20) located in four genes shown to have association with sCJD

Results from a European association study of sCJD were cross referenced with the WGS data to see if suggested associated variants for sCJD were enriched in elderly women. This association study has found supporting evidence for the role of STX6 and GAL3ST1 through additional exome analysis and some animal knock down experiments. After filtering for CADD scores >20 in the genomic regions of genes associated with sCJD, 5 genetic candidates remained. One of these in GAL3ST1 was found to be a common polymorphism in GnomAD populations and two others were found in multiple genomes of individuals from other highlands populations and eliminated as likely being related to kuru disease. Of the two remaining one is the already documented *PRNP* 127V variant. The fact this does not appear in any other populations outside of EHPNG further strengthens arguments that it is of recent origin and restricted to the kuru region. The variant that is remaining in PDIA4 after these filters is only present in a single elderly female and would require further

validation in other samples to verify the extent of its presence within disease affected communities and any association between kuru case/control status.

4.4 Discussion

Analysis using the Chromomatcher tool in conjunction with available WGS data has successfully provided a list of genomic regions possibly under selection in the Fore population. Some of these variants may be of interest to investigators in future studies to establish if these signals are a direct result of selection pressure due to kuru. Given the recent nature of the kuru epidemic and the substantial linkage equilibrium observed in the affected populations a strong emphasis was placed on haplotype based approaches as oppose to solely allele frequency based approaches. It was felt this would help focus analysis of findings on selected variants in relation to kuru.

Use of Chromomatcher in this analysis demonstrated its advantages over traditional tools XP-EHH and Relate when applied to a population under very recent selection pressure like the kuru-affected communities. XP-EHH analysis does not provide information on which individuals within a population are driving high EHH scores. There is an implicit understanding that the selection pressure has existed within the population for a large time and positively selected variants have increased to high frequencies as a result. When haplotype homozygosity is driven by a subpopulation in the analysis, Chromomatcher clustering of output immediately reveals which individuals are driving high scoring at loci. This functionality is not possible with Relate which also showed limited power to detect *PRNP* 127V when present at low frequencies in an analysis. Chromomatcher clustering is particularly useful for recent, limited selection pressures like the kuru epidemic where it is unlikely that a recent variant and its associated haplotype background will have risen to a sufficient frequency to be detected by Relate and XP-EHH. The clustering approach allows the inclusion of individuals from different epidemiological cohorts which allow to control for false

positives. This was done with the inclusion of kuru cases in these analyses and the exclusion of regions with enrichment of such individuals.

The strategy undertaken of basing the optimisation of the Chromomatcher tool based on the *PRNP* 127V will have limited the capacity to capture variants with a completely distinct evolutionary history. An alternative variant less recent in origin and residing on multiple haplotype backgrounds would have been far more difficult to detect and would likely have resulted in very heterogeneous clustering of output. Optimisation of Chromomatcher to find variants with longer evolutionary history and heterogeneous presentation will be benefitted by performing a series of simulations that predict the presentation of haplotypes in a population under selection pressure like that exerted by the kuru epidemic and further informed by population demographic factors like those discovered in this investigation. Haplotype based approaches may be of limited value in finding variants with deep origins and instead allele frequency methods would be of more use.

Simulations would have to be performed under neutral selection and selection pressure to generate a distribution of haplotypes like those anticipated in the kuru affected communities with a suitable demographic model. Many demographic models are based on simple assumptions of population size and bottlenecks in human history. This comes in the form of a bottleneck 50-70kya to represent the out of Africa expansion and population increases since. The kuru affected populations in EHPNG have complex population histories and would be expected to have had multiple population bottlenecks which would give rise to the extreme genetic drift and population heterogeneity observed in the demographic analysis.

Simulations would provide limits of Chromomatcher to elucidate disease relevant variants on long haplotype backgrounds. Chromomatcher itself may be of better use in application to other datasets and understanding other recent selection pressures with genetic architectures more suitable for analysis. One can imagine an idealised population with less genetic drift than the Fore population, and experiencing a strong recent selection pressure over a longer period than that of kuru without

the numerous practical and epidemiological challenges of kuru leaving much clearer genetic signals of selection in relation to this selection pressure. A possible example of this may be South American populations that have been exposed to novel pathogens since admixture with Europeans since 1492 in this hemisphere. Chromomatcher may be of best use when populations are still poorly understood and when there are fears of subpopulations masking true allele frequencies of variants during and after selection pressure. There is a clear need for a tool like Chromomatcher that can incorporate phenotypic and epidemiological descriptors of individuals into its analysis. It must be remembered that the Chromomatcher tool is still in its infancy and application to publicly available datasets which contain carriers of well characterised and validated variants that have been under recent selection pressure will help further optimise the tool. Optimisation and further application will reduce the presence of false positive findings due to population structure, ascertainment bias and genomic aberrations.

A lack of sufficient samples and natural genetic heterogeneity in EHPNG prevented the use of a control group in the Chromomatcher analysis. Use of a control group would help eliminate signals of selection driven by selection pressures other than kuru that were experienced in the region historically, or signals driven by population bottlenecks experienced in common ancestral populations. Selecting such a group that is sufficiently similar to the Fore but not affected by kuru is challenging as the differences between groups inside and outside of the kuru region will result in a greater the proportion of long haplotypes presenting in Chromomatcher due to population demographic differences. Additionally it would be ideal to match a sample of individuals in a comparison group to the Fore sample in terms of relatedness and age distribution. Different groups within the highlands have intrinsically higher levels of relatedness due to population history, demography and residential patterns.

Validation of findings in Chromomatcher using the available WGS data from 30 individuals would be enhanced by the addition of a suitable comparison group to further filter false positive results.

Prioritisation of Chromomatcher findings was performed by assessing population allele frequencies globally and geographical distribution of variants within PNG to exclude believed common variants that would have been tagged in either previous sCJD and vCJD association studies or the meta-analysis of kuru performed in this chapter. More refined measures that give more confidence over the population frequencies of suggested variants would allow more optimal prioritisation of possible kuru related variants.

Using CADD scores as a filter for assessing the likelihood of variants in regions playing a role in kuru disease may also have led to elimination of true findings. The CADD score filter may not have been suitably sensitive to incorporate the subtleties in kuru disease processes that the CADD score measure is not equipped to measure. Limitations of CADD scores to have validity for specific diseases have been highlighted previously(200). The *PRNP* 127V variant which has been shown to have a complete protective role against all prion diseases has a CADD score of 26.7. Despite a high functional effect in disease the change in amino acid sequence within the gene is not gauged to be sufficiently impactful to warrant a high CADD score.

Optimisation analyses with *PRNP* 127 revealed this clustering process is not always precise.

Individuals were observed in addition to *PRNP* 127V carriers forming part of the cluster of the highest scoring SNP upstream of *PRNP*. Final filtration of any suggested pathogenic variants by checking their presence in WGS data from individuals that died of kuru (preferably several younger cases with a shorter incubation and assumed higher genetic load) would definitively rule out a candidate in having a strong effect on kuru risk.

Examination of the presence of possible disease causing mutations in 30 individuals heavily exposed to kuru has led to the proposal of multiple genetic variant candidates that could play a causal role in kuru. Variants that appeared extremely rare in global populations were found to be widespread and common in the available PNG genomes. This reflects the distinct population history and demography of the region further with some variants being at high frequencies despite being absent or

exceedingly rare (<0.1%) in all other studied populations. Gaging the frequency of globally rare polymorphisms in PNG populations was limited in this analysis due to the WGS data being limited to 55 individuals from 6 different PNG Highland populations. It could not be stated with utmost confidence if a highly deleterious variant (according to annotated CADD score) residing in one of our genomes was truly rare in our populations, or in fact may be polymorphic and not under negative selection in these populations. Expansion of the set of the WGS data in future studies would improve accuracy of allele frequency estimates and increase confidence in assertions as to whether variants are rare or common. The regular occurrence of finding genes with elevated minor allele frequencies in PNG populations despite being rare in the GnomAD database (<1%) reflects the divergent population histories of these populations further. Previously it was suggested that the elevated *PRNP* codon 129 minor allele frequency in EHPNG was reflective of possible past kuru-like epidemics(178, 179). Examinations of allele frequencies of candidate variants in this study highlight the regularity of population frequency divergence between EHPNG and other global populations.

Future inclusion of kuru affected individuals in the WGS panel of individuals could allow expansion of the analysis to variants with CADD scores much lower and close to the CADD score of *PRNP* 127V (26.9), the known precedent. When a CADD score filter is applied of 30 and pseudogenes are removed, over 950 variants remained in the analysis which would lead to concerns regarding the presence of false positives being suggested and difficulty in effectively prioritising candidate variants. A kuru-case enrichment filter would reduce this number and help focus efforts on candidates more likely to be involved with kuru. The filtration process could also be enhanced by using tools to analyse for gene enrichment of deleterious variants in individuals in disease cohorts compared to controls if a suitable set of control data became available. Two genetic candidates in this study are in large genes, *DST* with 95 exons and *SYNE1* with 145 exons. The genomic average number of exons for a gene in the genome is 8.8(201). Modification of the process to take into account ascertainment of variants from larger genes could be done in a formalised manner to ameliorate this tendency.

Expanding the strategy of finding rare variants that may have played a role in the genetic architecture of kuru disease would prove extremely challenging with the data available. This would require a large WGS dataset of the affected populations in order to ascertain population allele frequencies of variants. The dataset in this investigation of the South Fore linguistic group is 585 individuals which would permit estimation of allele frequencies $<0.09\%$ in the population. WGS strategies although prodigiously expensive in comparison to genotyping approaches would alleviate many of the limitations of this study.

Many of the fine mapping problems that are apparent in the Chromomatcher analysis using genotype arrays in this studies would be alleviated using a WGS approach not requiring that a high scoring variant is 'tagging' a causal variant through linkage disequilibrium. Ascertainment of accurate allele frequencies to a high resolution would result in the possibility to incorporate the use of tools that ascertain genome wide patterns of singleton mutations in a population as a signature of selection(202). These methods have been shown to be more powerful in detecting positive selection than traditional methods based on patterns of linkage disequilibrium and allele frequency differences.

Many unknown elements remain regarding the epidemiology of kuru and may never be known. This includes true exposure to kuru prions. Estimates thus far have been based on date of birth and the historical frequency of kuru in village of residence obtained through interviews with villagers. Kuru surveillance only began in 1957 and estimates of surveillance prior to this have been largely based on interviews with ethnographers and cannot be considered precise in any form. Additionally dates of birth have been estimated based on interviews and linking birth to regional events. Lack of precision over key parameters will place limitations on downstream modelling of the epidemic. Additional epidemiological factors including variation in mortuary practices temporally and spatially will have had an impact on prion exposure. It is unknown if the contribution of differential exposure contributed to variation in incubation periods observed for kuru. Participation in multiple feasts and

increased dosage of kuru prions may have contributed to incubation and eventual development of kuru.

The meta-analysis performed in this chapter combining two independent GWAS studies in order to maximise power did not find genome-wide significant variants at common allele frequencies exerting a strong effect on kuru. Association studies of diseases like kuru with a relatively low sample set are made more complicated by unknown parameters like those mentioned in the introduction and methods section. Simple categorisation of factors may obscure more subtle relationships and dynamics that play a role in kuru pathogenesis. Association analysis using mixed model methods that can incorporate factors in a more quantitative way including incubation, population stratification, and exposure differences could provide more impactful findings.

This would be further complimented by a WGS data approach instead of imputed data that was used in this analysis. Imputation was performed using 1000 Genomes Project Phase 3 data from a multitude of global populations. These populations have quite a divergent evolutionary history and patterns of linkage disequilibrium, and allele frequencies due to genetic drift. This imputation panel would have been unable to impute novel variants that have occurred in the long population history since diversion from common ancestors of member of populations in the imputation panel. It may also have been problematic for the imputation process to accurately impute variants that are common in the South Fore population but rare in global populations. Imputation inaccuracy was partially ameliorated by excluding variants with an info score < 0.9 although this score may not have completely captured inaccuracies in the process. If this is the case then the conclusion of ruling out a common variant with a high effect size after the meta-analysis performed may be erroneous. Even if the whole dataset is not fully sequenced, an enhanced imputation approach with a substantial panel that reflects the genetic heterogeneity in the region could give greater certainty than that provided by the approach in this investigation. This would particular benefits for variants at low allele

frequencies as error rates in imputation in this analysis were clearly visible at minor allele frequencies < 5%.

Approaches like the one taken in this study could prove useful in finding genetic regions under selection for other traits and diseases. This could include specific traits and diseases particular to EHPNG and wider PNG or selection pressures that are exerted in other populations globally leading to convergent evolution processes. Isolated and genetically drifted populations like those in EHPNG with distinct population histories, allele frequencies, novel mutations and patterns of linkage disequilibrium will have distinct genetic loads for traits. This could lead to research in these populations providing novel insights into disease processes not gleaned from studies thus far. Understanding the genetic architecture of various diseases and traits will be of great benefit to the EHPNG region and PNG nation going forward. This may be of acute importance particularly of common western diseases that are now increasing in prevalence in EHPNG with the transition to more western lifestyles and a globalised economy. Demands placed by diabetes, obesity and heart disease will be relatively new to the region. Understanding of population genetic risk as a result of these diseases will permit the allocation of scarce healthcare resources to target diseases most likely to place a strain on communities and public services in the region.

In recent years work has begun to understand the role in variation in epigenetic markers and process in disease processes including neurodegeneration(203, 204). Epigenetic experimental designs have many challenges compared to traditional genetic association studies. Genetic association has often been referred to as a proxy for causality rather than just a correlation(205). This is due to the assignment of alleles and genotypes at birth prior to the presence of disease symptoms and development. Epigenetic studies require additional information to understand if variation in epigenetic markers is a factor that is causal of disease or is in fact a change brought about by disease onset itself. If the latter is the case this does not preclude findings of utility as much can be learned of disease biology and processes through their action on epigenetic mechanisms. With neurological

disease and kuru in particular, epigenetic experimental designs may have limitations due to primary site of disease action being in the brain. The majority of epigenetic studies have been based on retrieving samples from easily accessible tissues including blood and saliva. Epigenetic changes that may be restricted to the site of disease, in the case of kuru the central nervous system will require extraction of DNA from these tissues to observe such changes. Extraction of cells and DNA from the brain and central nervous system is less practical and in the case of kuru impossible due to logistical difficulties, cost and the fact that the overwhelming majority of exposed individuals to kuru are now deceased.

An omission in this investigation was not looking for possible selection candidates in sex chromosomes. Given the observed incidence bias towards females in kuru disease, use of this data in particular would perhaps give a more highly statistically powered analysis on any variants located on the X and Y chromosome.

Difficulties in access to disease specific tissues will also place limitations on obtaining an understanding of the role of somatic mutations on the development of kuru. These studies involve extraction of different tissue samples from different sites and examine the presence of genetic differences between cells. Somatic mutations could additionally explain variation in incubation period of individuals being asymptomatic carriers of kuru prions and prion propagation only initialising when a somatic mutation occurs.

A general assumption when prioritising genetic candidates for possible association to kuru disease was to focus on genes expressed in the brain. This reasoning was based on the neurodegenerative nature of kuru disease and the precedent of *PRNP* being heavily expressed in the brain(34). This assumption may inadvertently lead to the elimination of true associations. The acquired dietary form of kuru may mean that genes more peripherally expressed to the site of neurodegeneration may also play a role in disease pathogenesis. This could include genes relating to protein digestion, and transport of materials across the blood-brain barrier.

It may in the future have to be conceded that variants with large effect sizes on kuru development do not exist outside of the *PRNP* locus. Thus far in studies of other prion diseases including sporadic, variant and inherited forms, no genes with strong effect sizes have been observed outside of the *PRNP* locus (genome-wide associated hits have been found in an unpublished GWAS of sCJD), although such a locus may exist in kuru due to differences in disease processes and its genetic architecture in the specific affected population.

4.5 Summary of findings

- Meta-analysis of kuru does not provide evidence of variants at common population frequencies in the effected region that exert a strong effect on kuru risk.
- Use of a new Chromomatcher tool in conjunction with WGS data in analysis of highly exposed individuals has resulted in candidate variants that may play a role in kuru risk.
- WGS data analysis of highly exposed individuals has found variants that appear to be rare in global populations and other PNG populations. Some of these variants have functional roles in neurological disease and merit further investigation for possible roles in kuru.

4.6 References

1. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
2. Whitfield JT, Pako WH, Collinge J, Alpers MP. Cultural factors that affected the spatial and temporal epidemiology of kuru. *Royal Society open science*. 2017;4(1):160789.
3. Alpers MP. Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3707-13.
4. Collinge J, Whitfield J, McKintosh E, Beck J, Mead S, Thomas DJ, et al. Kuru in the 21st century--an acquired human prion disease with very long incubation periods. *Lancet (London, England)*. 2006;367(9528):2068-74.
5. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A Novel Protective Prion Protein Variant that Colocalizes with Kuru Exposure. *New England Journal of Medicine*. 2009;361(21):2056-65.
6. Collinge J, Alpers MP. Reminiscences and reflections on kuru, personal and scientific. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3613.
7. Anderson W. *The collectors of lost souls : turning Kuru scientists into whitemen*. Baltimore, Md.: Johns Hopkins University Press; 2008.
8. Lindenbaum S. *Kuru sorcery : disease and danger in the New Guinea highlands*. 2016.
9. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nature reviews Genetics*. 2014;15(6):379-93.
10. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Beck J, et al. Genetic susceptibility, evolution and the kuru epidemic. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3741-6.
11. Lindenbaum S. Review. Understanding kuru: the contribution of anthropology and medicine. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3715-20.
12. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. *Briefings in functional genomics*. 2014;13(5):371-7.
13. Mead S, Uphill J, Beck J, Poulter M, Campbell T, Lowe J, et al. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Human molecular genetics*. 2012;21(8):1897-906.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007;81(3):559-75.
15. Liu X, Ong RT, Pillai EN, Elzein AM, Small KS, Clark TG, et al. Detecting and characterizing genomic signatures of positive selection in global populations. *American journal of human genetics*. 2013;92(6):866-81.
16. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*. 2014;31(10):2824-7.
17. Albrechtsen A, Sand Korneliusen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol*. 2009;33(3):266-74.
18. Synofzik M, Smets K, Mallaret M, Di Bella D, Gallenmüller C, Baets J, et al. SYNE1 ataxia is a common recessive ataxia with major non-cerebellar features: a large multi-centre study. *Brain : a journal of neurology*. 2016;139(5):1378-93.

19. Edvardson S, Cinnamon Y, Jalas C, Shaag A, Maayan C, Axelrod FB, et al. Hereditary sensory autonomic neuropathy caused by a mutation in dystonin. 2012;71(4):569-72.
20. Mather CA, Mooney SD, Salipante SJ, Scroggins S, Wu D, Pritchard CC, et al. CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genet Med*. 2016;18(12):1269-75.
21. Hardy J, Scholz S, Evans W, Goldfarb L, Singleton A. Prion genotypes in Central America suggest selection for the V129 allele. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2006;141b(1):33-5.
22. Mead S, Stumpf MP, Whitfield J, Beck JA, Poulter M, Campbell T, et al. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science (New York, NY)*. 2003;300(5619):640-3.
23. Sakharkar MK, Chow VT, Kanguene P. Distributions of exons and introns in the human genome. *In silico biology*. 2004;4(4):387-93.
24. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015;528(7583):499-503.
25. Hwang JY, Aromolaran KA, Zukin RS. The emerging field of epigenetics in neurodegeneration and neuroprotection. *Nature reviews Neuroscience*. 2017;18(6):347-61.
26. Berson A, Nativio R, Berger SL, Bonini NM. Epigenetic Regulation in Neurodegenerative Diseases. *Trends in neurosciences*. 2018;41(9):587-98.
27. Balding DJ. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*. 2006;7(10):781-91.
28. Aguzzi A, Baumann F, Bremer J. The prion's elusive reason for being. *Annual review of neuroscience*. 2008;31:439-77.

Appendix 1 – Sources of data in analyses

Analysis	Section	Figure	Number of Individuals	Name of Primary Dataset
Kuru GWAS	4.3.1	4.4	662	PNG merge
XP-EHH	4.3.2	4.6	44	PNG merge
Chromomatcher	4.3.3	4.9-10	341	PNG merge
WGS Analysis	4.3.4	4.12-14	30	WGS data

Chapter 5 – Exploring polygenic architecture of PD’s and kuru

5.1 Introduction

Work in the previous chapter focused on the search for previously undetected genetic variants that alter susceptibility to kuru. In this final chapter work was performed in an attempt to ascertain if contributions from multiple loci with much weaker effect sizes are detectable for PD’s. Recent developments in this field were outlined in the introduction (Section 1.5.2). Specific hypotheses that were tested in this chapter were;

- Substantial amount of sCJD risk is explained by the action of multiple low effect size variants reflecting a ‘polygenic architecture’.
- Shared polygenic architecture between sCJD and vCJD will be detectable due to molecular and clinical similarities.
- Use of summary sCJD GWAS statistics will provide explanatory power for kuru case/control status reflecting shared genetic aetiology.

Further information pertaining to the particular challenges of applying these techniques to PD’s including kuru is presented below.

Understanding the genetic architecture of traits can give insights into the breadth of biological pathways involved in their manifestation(129, 149). This is of clear importance for PD as much still remains unknown about both the natural function of PrP and also a comprehensive understanding of processes and biological pathways related to PD (see section 1.1.6)(38).

To date, no treatments exist for PD’s which are completely fatal and debilitating. Applying novel techniques to the pursuit of further cataloguing the full genetic architecture of PD’s may lead or contribute to clinical advances in the future. Variants with a small effect on trait risk can be a source

of potential drug targets, as therapeutic agents may manipulate the protein product of gene more drastically than the originally discovered variant effect size may suggest.

Trials of treatments for neurodegenerative diseases including PD's have been hampered by participants being recruited after the onset of symptoms of neurodegeneration. Identification of individuals with particularly high genetic risk prior to onset of symptoms could provide a pool of candidates to recruit into longitudinal trials that could measure the efficacy of drugs prior to onset of symptoms and diagnosis.

Advances in the implementation of polygenic risk score (PRS) methods in recent years have come about largely due to vast increases in the number of participants in GWAS permitted by improved recruitment processes, data processing and the reduction in cost of genotyping samples(206). This is particularly evident with common diseases including cardiovascular disease, Alzheimer's disease and type 2 diabetes. PRS methodologies are seen as a way of explaining a large portion of the 'missing heritability' between heritability explained by associated individual variants and heritability predicted from twin studies and more recent computational methods(207, 208). Similar increases in the scale of PD GWAS are unlikely to be a reality in the near future. PD's are rare in nature with between 100-150 cases of sCJD presenting annually in the United Kingdom(25). PRS based on existing PD GWAS data that provides estimates of variant effect sizes will not be as accurate as for common disease GWAS with more participants and will have an effect on downstream analysis and interpretation of findings.

The utility of PRS techniques also depend on the heritability of the trait in addition to its genetic architecture. Highly heritable diseases with polygenic architecture (e.g. height) are likely to be best characterised by PRS techniques. sCJD in contrast is less heritable in nature due to it being partly defined by the lack of a family history of PD, and an average age of onset of 67(209). Variation in the appearance of sCJD within a population may be due to stochastic neurodevelopmental processes, somatic mutations and epigenetic mechanisms in addition to any genetic variation(210, 211).

Accuracy of PRS techniques have been shown to be effected by population differences between the initial training data set and target data sets that it is applied to. Population stratification has been a factor to contend with for GWAS of sCJD (the only large GWAS of PD conducted to date have been for sCJD). In order to recruit sufficient participants to increase statistical power for detection of common variants in GWAS approaches, large consortia have been established with institutions and public health bodies across Europe and North America. Approaches that involve partitioning this GWAS dataset into training and target sets will have to take such population heterogeneity into account.

Despite the lack of clinical utility for PRS approaches for PD's compared to common diseases (threefold PRS increase for sCJD will not result in lifetime risk increasing to a substantial level), information can be obtained using these approaches to further understand disease processes and risk of individuals. Shared genetic architecture between traits can reveal shared dynamics at the genetic can help highlight potential therapeutic targets or biological pathways to investigate.

The reported controversy in relation to PRS and the current deficit of accurate prediction for individuals of divergent ancestries to the original discovery association data has relevance for PD's. Historical GWAS of PD that form the basis of any creation of PRS have been centred on individuals of European ancestry in order to reduce the impact of population stratification on findings. This means that PRS tailored on these GWAS data will be less accurate in predicting risk in individuals in divergent ancestries. This presents challenges for assessing shared genetic architecture between kuru and other PD's.

With this knowledge in mind, initial efforts applying PRS techniques to PD focused on developing a PRS score for sCJD and checking for any shared aetiology between sCJD, vCJD and kuru on a case control basis. Application to vCJD and kuru were felt to be challenging due to the lack of samples available (vCJD and kuru) and extreme population differentiation between training and target data

set (kuru). Future application of PRS techniques to other aspects of PD including incubation period and associated challenges are discussed in section 5.2.4.

5.2 Materials and methods

5.2.1 Approach taken to investigate PRS of PD

The polygenic nature of sCJD and shared genetic aetiology between sCJD and other PD's was investigated using the application PRSice(144). PRSice operates by calculating PRS in individuals in a 'target' data set. This is done by summing effect sizes at each locus in each individual in the target data based on effect size contributions for SNPs found in GWAS summary data, referred to as the 'base data'. Effect sizes for each SNP in the base phenotype will vary and are represented by a beta value which can be summarised as a value that represents the effect of an increase in genotype at the locus on the phenotype. This is taken from a univariate regression of the base phenotype on each individual SNP.

Some SNPs will not have an effect on the base phenotype and inclusion of these SNPs will not be informative and may lead to erroneous findings and interpretations. PRSice calculates PRS in individuals in the target data sets several times, removing SNP's in the base data set sequentially based on their p-value association for the base phenotype in the base GWAS summary data. The optimum p-value threshold is decided on evaluation of association of PRS and individuals case/control status in the target data set by an appropriate regression model (linear regression for a continuous phenotype and logistic regression for a case/control phenotype). Results of this evaluation are summarised in an R^2 value that represents the ability of the regression analysis to predict the target phenotype in the target data set. For case/control phenotype (which is what was used in the analyses in this chapter) Nagelkerke R^2 is calculated which evaluates the goodness of fit of the regression model and its explanatory power. The use of multiple p-value thresholds to find the optimum collection of SNPs to calculate PRS in the target data set will lead to likely inflation of the R^2 value obtained in the p-value threshold selected. To ascertain if the PRS model produced is

explaining true genetic differences as a result of phenotype in the target data set, PRSice has a permutation procedure that performs the regression analysis multiple times, each time randomly permuting the phenotype labels in individuals in the target data set. This allows the creation of an R^2 value for each imputed regression analysis and the creation of an empirical p-value for the observed R^2 value. This is based on its placement within the distribution of the multiple R^2 values produced the imputation process.

The approach taken in this chapter using PRSice was first to create a PRS for sCJD, the only PD with sufficient sample size to permit such an analysis. This was done by creating a base and target data set by splitting data from a recently performed GWAS by colleagues on sCJD. Eighty percent of individuals in the original GWAS were used to create new summary GWAS data that would be independent of the individuals in the target data set. The specific quality control procedures and parameters used are outlined below in section 5.2.2.

The next stage of the analysis was to evaluate if use of PRSice could provide information on any shared genetic aetiology between sCJD and other PD's. The approach taken was based on a previous analysis performed that looked for shared genetic aetiology between schizophrenia and major depressive disorder (MDD)(144). Inferences about shared genetic aetiology between the two traits were made by using GWAS summary data from a schizophrenia study (base data) and applying it to target data of case/control individuals in a MDD study. The extent of PRSice to provide explanatory power of case/control status in the MDD target data through scoring of PRS based on GWAS summary data of schizophrenia at multiple p-value thresholds was presented as a shared polygenic architecture between the traits. The analysis performed in this chapter provided a similar approach by evaluating any shared polygenic architecture between sCJD and vCJD (see section 5.2.3 for parameters used), sCJD and an unrelated trait (5.2.4) and sCJD and kuru (5.2.5). In each instance summary GWAS data provided from a GWAS performed on sCJD (section 5.2.2) was used as base data and applied to phenotypic data for each of these target data sets.

5.2.2 sCJD GWAS and PRS

Plink data was provided by colleagues from work towards an upcoming GWAS publication in sCJD and vCJD. A GWAS was performed previously in 2012 of all PD's by the same group across multiple PD's(72). A meta-analysis was performed and the only shared locus across all disease categories was found at *PRNP* codon 129. Separate GWAS were performed on sCJD, vCJD, iCJD, inherited PD and kuru resistance despite attending mortuary feasts. The previously published GWAS performed in 2012 for sCJD contained 1,259 cases of sCJD and 6,000 control individuals. In the unpublished GWAS of sCJD this number has increased to 4,110 cases and 13,569 control individuals. Preliminary analysis has revealed multiple loci outside of the *PRNP* locus that appear associated with sCJD. This increase was achieved through the involvement of new collaborators and previously uninvolved countries. The number of variants used in the unpublished GWAS has increased from the previous study (6,214,568 variants compared to 511,862 variants in 2012). This was achieved through imputation of samples using the Michigan imputation server(212). Only variants with an info score >0.9 (test for imputation quality) were used in subsequent association analyses. This permitted a greater intersection of variants between the multiple sources of data that were genotyped on numerous genotyping arrays.

To ensure highest power and accuracy in developing the sCJD PRS, a strategy of splitting this original sCJD GWAS data was employed. Eighty percent of samples were randomly selected and used to produce new GWAS summary statistics to apply to a training set that was composed of the remaining 20% of samples. Ideally this trained PRS would then be tested on a replication cohort from the same population as the training data. This is not possible for a sCJD GWAS given the lack of availability of alternative data.

GWAS analysis on the 80% base sCJD dataset was performed using plink 1.9 (--logistic command). Twenty principal components were added as covariates to the analysis to control for effects as a result of population structure (--pca command in plink 1.9).

The summary statistics were applied to the target 20% sCJD dataset. This was done in PRSice and 20 principal components were added from a PCA analysis using plink 1.9 to control for population stratification in the target dataset.

5.2.3 Shared aetiology between sCJD and vCJD

The sCJD dataset was then applied to a vCJD dataset. This contained 137 individuals with vCJD and 818 control individuals with British ancestry taken from the Wellcome Trust Case Control Consortium (WTCCC). Population covariates were included for this target data set in the form of 20 principal components to control for population stratification.

5.2.4 Shared aetiology between sCJD and an 'unrelated trait'

To check for artefactual population stratification on inflating separation of cases and controls in the sCJD training set, GWAS summary data from a study of a trait felt to have no biological link to sCJD was used. The GWAS summary data chosen was from a study of facial attractiveness on individuals with north European ancestry from the United States(213).

5.2.5 Shared aetiology between sCJD and kuru

Chapter 2 in this thesis outlined extreme population differentiation observed between groups within EHPNG, and extreme differentiation between EHPNG populations and the rest of the world. This means that any application of sCJD PRS developed from a study of European individuals will be greatly affected by population ancestral differences. This problem will extend to interpretations in differences in PRS applied to EHPNG generally and also differences between the various linguistic groups within EHPNG. Any attempted modelling or application of PRSice would be additionally hampered by the lack of samples (89 kuru cases) on which to base the model.

For this reason a limited exploration of PRS differences among EHPNG individuals for sCJD was undertaken. sCJD GWAS summary data again was used as the base data and applied to the *Imputed merge* PNG data set to test if an observable difference existed between kuru-case individuals and resistant individuals. Only individuals within the South Fore linguistic group were used to control for

the strong population stratification present in EHPNG (Chapter 2). Resistant individuals were taken from the Elderly women epidemiological cohort defined in chapter 4. After controlling for relatedness (PLINK1.9 PIHAT < 0.1875), 51 kuru cases of kuru and 59 elderly women were used in the analysis. A summary of the sources of data used in analyses in this chapter can be found in Appendix 1.

5.3 Results

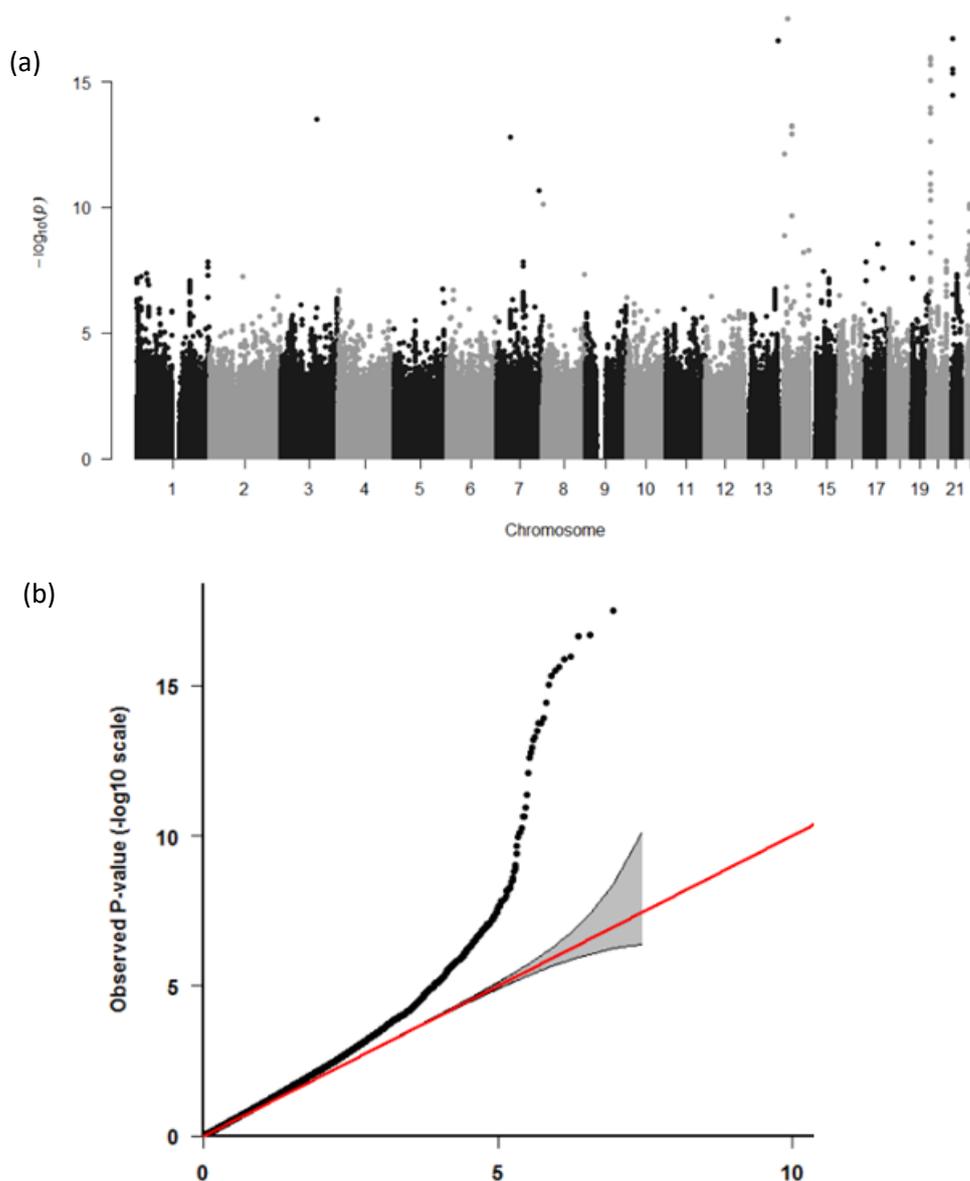


Figure 5.1. (a) Manhattan plot of GWAS of 3,272 cases of sCJD and 10,168 control individuals . The summary statistics from this GWAS were used as the base data for creating a PRS for sCJD. (b) qqplot of sCJD GWAS showing inflation of observed p-values compared to expected p-values. Lambda measure to quantify p-value inflation was measured (1.10).

Figure 5.1 shows manhattan and qqplots for GWAS performed on 3,272 individuals with sCJD and 10,168 control individuals. Lambda was measured at 1.10 between median and observed p-values when chi-squared distributed. Highest associated variants include variants implicated in analysis of the full dataset by colleagues including *PRNP* codon 129. Additional highly associated variants found in this study may be a result of population stratification present.

PRSice has optimal explanatory power (R^2 0.48) classifying sCJD case/control status in a target data set of 818 sCJD cases and 2,542 control individuals (Figure 5.2 a) when a p-value threshold of 0.25 is applied to sCJD GWAS summary stats are used as the base data (104,978 SNP's after clumping). R^2 values for all p-value thresholds were found to be statistically significant (conventional value of $p < 1 \times 10^{-5}$, see appendix 2 for plots of p-values for R^2 values in each analysis). Imputation of case/control status in the target data 10,000 times gave a highly significant empirical p-value ($p < 1 \times 10^{-50}$).

Figure 5.3 shows PRSice results for application of the same sCJD GWAS summary data (that used to produce Figure 5.1) to a vCJD target dataset. A reduced optimal correlation is observed (R^2 0.32) between PRS and case/control status. R^2 values gradually decline at lower p-value thresholds (Figure 5.3 b). A p-value threshold of 1 is optimal with the inclusion of 182,505 variants. Differences between PRS scores of cases and controls were statistically significant ($p = 2.04 \times 10^{-45}$, paired t-test). Imputation of case/control status in the vCJD target data set resulted in a highly significant empirical p-value for the observed data ($p = 9.9 \times 10^{-5}$).

Application of GWAS summary data from a study of a facial attractiveness phenotype to the sCJD target data set of 818 sCJD cases and 2,452 control shows a greatly reduced ability to stratify cases and controls (R^2 0.0029),(Figure 5.4a). Application of different p-value thresholds shows erratic R^2 with an optimal p-value threshold of 0.0085 and 5,349 variants included (Figure 5.4b). A higher R^2 value is obtained at $p = 5 \times 10^{-5}$ but this has a high standard error and the resulting p-value is not significant. No p-value threshold was statistically significant ($p < 1 \times 10^{-5}$). Differences of PRS

between cases and controls did not show statistically significant differences for facial attractiveness perception phenotype ($p=0.26$, paired t-test).

Application of the sCJD base summary data to the *Imputed merger* data set of individual from EHPNG showed no statistically significant explanatory power for kuru case/control status ($p = 0.10$, paired t-test) (Figure 5.5a). An individual was removed after inspection of the first two principal components in a plot. None of the p-value thresholds applied resulted in an R^2 were found to be statistically significant (see appendix 3). Cases were further stratified between young (25 years or younger at the time of death) and old (older than 25 years) to check for any possible polygenic component to kuru incubation. No statistically significant p-value was obtained ($p = 0.88$, paired t-test) (plot Appendix 2).

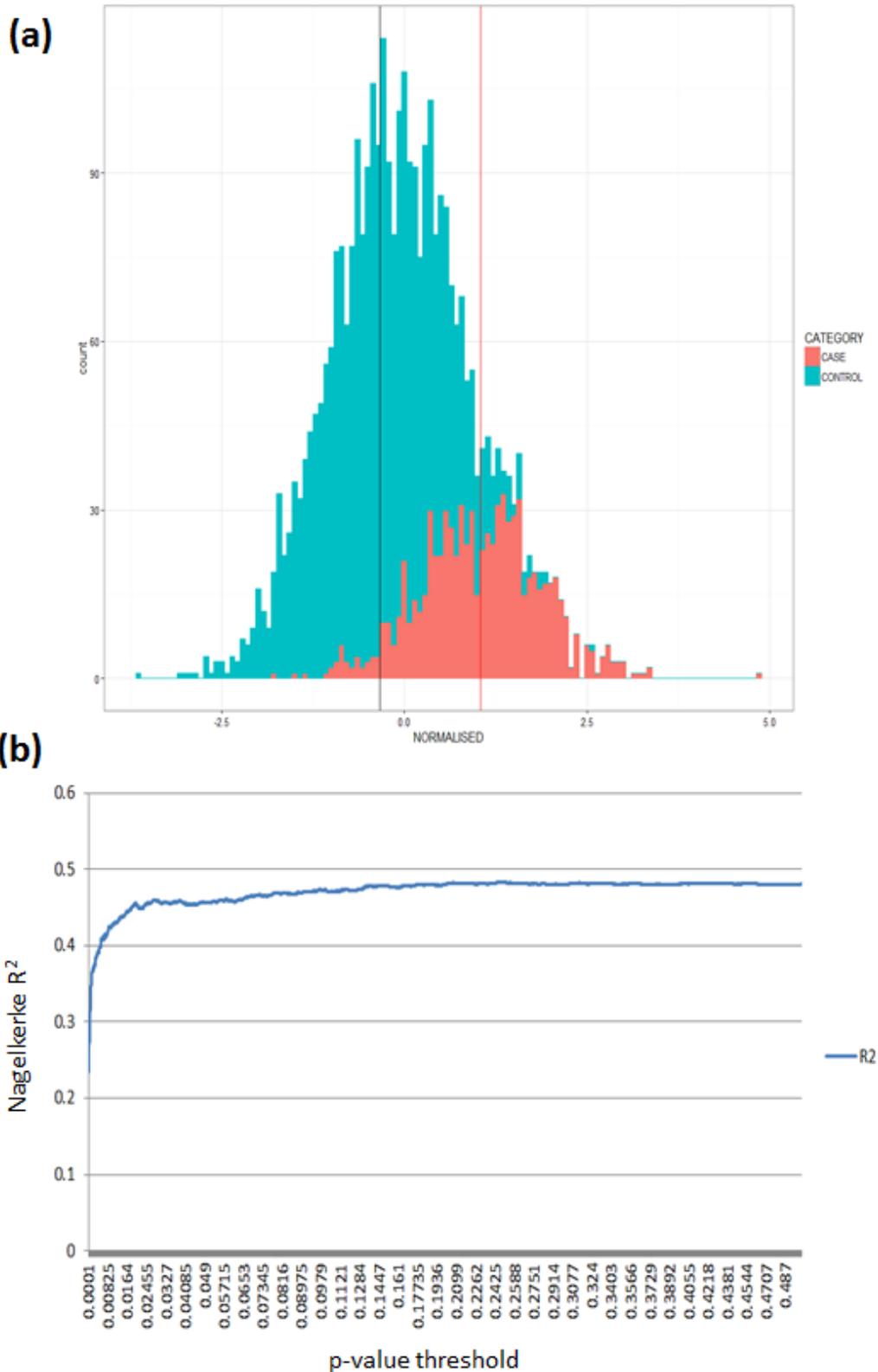
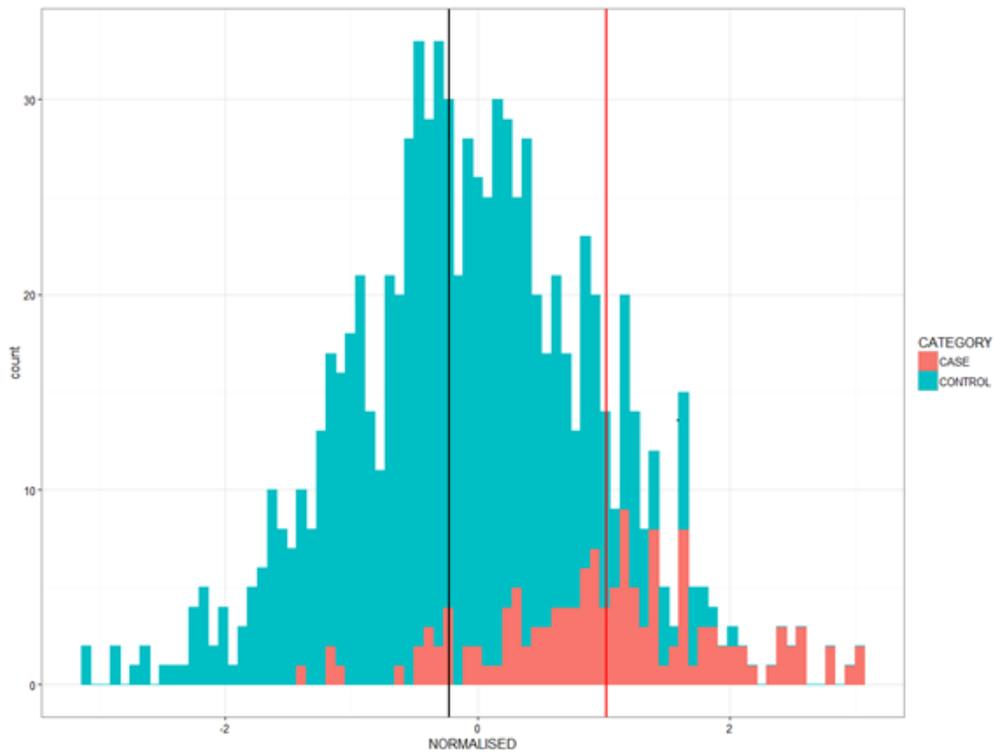


Figure 5.2. (a) Histogram of normalised PRS scores from summary sCJD GWAS data are applied to a cohort of 818 cases of sCJD and 2,542 control individuals. Red vertical line shows the mean of normalised PRS for cases of sCJD and the black line the mean for controls. (b) p-value thresholds show a dramatic decrease in case/control explanatory power at $p < 0.015$.

(a)



(b)

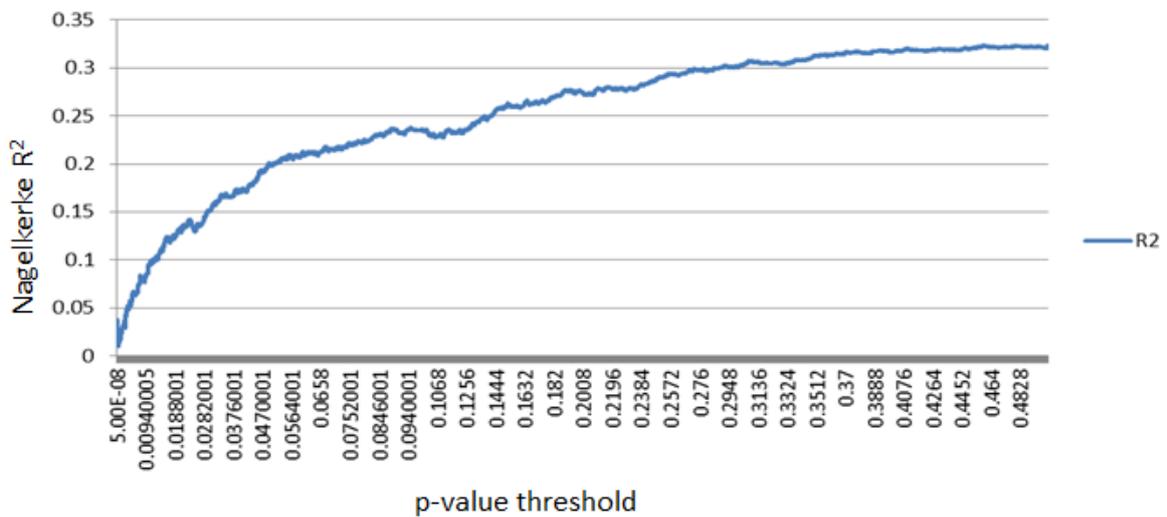


Figure 5.3 (a) Histogram of normalised PRS scores applied to a cohort of 137 cases of vCJD and 729 UK control individuals. Red vertical line shows the mean of normalised PRS for cases of vCJD and the black line the mean for controls. (b) p-value thresholds show a reduction in explanatory power at lower p-value thresholds.

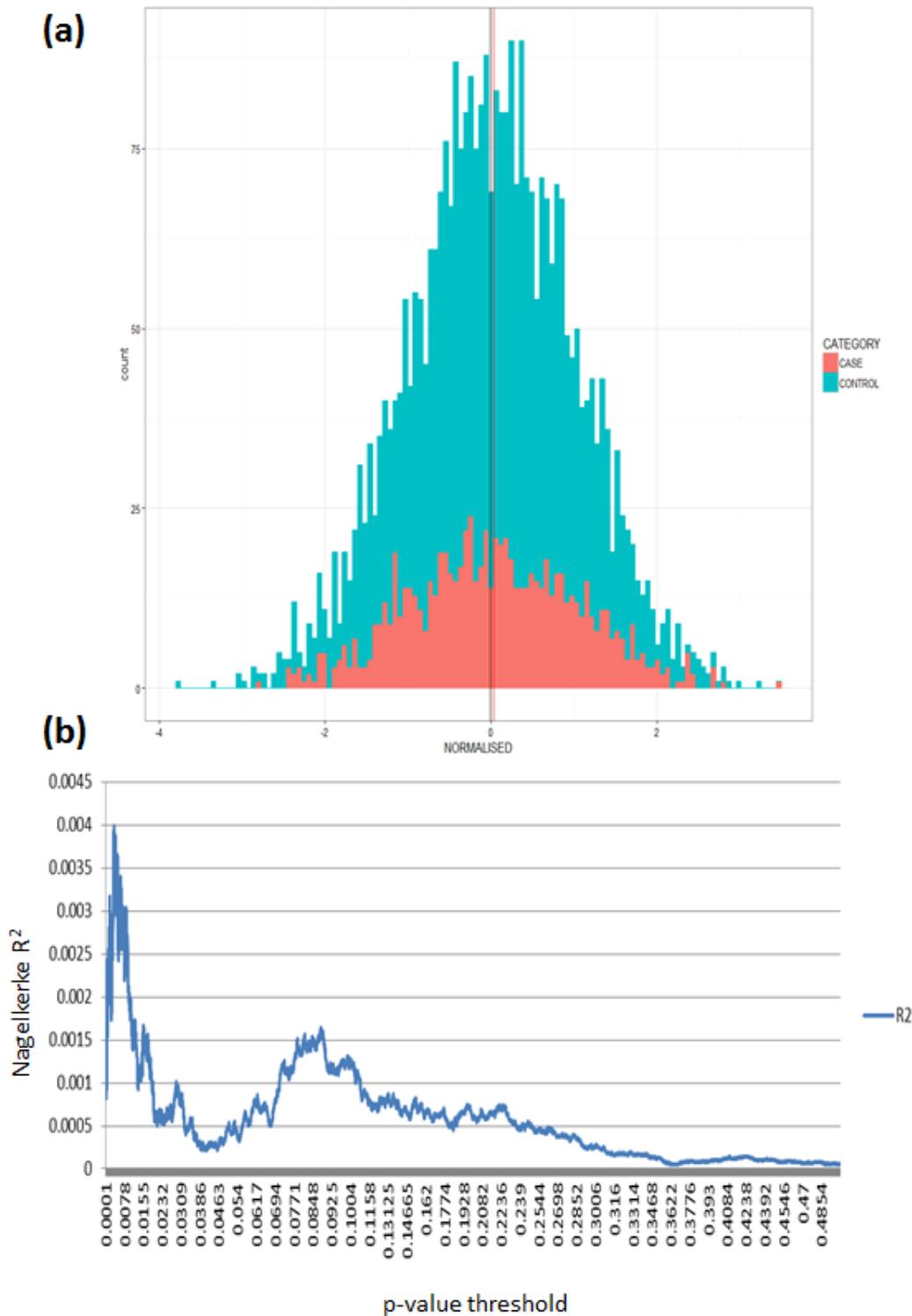


Figure 5.4.(a) Histogram of normalised PRS scores applied to a cohort of 818 cases of sCJD and 2,542 control individuals when a PRS applied from summary GWAS data for facial attractiveness is applied. Red vertical line shows the mean of normalised PRS for cases of vCJD and the black line the mean for controls. (b) Application of different p-value thresholds shows low explanatory power for facial attractiveness sCJD GWAS data to designate case/control status in sCJD case/control target data.

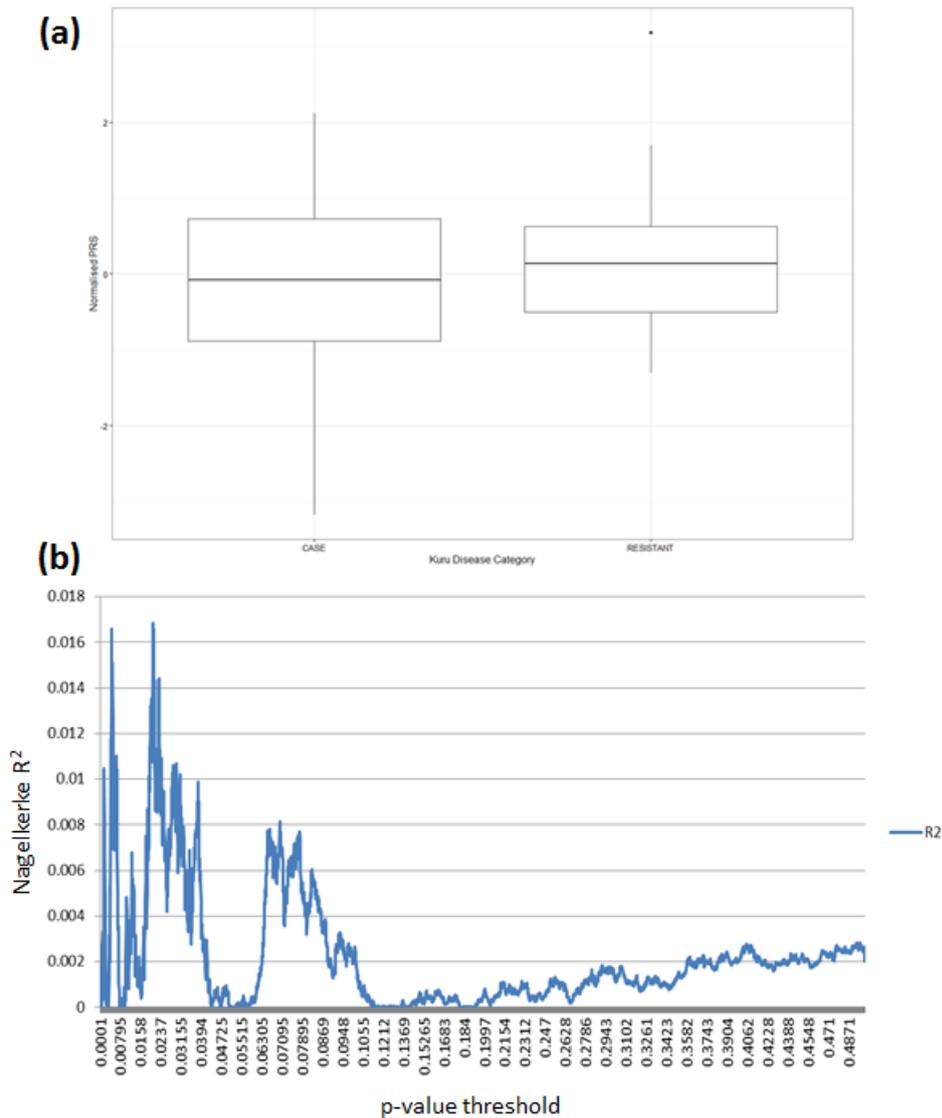


Figure 5.5. (a) Boxplot of normalised PRS scores applied to a cohort of 51 cases of kuru AND 58 kuru resistant individuals shows no statistical significance in average PRS values (b) Plot showing PRSice explanatory power for kuru case/control status when using sCJD GWAS summary base data at various p-value thresholds. Substantial standard errors were observed for all thresholds. The optimum p-value threshold was 0.02 when taking this into account (See appendix 2).

5.4 Discussion

SNP-based heritability estimate tool LDK has estimated heritability for sCJD to be 0.38. So far in published GWAS, only the *PRNP* locus has been associated with sCJD and this association only explains a small fraction of the heritability that has been estimated. Other variants have been implicated in the recently performed unpublished GWAS, but all new variants have modest effect sizes and together will only explain a small amount of the total heritability. Although GWAS for sCJD

have historically been underpowered compared to common disorders the findings in this chapter that variance in sCJD in cases and controls is explained by a polygenic component is not surprising.

The PRS model ability to predict case/control status may have been inflated by population stratification in the analysis. The GWAS performed on the base data in this analysis showed evidence of inflated observed p-values compared to what would be expected under neutrality. Lambda measurement confirmed this with a value of 1.10. Although some population stratification will inevitably be present in any analysis with individuals with different ancestral histories, a lambda value of 1.10 reflects a substantial component of population stratification. Some degree of caution should be taken when interpreting the R^2 values obtained in this analysis. This issue would have been compounded by any population stratification in the sCJD target data set. In this dataset an association analysis was performed to confirm this and again a lambda value of 1.10 was obtained. This issue was highlighted in a study that showed a previous finding of polygenic adaptation to height in Europeans was completely explained by population stratification in the analysis(214).

Population stratification could be further improved in the future with use of a mixed model based approach to generate sCJD GWAS summary data. This approach has been shown to ameliorate this greatly and could provide assurances over the contribution of population stratification to case/control classification in the target data set and any interpretations made. Additional covariates could be added to improve the analysis including sex and age(215).

Distribution of PRS for cases and control individuals in the dataset confirms the expected sufficient lack of distinction to envisage any clinical, diagnostic merit to applying this approach in healthcare settings (Section 1.6 has further discussion). Individuals in the analysis of sCJD have PRS values more than 3 standard deviations away from the mean. Lifetime risk for sCJD has been estimated at 1/5,000, meaning individuals with this increased genetic risk are still not likely to be at substantial risk. This level of risk is still at the level of a rare disease like Huntingdon's disease in the general UK population and would not merit drastic changes in lifestyle or medical intervention.

Discovery of a significant polygenic component to the genetic architecture of sCJD could inform future strategies for GWAS design and implementation. To date, great efforts and resources have been expended in forming international collaborations in order to obtain sufficient numbers of samples of sCJD to create sufficiently powered studies to find common variants with moderate effect sizes. The realisation that a large proportion of the genetic architecture and heritability of sCJD could be caused by the multiplicative effect of many weak variants could place a realistic ceiling on the ability of GWAS approaches to further understanding of the contribution of single loci to sCJD association.

Application of the sCJD base data to vCJD reflects a stability of PRS for sCJD to predict case/control status across these traits. Any shared genetic aetiology may not be surprising given similarities at the molecular and phenotypic level (See section 1.1 for more on PD's). Additional genetic factors may play a role in vCJD due to the dietary mode of prion acquisition, the more prolonged disease course, and different sites of action in the brain. It is difficult to know from this analysis alone if differences in the ability to explain case/control status in vCJD case/control target data can be attributed to differences in the diseases or is a result of differences in the two analyses. Fewer individuals were used in the vCJD analysis, and the vCJD analysis was comprised of individuals from the United Kingdom only. These two factors would have impacted on the model and PRS scoring to some degree. Population stratification in the vCJD target data was lower compared to the sCJD target data due to individuals being from the United Kingdom only.

One can speculate to the underlying bases to such a shared genetic architecture between sCJD and vCJD. The regular biological role of *PRNP* in mammals is still not fully understood(1, 11). Its high expression in the peripheral and central nervous systems have led to proposed functions in these areas as well as possible roles in memory and sleep regulation(216). Model organism experiments involving the modulation and knockout of *PRNP* expression have implicated the role of other

paralogue genes in *PRNP* function. This suggests *PRNP* being involved in multiple complex, dynamic pathways which may be effected by expression at multiple loci.

The inability of the summary GWAS data from a separate GWAS using facial attractiveness as a genotype to classify case/control status or derive statistically significant PRS values based on sCJD case/control status in target data helps remove some of the concerns arising from population stratification present in the study. The facial attractiveness GWAS was performed using 4,883 samples using a quantitative trait measure instead of the categorical one used in the sCJD, vCJD and kuru analyses. These differences may have affected the regression analysis and PRS scoring, although the degree of difference observed does provide confidence that the sCJD and vCJD analyses results reflect genuine shared genetic architecture between the traits.

Application of sCJD GWAS summary data to the *Imputed merger* data set does not provide any evidence for any shared polygenic component between sCJD and kuru. Given the extent of overlap seen between sCJD and vCJD and the similarities observed between kuru and both diseases, it is unlikely that a polygenic component does not exist for kuru. The lack of observation is more likely due to extensive challenges in applying PRS methodologies across such divergent populations in addition to the small number of individuals used in the analysis.

Acquisition of infectious prions in kuru through dietary exposure presents problems when applying PRS techniques based on sCJD GWAS summary data. A complex interplay of cultural and demographic factors contribute to exposure to kuru prions. This, in combination with other unknown factors about exposure to kuru prions including dosage of material and variation in mortuary feast practices, make modelling disease exposure challenging. An approach that could further incorporate these factors into an analysis beyond the case/resistant individual analysis performed in this chapter may be able to leverage information about the broader genetic architecture underlying kuru.

Previous use of PRS technology to predict risk for traits in diverse populations has resulted in large differences in average PRS for different populations(146). Many of these differences seem implausible given the traits in question and the relatively small amount of genetic diversity and differentiation between human populations, even on different continents. This issue is encapsulated with exploratory attempts to apply sCJD PRS to kuru case/control data.

Other difficulties when comparing between divergent populations will be differences in environments which will have impacts on the heritability of the trait. This can be illustrated when comparing a comparison of GWAS results for elevated cholesterol between Uganda and the United Kingdom(139). A large component of environmental variation in relation to this trait is variation in diet and in particular dietary consumption of lipids. Diets in Northern Uganda where one of the GWAS was conducted are less varied than in the United Kingdom which means that heritability of the trait will be expected to be higher in Uganda. Environmental contributors to prion diseases and sCJD are poorly understood although postulated environmental risk factors have included exposure to blood transfusions, and surgical procedures(217). The level of exposure to environmental exposures between variables such as these will vary between populations and in particular between the United Kingdom and PNG.

Past studies of kuru have been previously motivated by the possibility of kuru acting as a model for understanding and predicting the likely incidence of vCJD in affected populations during the BSE crisis. This work was largely performed based on understandings of levels of incidence of kuru and population exposures to the infectious agent. If this understanding could be furthered by developing polygenic population screens for risk for kuru stratified by exposure and outcome groups. Findings from such approach could potentially be applied to other populations to assess population wide risks of developing PD under scenarios of other orally acquired or similar kuru-like epidemics.

Little is still known about the great variance in incubation period that is apparent in both iatrogenic CJD, vCJD and kuru after exposure to infectious prions(23). This is further complicated for vCJD and

kuru where date of exposure is not precisely understood(27). *PRNP* codon 129 genotype has been associated with longer incubation periods in iatrogenic CJD (Section 1.1.4). In vCJD there has only been a single case with a non-homozygous methionine genotype at *PRNP* codon 129 (Section 1.1.5). It is likely that other genetic factors will play a role in the greatly observed variance in incubation period for these three diseases. Expanding PRS approaches beyond simple case/control phenotypic categorisation of individuals could in future inform more about variation in these areas of PD development.

There has been evidence for shared aetiology between PD and other neurodegenerative diseases(217, 218). Application of PRS methods to ascertain the extent of shared polygenic architecture would be of interest and of possible future clinical utility. Approaches could involve the application of GWAS summary data from previously conducted studies of neurodegenerative diseases to target data comprised of PD case/control individuals. PD GWAS summary statistics could also be applied to target data of individuals with other neurodegenerative diseases and controls.

An important avenue of future work will be to replicate the findings in this chapter with an independent methodology. This could help rule out the possibility of artefactual elements of the particular software contributing to results(219).

5.5 Summary of findings

- PRSice successfully classifies case/control status in both sCJD and vCJD target data using sCJD GWAS summary statistics. This provides evidence of a polygenic architecture between both traits and a shared genetic aetiology between the two.
- Application of the model to the *Imputed merger* containing kuru case/control individuals data set shows no evidence of shared genetic aetiology with the approach used in this chapter.

5.6 References

1. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*. 2017;19:110.
2. Mackay TFC. The Genetic Architecture of Quantitative Traits. *Annual review of genetics*. 2001;35(1):303-39.
3. Vázquez-Fernández E, Vos MR, Afanasyev P, Cebey L, Sevillano AM, Vidal E, et al. The Structural Architecture of an Infectious Mammalian Prion Using Electron Cryomicroscopy. *PLOS Pathogens*. 2016;12(9):e1005835.
4. Baranzini SE. The era of GWAS is over - Commentary. *Mult Scler*. 2018;24(3):260-1.
5. Verweij KJ, Mosing MA, Zietsch BP, Medland SE. Estimating heritability from twin studies. *Methods in molecular biology (Clifton, NJ)*. 2012;850:151-70.
6. Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. *Nature genetics*. 2017;49(7):986-92.
7. Budka H, Will RG. The end of the BSE saga: do we still need surveillance for human prion diseases? *Swiss medical weekly*. 2015;145:w14212.
8. Schelzke G, Kretzschmar HA, Zerr I. Clinical aspects of common genetic Creutzfeldt-Jakob disease. *European journal of epidemiology*. 2012;27(2):147-9.
9. Manjrekar J. Epigenetic inheritance, prions and evolution. *Journal of genetics*. 2017;96(3):445-56.
10. Alzualde A, Moreno F, Martinez-Lage P, Ferrer I, Gorostidi A, Otaegui D, et al. Somatic mosaicism in a case of apparently sporadic Creutzfeldt-Jakob disease carrying a de novo D178N mutation in the PRNP gene. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2010;153b(7):1283-91.
11. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*. 2015;31(9):1466-8.
12. Mead S, Uphill J, Beck J, Poulter M, Campbell T, Lowe J, et al. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Human molecular genetics*. 2012;21(8):1897-906.
13. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nature genetics*. 2016;48(10):1284-7.
14. Hu B, Shen N, Li JJ, Kang H, Hong J, Fletcher J, et al. Genome-wide association study reveals sex-specific genetic architecture of facial attractiveness. *PLoS genetics*. 2019;15(4):e1007973.
15. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*. 2018:355057.
16. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017;95:1.22.1-1..3.
17. Ironside JW, Ritchie DL, Head MW. Prion diseases. *Handbook of clinical neurology*. 2017;145:393-403.
18. Collinge J. Prion diseases of humans and animals: their causes and molecular basis. *Annual review of neuroscience*. 2001;24:519-50.
19. Wulf M-A, Senatore A, Aguzzi A. The biological function of the cellular prion protein: an update. *BMC Biology*. 2017;15(1):34.
20. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric genetics*. 2018;28(5):85-9.
21. Kuchenbaecker K, Telkar N, Reiker T, Walters RG, Lin K, Eriksson A, et al. The transferability of lipid loci across African, Asian and European cohorts. *Nature communications*. 2019;10(1):4330.
22. Castellani RJ, Perry G, Smith MA. Prion disease and Alzheimer's disease: pathogenic overlap. *Acta neurobiologiae experimentalis*. 2004;64(1):11-7.

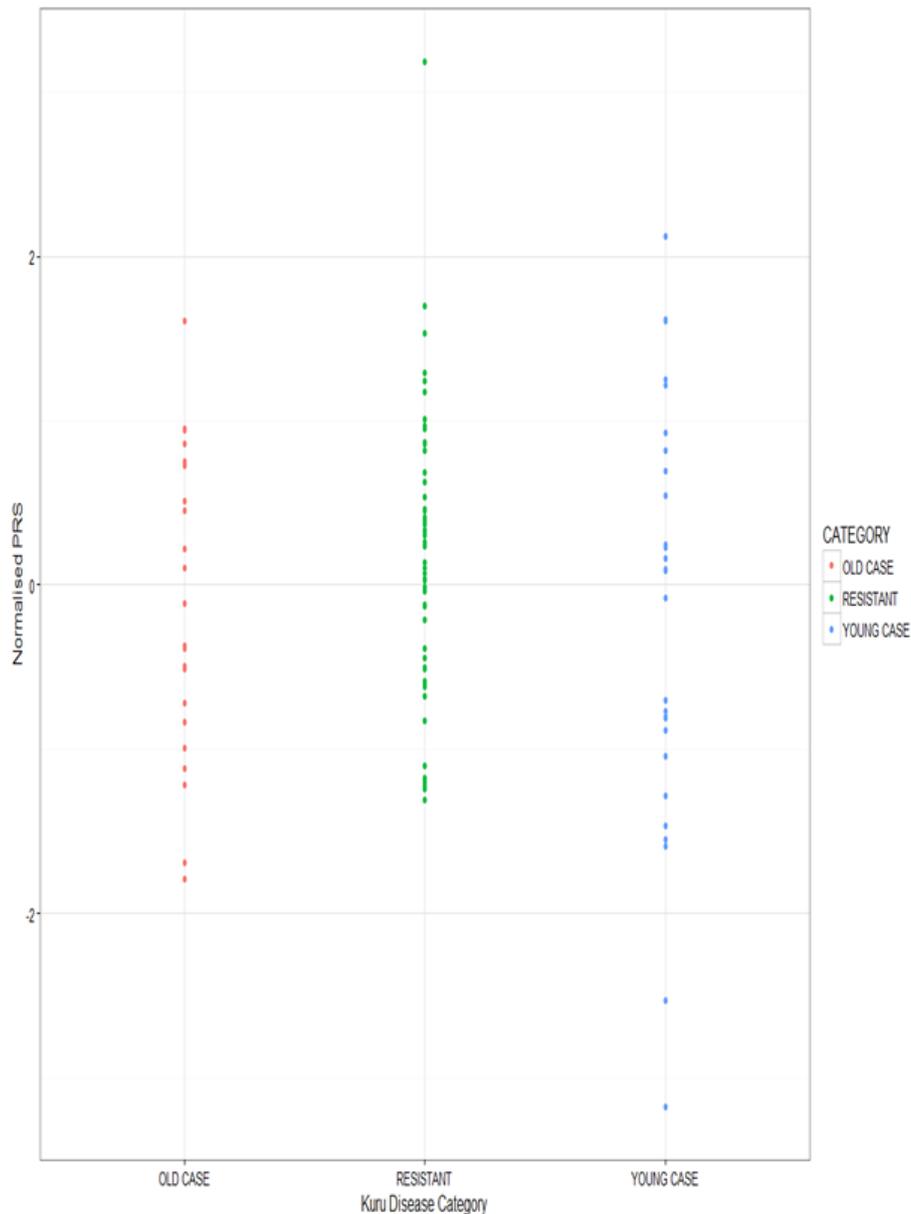
23. Bonda DJ, Manjila S, Mehndiratta P, Khan F, Miller BR, Onwuzulike K, et al. Human prion diseases: surgical lessons learned from iatrogenic prion transmission. *Neurosurgical focus*. 2016;41(1):E10.
24. Collinge J. Variant Creutzfeldt-Jakob disease. *Lancet (London, England)*. 1999;354(9175):317-23.
25. Zhou J, Liu B. Alzheimer's disease and prion protein. *Intractable Rare Dis Res*. 2013;2(2):35-44.
26. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American journal of human genetics*. 2015;97(4):576-92.

5.7 Appendices

Appendix 1 – Sources of data used in analysis

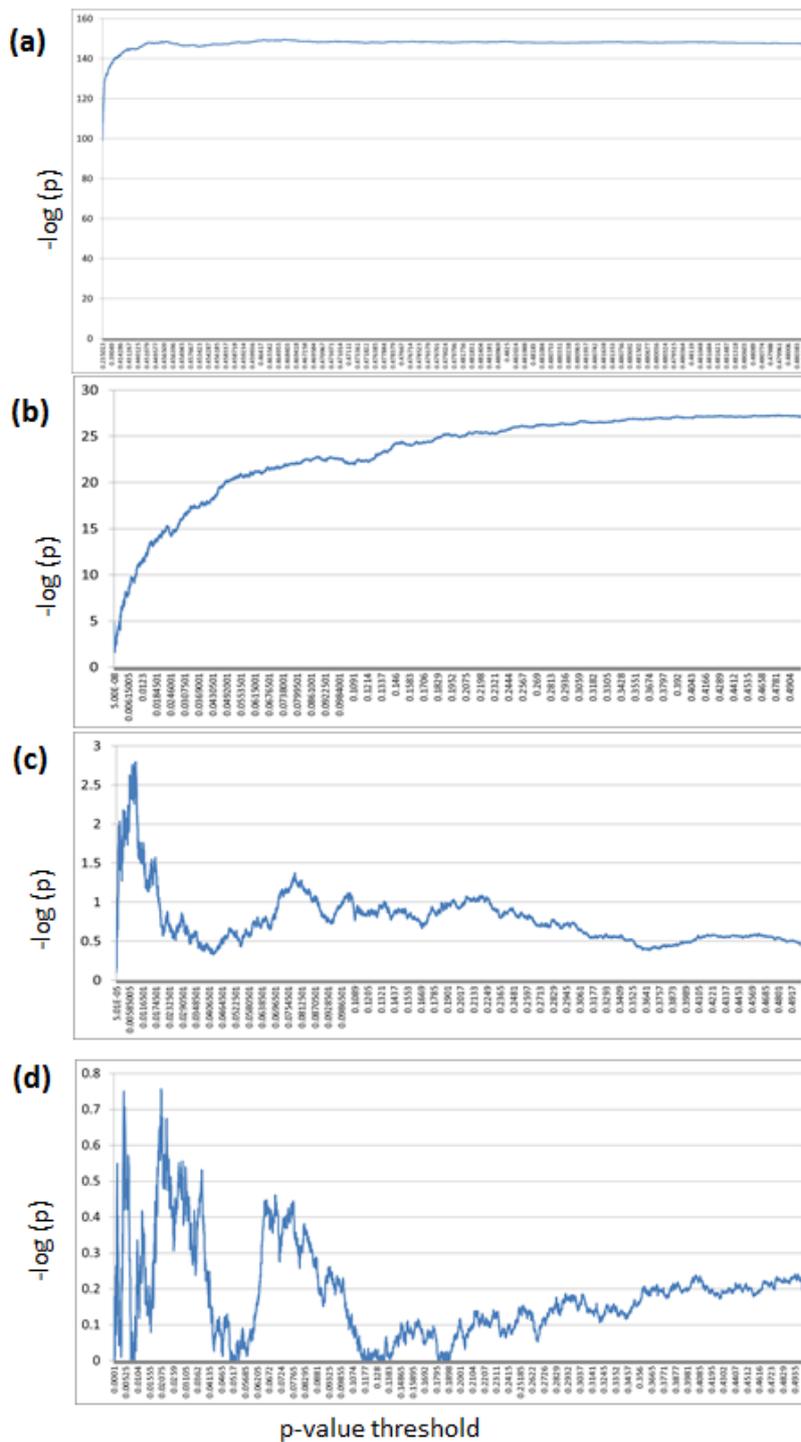
Analysis	Section	Figure	Number of Individuals	Name of Primary Dataset
Split sCJD GWAS	5.3	5.1	13440	sCJD Data
sCJD PRS	5.3	5.2	3360	sCJD Data
vCJD PRS	5.3	5.3	866	vCJD data
facial attractiveness PRS applied to sCJD	5.3	5.4	3360	GWAS catalog
Kuru PRS	5.3	5.5	109	PNG merge

Appendix 2 – Scatter plot of sCJD PRS applied to South Fore kuru cases (young and old) and kuru resistant individuals



Boxplot of normalised PRS scores applied to a cohort of 51 cases of kuru AND 58 kuru resistant individuals.

Appendix 3 – Plots of R^2 versus $-\log p$ values for p value thresholds for four PRSice analyses



PRSice p-value thresholding using sCJD GWAS summary data applied to (a) sCJD target data, (b) vCJD target data, (c) applying facial attractiveness summary GWAS data to sCJD target data and (d) applying sCJD GWAS summary data to kuru case/control data from *Imputed merger* data

6 Summary and Future Work

6.1 Summary

Work in this doctoral thesis was based on a novel dataset of 943 individuals from EHPNG. This has allowed the investigation of the kuru epidemic using the latest computational methods available. Extensive work has been performed in other fields of study and incorporation of these findings has been informative to this investigation(48, 59, 75, 100, 104, 151, 158, 220). A general approach has been to use major findings from other fields, particularly epidemiology, anthropology and linguistics, to construct hypotheses that could be tested using genetic data.

An understanding and awareness of methodologies and practices of population genetics, genetic epidemiology, evolution and natural selection was necessary in order to approach the major research questions rigorously. Additionally, given the very specific and bespoke populations at the centre of this study, these understandings had to be adapted to the specific setting. This was aided by understanding of approaches taken in investigations of other population isolates(221-223).

The major findings in this thesis are summarised below:

1. *EHPNG shows a distinct population structure*

Linguistic groups are clearly differentiated from one another genetically. Differences between groups and communities especially linguistic, geographical and cultural appear to be huge drivers of the amount of differentiation between groups in the region. Complex systems of migration that have been observed from field studies(101, 104) and interview of individuals in the region was confirmed through analysis of CP ancestry profiles of individuals. The presence of 'isolates within a population isolate' are apparent genetically with groups that had been noted for their differing linguistic and cultural practices with notable divergent genetic ancestry profiles compared to other groups in the region(54, 224).

2. *Kuru resulted in an observable impact in patterns of genetic diversity of the affected communities.*

Kuru claimed the lives of ~2,400 individuals over the two decades of highest incidence in communities with a standing population of ~40,000(51). A reduction in genetic diversity in the most affected South Fore linguistic group was observed through time-dependent analysis of genetic diversity.

3. *Analysis of individuals highly exposed to kuru has revealed potential genetic candidates in relation to kuru predisposition and protection.*

Application of a novel technique called 'Chromomatcher' that aims to capture variants under recent positive selection has revealed potential candidates under selection in relation to kuru. Development of Chromomatcher and interpretation of its output reveals the challenges of looking for long, shared genomic regions of epidemiological relevance in population isolates. These populations harbour numerous sites of this nature due to the particular demographic parameters. This novel technique was complimented with standardised and established approaches to prioritise findings including GWAS, XP-EHH, Relate and WGS data analysis(160, 185, 193).

4. *Presence of shared polygenic architecture between sCJD and vCJD. Extensive challenges when applying PRS methodology to kuru*

Use of PRSice demonstrated an ability to explain case/control status in sCJD case/control data through the combined effect of thousands of variants taken from sCJD GWAS summary statistics. This explanatory power was retained when the same sCJD summary statistics were applied to vCJD case/control data, reflecting a shared polygenic architecture between the traits. A similar approach applying sCJD summary statistics to kuru case/control data did not

reveal any shared polygenic architecture between sCJD and kuru. Further work to validate and better understand shared genetic architecture between the traits will be of value.

The work in this thesis presents attempts to obtain clearer understanding of a unique set of communities undergoing a remarkable transformation in their way of life and in the process of enduring a cataclysmic disease epidemic. Kuru had observable effects on community relations, belief systems and the demographic composition of villages at the epicentre of kuru. It is hoped the work in this thesis can act as a blueprint for others attempting to characterise isolated populations and applying these findings to specific epidemiological and evolutionary research questions that further understanding of disease and its legacy on affected communities.

6.2 Future work

In the sections below I would like to firstly discuss specific areas to address to further specific research questions posed in this thesis and then a broader discussion of future possible directions of kuru research and beyond.

6.2.1 Suggestions for development of investigations performed in this thesis

Limited time and resources naturally place restrictions on any doctoral study. Limitations of methodologies and available data have been discussed at length in individual chapters. Immediate priorities for future investigators attempting to advance understanding of kuru and in particular attempts to find specific genetic variants that affect kuru risk include:

- Laboratory assay of candidate variants suggested in Chapter 4. Genotype assay of candidate variants from DNA samples of the entire data set will allow a statistically robust examination if candidate variants are associated with kuru disease status. Associated variants could be further investigated using model organisms, cell cultures and computational analysis of variants on prion propagation processes.

- Use of X-Chromosome data to assess if kuru resulted in a sex-biased reduction in genetic diversity.
- Genome sequencing of individuals who died of kuru at a young age to act as a control group when filtering potential candidate variants.
- Use of mixed-model approaches to GWAS that will reduce concerns over population stratification.
- Genome sequencing of several hundred individuals to form part of an improved imputation panel. Analysis of imputation based on a panel comprised of 1000 genomes project individuals showed increased inaccuracies for imputation of variants at lower population allele frequencies in EHPNG. Use of a population specific imputation panel could help address this concern.

6.2.2 Future avenues of kuru research

6.2.1 Refinement of Chromomatcher tool

Refinement of the Chromomatcher tool used in this research project would help further understand the genetic signature of selection pressures like kuru. Sizeable challenges exist when studying population isolates with highly divergent ancestral histories and extreme parameters for genetic drift and heterogeneity between groups. Chromomatcher the ability to find clusters of individuals within a population with exceptionally long, shared haplotypes. An ongoing challenge is to distinguish which of these have arisen to a relatively high frequency in a short space of time due to selection or due to chance demographic factors like drift. Characterisation of the positively selected *PRNP* 127V variant has further highlighted these challenges. Understanding the lower limits of sensitivity Chromomatcher will be helped by a comprehensive simulation approach that will take into account a multitude of factors including the time of origin for a variant, its frequency prior to selection pressure and the multitude of demographic factors and epidemiological parameters that would influence an analysis. Chromomatcher will be suitable for analysis of other populations under recent selection pressure.

6.2.2.2 Archaic Introgression

The genomes of modern humans living in present-day Melanesia have been shown to harbour between 4-6% of Denisovan derived DNA from a potential admixture event 50-60,000 years ago(190). Attempts have been made to identify specific tracts of the genomes of modern day Melanesians that represent DNA inherited from this admixture event(225). It has been suggested that enrichment in particular parts of the genomes of Denisovan derived DNA may reflect processes of adaptive archaic introgression. Negative selection of introgressed Denisovan material will also result in detectable 'archaic deserts'(226). The large dataset of Melanesians used in this research project of 942 Melanesian individuals living in EHPNG would present an opportunity to more fully research these processes. It may be in the future to develop a comprehensive atlas of Denisovan introgression in modern humans. Techniques that accurately map these derived segments require WGS data to differentiate between modern human and Denisovan derived segments.

6.2.2.3 Application of techniques to other selection pressures

The focus of genetic adaptation to selection pressure in this thesis has been restricted to the population response to kuru. It is likely that the ancestors of populations of the highlands of Papua New Guinea including ancestors of EHPNG individuals have experienced numerous selection pressures in the many millennia of habitation and relative isolation in the region. Such adaptations would likely include responses to pathogenic based selection pressure, zoonotic based selection pressures as the result of animal husbandry, and adaptations to the specific climate of EHPNG including the elevation of settlements in the region. Early western observers noted a wide array of causes of mortality in the region including warfare, heart disease, stroke, and infectious diseases(51). When inhabitants of EHPNG were first relocated to lowland coastal areas as part of new labour exchange schemes many suffered and died of malaria as a result of a lack of protection to the disease causing parasite(54). These examples reflect a wide range of possible research enquiries into genetic responses of these isolated populations to specific pressures restricted to the region like kuru and common selection pressures observed elsewhere in other regions of the world. These common selection pressures may have resulted in different mechanisms appearing to

increase fitness and represent possible convergent evolution events that have been observed elsewhere(227). Additionally, given the great change in the way of life of inhabitants of EHPNG since European contact in the 20th century and their greatly differentiated ancestral history compared to commonly studied European populations there will likely be differences in communities genetic risks to diseases and health concerns as a result of changes in lifestyles since European contact. As the nation of PNG further develops, enhanced knowledge of the elevated and reduced risks of communities to health exposures including western diets, a larger elderly population and antibiotic resistance will allow medical authorities to allocate already limited resources more effectively to the areas of highest public health concern. Approaches to understanding the evolutionary and genetic response to kuru used in this thesis have centred on ascertaining the response of populations to a recent and short-lived selection pressure. The research questions relating to other selection pressures faced by these communities in the past will focus on a much wider temporal period and selection periods that may have lasted dozens if not hundreds of generations. This will involve the use of different tools, techniques and approaches to frame these different evolutionary scenarios. Promising tools include those that utilise WGS data and look for genetic signatures of positive selection in the genomes of modern individuals by assaying the presence of 'singletons' present in these genomes at sites in the genome(228). Given the large collection of individuals in this data set, it would be possible to infer these parameters if more data is whole genome sequenced. Such analyses will also be enhanced with merger with public data, with more relevant comparison populations likely to become available in the future.

References

1. Ironside JW, Ritchie DL, Head MW. Prion diseases. *Handbook of clinical neurology*. 2017;145:393-403.
2. Soto C. Prion hypothesis: the end of the controversy? *Trends in biochemical sciences*. 2011;36(3):151-8.
3. Liberski PP, Gajos A, Sikorska B, Lindenbaum S. Kuru, the First Human Prion Disease. *Viruses*. 2019;11(3).
4. Mackenzie G, Will R. Creutzfeldt-Jakob disease: recent developments. *F1000Research*. 2017;6:2053-.
5. Unit TNCRS. CREUTZFELDT-JAKOB DISEASE SURVEILLANCE IN THE UK. 2018.
6. Pocchiari M, Puopolo M, Croes EA, Budka H, Gelpi E, Collins S, et al. Predictors of survival in sporadic Creutzfeldt-Jakob disease and other human transmissible spongiform encephalopathies. *Brain : a journal of neurology*. 2004;127(Pt 10):2348-59.
7. Lo RY, Shyu WC, Li H. Long-duration sCJD with PRNP codon 129 methionine homozygosity and cerebral cortical plaques. *Neurology*. 2006;66(12):1944-5.
8. Palmer MS, Dryden AJ, Hughes JT, Collinge J. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. *Nature*. 1991;352(6333):340-2.
9. Van Everbroeck B, Michotte A, Sciot R, Godfraind C, Deprez M, Quoilin S, et al. Increased incidence of sporadic Creutzfeldt-Jakob disease in the age groups between 70 and 90 years in Belgium. *European journal of epidemiology*. 2006;21(6):443-7.
10. Mok T, Jaunmuktane Z, Joiner S, Campbell T, Morgan C, Wakerley B, et al. Variant Creutzfeldt-Jakob Disease in a Patient with Heterozygosity at PRNP Codon 129. *New England Journal of Medicine*. 2017;376(3):292-4.
11. Collinge J. Prion diseases of humans and animals: their causes and molecular basis. *Annual review of neuroscience*. 2001;24:519-50.
12. Mead S, Poulter M, Beck J, Webb TE, Campbell TA, Linehan JM, et al. Inherited prion disease with six octapeptide repeat insertional mutation--molecular analysis of phenotypic heterogeneity. *Brain : a journal of neurology*. 2006;129(Pt 9):2297-317.
13. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying prion disease penetrance using large population control cohorts. *Science translational medicine*. 2016;8(322):322ra9.
14. Prusiner SB. Inherited prion diseases. *Proceedings of the National Academy of Sciences of the United States of America*. 1994;91(11):4611-4.
15. Soto C, Satani N. The intricate mechanisms of neurodegeneration in prion diseases. *Trends Mol Med*. 2011;17(1):14-24.
16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285.
17. Minikel EV, Vallabh SM, Orseth MC, Brandel J-P, Haik S, Laplanche J-L, et al. Age of onset in genetic prion disease and the design of preventive clinical trials. *bioRxiv*. 2018:401406.
18. Rudge P, Jaunmuktane Z, Adlard P, Bjurstrom N, Caine D, Lowe J, et al. Iatrogenic CJD due to pituitary-derived growth hormone with genetically determined incubation times of up to 40 years. *Brain : a journal of neurology*. 2015;138(Pt 11):3386-99.
19. Jaunmuktane Z, Mead S, Ellis M, Wadsworth JDF, Nicoll AJ, Kenny J, et al. Evidence for human transmission of amyloid- β pathology and cerebral amyloid angiopathy. *Nature*. 2015;525:247.
20. Purro SA, Farrow MA, Linehan J, Nazari T, Thomas DX, Chen Z, et al. Transmission of amyloid- β protein pathology from cadaveric pituitary growth hormone. *Nature*. 2018;564(7736):415-9.

21. Heath CA, Barker RA, Esmonde TFG, Harvey P, Roberts R, Trend P, et al. Dura mater-associated Creutzfeldt-Jakob disease: experience from surveillance in the UK. *Journal of neurology, neurosurgery, and psychiatry*. 2006;77(7):880-2.
22. Kim HL, Do JY, Cho HJ, Jeon Y-C, Park SJ, Ma HI, et al. Dura mater graft-associated Creutzfeldt-Jakob disease: the first case in Korea. *Journal of Korean medical science*. 2011;26(11):1515-7.
23. Bonda DJ, Manjila S, Mehndiratta P, Khan F, Miller BR, Onwuzulike K, et al. Human prion diseases: surgical lessons learned from iatrogenic prion transmission. *Neurosurgical focus*. 2016;41(1):E10.
24. Trevitt CR, Singh PN. Variant Creutzfeldt-Jakob disease: pathology, epidemiology, and public health implications. *The American Journal of Clinical Nutrition*. 2003;78(3):651S-6S.
25. Budka H, Will RG. The end of the BSE saga: do we still need surveillance for human prion diseases? *Swiss medical weekly*. 2015;145:w14212.
26. Bruce M, Chree A, McConnell I, Foster J, Pearson G, Fraser H. Transmission of bovine spongiform encephalopathy and scrapie to mice: strain variation and the species barrier. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 1994;343(1306):405-11.
27. Collinge J. Variant Creutzfeldt-Jakob disease. *Lancet (London, England)*. 1999;354(9175):317-23.
28. Gerodimos R. The UK BSE crisis as a failure of government 2004. 911-29 p.
29. Kaski D, Mead S, Hyare H, Cooper S, Jampana R, Overell J, et al. Variant CJD in an individual heterozygous for PRNP codon 129. *Lancet (London, England)*. 2009;374(9707):2128.
30. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A Novel Protective Prion Protein Variant that Colocalizes with Kuru Exposure. *New England Journal of Medicine*. 2009;361(21):2056-65.
31. Clarke P, Ghani AC. Projections of the future course of the primary vCJD epidemic in the UK: inclusion of subclinical infection and the possibility of wider genetic susceptibility. *J R Soc Interface*. 2005;2(2):19-31.
32. Gill ON, Spencer Y, Richard-Loendt A, Kelly C, Dabaghian R, Boyes L, et al. Prevalent abnormal prion protein in human appendixes after bovine spongiform encephalopathy epizootic: large scale survey. *BMJ : British Medical Journal*. 2013;347:f5675.
33. The L. The Phillips report on BSE and vCJD. *The Lancet*. 2000;356(9241):1535.
34. Aguzzi A, Baumann F, Bremer J. The prion's elusive reason for being. *Annual review of neuroscience*. 2008;31:439-77.
35. Mehrabian M, Hildebrandt H, Schmitt-Ulms G. NCAM1 Polysialylation: The Prion Protein's Elusive Reason for Being? *ASN neuro*. 2016;8(6).
36. Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wuthrich K. NMR structure of the mouse prion protein domain PrP(121-231). *Nature*. 1996;382(6587):180-2.
37. Acevedo-Morantes CY, Wille H. The structure of human prions: from biology to structural models-considerations and pitfalls. *Viruses*. 2014;6(10):3875-92.
38. Vázquez-Fernández E, Vos MR, Afanasyev P, Cebey L, Sevillano AM, Vidal E, et al. The Structural Architecture of an Infectious Mammalian Prion Using Electron Cryomicroscopy. *PLOS Pathogens*. 2016;12(9):e1005835.
39. Lloyd SE, Mead S, Collinge J. Genetics of prion diseases. *Current opinion in genetics & development*. 2013;23(3):345-51.
40. Giles K, Glidden DV, Patel S, Korth C, Groth D, Lemus A, et al. Human prion strain selection in transgenic mice. *Annals of neurology*. 2010;68(2):151-61.
41. MacCannell D. Bacterial strain typing. *Clinics in laboratory medicine*. 2013;33(3):629-50.
42. Wadsworth JDF, Collinge J. Molecular pathology of human prion disease. *Acta Neuropathologica*. 2011;121(1):69-77.

43. Aguzzi A, Heikenwalder M, Polymenidou M. Insights into prion strains and neurotoxicity. *Nature Reviews Molecular Cell Biology*. 2007;8:552.
44. Liberski PP. Electron Microscopy of Prion Diseases. In: Liberski PP, editor. *Prion Diseases*. New York, NY: Springer New York; 2017. p. 123-43.
45. Wadsworth JDF, Joiner S, Linehan JM, Asante EA, Brandner S, Collinge J. Review. The origin of the prion agent of kuru: molecular and biological strain typing. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3747-53.
46. Brandner S, Jaunmuktane Z. Prion disease: experimental models and reality. *Acta neuropathologica*. 2017;133(2):197-222.
47. Collinge J, Alpers MP. Reminiscences and reflections on kuru, personal and scientific. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3613.
48. Whitfield JT, Pako WH, Collinge J, Alpers MP. Cultural factors that affected the spatial and temporal epidemiology of kuru. *Royal Society open science*. 2017;4(1):160789.
49. Alpers MP. Review. The epidemiology of kuru: monitoring the epidemic from its peak to its end. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3707-13.
50. Collinge J, Whitfield J, McKintosh E, Beck J, Mead S, Thomas DJ, et al. Kuru in the 21st century--an acquired human prion disease with very long incubation periods. *Lancet (London, England)*. 2006;367(9528):2068-74.
51. Lindenbaum S. Kuru sorcery : disease and danger in the New Guinea highlands. 2016.
52. Gajdusek DC, Zigas V. Kuru; clinical, pathological and epidemiological study of an acute progressive degenerative disease of the central nervous system among natives of the Eastern Highlands of New Guinea. *The American journal of medicine*. 1959;26(3):442-69.
53. McLean CA. Review. The neuropathology of kuru and variant Creutzfeldt-Jakob disease. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3685-7.
54. Anderson W. *The collectors of lost souls : turning Kuru scientists into whitemen*. Baltimore, Md.: Johns Hopkins University Press; 2008.
55. Anderson WH. Early perceptions of an epidemic. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3675-8.
56. Gray G. 'You are... my anthropological children': AP Elkin, Ronald Berndt and Catherine Berndt, 1940–1956. *Aboriginal History*. 2005;29:77-106.
57. Yam P. The pathological protein mad cow, chronic wasting, and other deadly prion diseases. 2005.
58. Gajdusek DC, Zigas V. Degenerative disease of the central nervous system in New Guinea; the endemic occurrence of kuru in the native population. *The New England journal of medicine*. 1957;257(20):974-8.
59. Lindenbaum S. Cannibalism, kuru and anthropology. *Folia neuropathologica*. 2009;47(2):138-44.
60. Hadlow WJ. Kuru likened to scrapie: the story remembered. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3644-.
61. Adams DB. Evolutionary biology and the risk of scrapie disease in sheep. *Open veterinary journal*. 2018;8(3):282-94.
62. Gajdusek DC, Gibbs CJ, Alpers M. Transmission and Passage of Experimental 'Kuru' to Chimpanzees. *Science (New York, NY)*. 1967;155(3759):212.
63. Gajdusek DC, Gibbs CJ, Alpers M. Experimental Transmission of a Kuru-like Syndrome to Chimpanzees. *Nature*. 1966;209(5025):794-6.
64. Alpers MP. Some tributes to research colleagues and other contributors to our knowledge about kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3614-7.

65. Mathews JD, Glasse R, Lindenbaum S. Kuru and cannibalism. *Lancet* (London, England). 1968;2(7565):449-52.
66. Bolton DC, McKinley MP, Prusiner SB. Identification of a protein that purifies with the scrapie prion. *Science* (New York, NY). 1982;218(4579):1309-11.
67. Lantos PL. From slow virus to prion: a review of transmissible spongiform encephalopathies. *Histopathology*. 1992;20(1):1-11.
68. Collinge J, Whitfield J, McKintosh E, Frosh A, Mead S, Hill AF, et al. A clinical study of kuru patients with long incubation periods at the end of the epidemic in Papua New Guinea. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3725-39.
69. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, et al. A novel protective prion protein variant that colocalizes with kuru exposure. *The New England journal of medicine*. 2009;361(21):2056-65.
70. Lloyd SE, Onwuazor ON, Beck JA, Mallinson G, Farrall M, Targonski P, et al. Identification of multiple quantitative trait loci linked to prion disease incubation period in mice. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(11):6279-83.
71. Asante EA, Smidak M, Grimshaw A, Houghton R, Tomlinson A, Jeelani A, et al. A naturally occurring variant of the human prion protein completely prevents prion disease. *Nature*. 2015;522(7557):478-81.
72. Mead S, Uphill J, Beck J, Poulter M, Campbell T, Lowe J, et al. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Human molecular genetics*. 2012;21(8):1897-906.
73. Ross M. Pronouns as a preliminary diagnostic for grouping Papuan languages 2005. 15-66 p.
74. Wurm SA. *New Guinea area languages and language study*. 1, 1. Canberra: Dep. of Linguistics, Research School of Pacific Studies, Australian National Univ.; 1977.
75. Feil DK. *The evolution of highland Papua New Guinea societies*. Cambridge: Cambridge University Press; 2010.
76. L. Langness L. *Sexual Antagonism in the New Guinea Highlands: A Bena Bena Example* 1967. 161-77 p.
77. Watson JB. OTHER: New Guinea: The Last Unknown. Gavin Souter. *American Anthropologist*. 1967;69(6):789-90.
78. Summerhayes G, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, et al. Human Adaptation and Plant Use in Highland New Guinea 49,000 to 44,000 Years Ago 2010. 78-81 p.
79. Denham T, Fullagar R, Head L. Plant exploitation on Sahul: From colonisation to the emergence of regional specialisation during the Holocene 2009. 29-40 p.
80. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538:207.
81. Greenhill SJ. *TransNewGuinea.org: An Online Database of New Guinea Languages*. *PloS one*. 2015;10(10):e0141563-e.
82. Diamond J, Ordunio D, OverDrive I. *Guns, Germs, and Steel*. 2011.
83. Pawley A. *Papuan pasts cultural, linguistic and biological histories of Papuan-speaking peoples*. Canberra: Pacific Linguistics; 2005.
84. Bergstrom A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, et al. A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science* (New York, NY). 2017;357(6356):1160-3.
85. Greig K, Gosling A, Collins CJ, Boocock J, McDonald K, Addison DJ, et al. Complex history of dog (*Canis familiaris*) origins and translocations in the Pacific revealed by ancient mitogenomes. *Scientific reports*. 2018;8(1):9130.
86. *Travelling companions*. *Nature Plants*. 2016;2:16177.
87. Kirch PV. *Lapita and Its Aftermath: The Austronesian Settlement of Oceania*. *Transactions of the American Philosophical Society*. 1996;86(5):57-70.

88. Kayser M, van Oven M, Choi Y, Brauer S, Mona S, Suarkia D, et al. The Impact of the Austronesian Expansion: Evidence from mtDNA and Y Chromosome Diversity in the Admiralty Islands of Melanesia. *Molecular biology and evolution*. 2008;25(7):1362-74.
89. Brucato N, Fernandes V, Kusuma P, Cerny V, Mulligan CJ, Soares P, et al. Evidence of Austronesian Genetic Lineages in East Africa and South Arabia: Complex Dispersal from Madagascar and Southeast Asia. *Genome biology and evolution*. 2019;11(3):748-58.
90. Posth C, Nägele K, Collieran H, Valentin F, Bedford S, Kami KW, et al. Language continuity despite population replacement in Remote Oceania. *Nature ecology & evolution*. 2018;2(4):731-40.
91. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, et al. Population Turnover in Remote Oceania Shortly after Initial Settlement. *Current Biology*. 2018;28(7):1157-65.e7.
92. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature*. 2016;538(7626):510-3.
93. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010;468:1053.
94. Du A, Alemseged Z. Temporal evidence shows *Australopithecus sediba* is unlikely to be the ancestor of *Homo*. *Science advances*. 2019;5(5):eaav9038.
95. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. *Science (New York, NY)*. 2010;328(5979):710.
96. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS genetics*. 2012;8(10):e1002947.
97. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. 2014;512:194.
98. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*. 2016;26(9):1241-7.
99. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*. 2019.
100. Gajdusek DC, Alpers M. Genetic studies in relation of kuru. I. Cultural, historical, and demographic background. *American journal of human genetics*. 1972;24:Suppl:1-38.
101. Mandeville E. Agnation, Affinity and Migration Among the Kamano of the New Guinea Highlands 1979. 105 p.
102. Lindenbaum S. First impressions of the Fore. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3648-9.
103. Sorenson ER, Claessen HJM, du Torr BM, Griffith J, Hockings P, Jablonko A, et al. Socio-Ecological Change Among the Fore of New Guinea [and Comments and Replies]. *Current Anthropology*. 1972;13(3/4):349-83.
104. B. Watson J. Society as Organized Flow: The Tairora Case 1970. 107-24 p.
105. Connolly B, Anderson R. First contact. New York, N.Y., U.S.A.: Penguin Books; 1988.
106. Mandeville E. Kamano Adoption 1981. 229 p.
107. Littlewood RA. ISOLATE PATTERNS IN THE EASTERN HIGHLANDS OF NEW GUINEA. *The Journal of the Polynesian Society*. 1966;75(1):95-106.
108. Wiesenfeld SL, Gajdusek C. Genetic structure and heterozygosity in the kuru region, eastern highlands of New Guinea. *American journal of physical anthropology*. 1976;45(2):177-90.
109. Bhatia K, Prasad ML, Barnish G, Koki G. Antigen and haplotype frequencies at three human leucocyte antigen loci (HLA-A, -B, -C) in the Pawaia of Papua New Guinea. *American journal of physical anthropology*. 1988;75(3):329-40.
110. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
111. Senn N, Riddell M, Omena M, Siba P, Reeder JC, Clements CJ, et al. Measles in Papua New Guinea: an age-specific serological survey. *Vaccine*. 2010;28(7):1819-23.

112. Watson JB. THE SIGNIFICANCE OF A RECENT ECOLOGICAL CHANGE IN THE CENTRAL HIGHLANDS OF NEW GUINEA. *The Journal of the Polynesian Society*. 1965;74(4):438-50.
113. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*. 2015;526(7571):112-7.
114. Andrews TM, Price RM, Mead LS, McElhinny TL, Thanukos A, Perez KE, et al. Biology undergraduates' misconceptions about genetic drift. *CBE Life Sci Educ*. 2012;11(3):248-59.
115. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. *American journal of human genetics*. 2004;74(6):1111-20.
116. Luzzatto L. Genetic factors in malaria. *Bull World Health Organ*. 1974;50(3-4):195-202.
117. Slade RW, McCallum HI. Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*. 1992;132(3):861-4.
118. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic adaptive introgression in humans. *Nature reviews Genetics*. 2015;16(6):359-71.
119. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*. 2013;193(3):929-41.
120. Hersch EI, Phillips PC. Power and potential bias in field studies of natural selection. *Evolution; international journal of organic evolution*. 2004;58(3):479-85.
121. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays : news and reviews in molecular, cellular and developmental biology*. 2013;35(9):780-6.
122. Pedersen CT, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, et al. The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics*. 2017;205(2):787-801.
123. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research*. 2011;21(6):821-9.
124. Stearns SC, Byars SG, Govindaraju DR, Ewbank D. Measuring selection in contemporary human populations. *Nature Reviews Genetics*. 2010;11:611.
125. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. *Briefings in functional genomics*. 2014;13(5):371-7.
126. Foll M, Gaggiotti Oscar E, Daub Josephine T, Vatsiou A, Excoffier L. Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. *The American Journal of Human Genetics*. 2014;95(4):394-407.
127. R. Vahdati A, Wagner A. Parallel or convergent evolution in human population genomic data revealed by genotype networks. *BMC Evolutionary Biology*. 2016;16(1):154.
128. Casanova J-L. Adaptive immunity by convergent evolution. *Nature Reviews Immunology*. 2018;18:294.
129. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*. 2017;19:110.
130. Bearn AG, Miller ED. Archibald Garrod and the development of the concept of inborn errors of metabolism. *Bulletin of the history of medicine*. 1979;53(3):315-28.
131. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-86.
132. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*. 2018;19(9):581-90.

133. Illingworth DR, Stein EA, Mitchel YB, Dujovne CA, Frost PH, Knopp RH, et al. Comparative effects of lovastatin and niacin in primary hypercholesterolemia. A prospective trial. *Archives of internal medicine*. 1994;154(14):1586-95.
134. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013;45(11):1274-83.
135. Plotnikov D, Guggenheim JA. Mendelian randomisation and the goal of inferring causation from observational studies in the vision sciences. *Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*. 2019;39(1):11-25.
136. Biedrzycki RJ, Sier AE, Liu D, Dreikorn EN, Weeks DE. Spinning convincing stories for both true and false association signals. *Genet Epidemiol*. 2019.
137. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews Genetics*. 2011;13(1):36-46.
138. Yan G, Guo T, Xiao S, Zhang F, Xin W, Huang T, et al. Imputation-Based Whole-Genome Sequence Association Study Reveals Constant and Novel Loci for Hematological Traits in a Large-Scale Swine F2 Resource Population. *Frontiers in genetics*. 2018;9:401.
139. Kuchenbaecker K, Telkar N, Reiker T, Walters RG, Lin K, Eriksson A, et al. The transferability of lipid loci across African, Asian and European cohorts. *Nature communications*. 2019;10(1):4330.
140. Hizume M, Kobayashi A, Teruya K, Ohashi H, Ironside JW, Mohri S, et al. Human prion protein (PrP) 219K is converted to PrPSc but shows heterozygous inhibition in variant Creutzfeldt-Jakob disease infection. *The Journal of biological chemistry*. 2009;284(6):3603-9.
141. Misra A, Chakrabarti SS, Gambhir IS. New genetic players in late-onset Alzheimer's disease: Findings of genome-wide association studies. *The Indian journal of medical research*. 2018;148(2):135-44.
142. Karimi-Moghadam A, Charsouei S, Bell B, Jabalameli MR. Parkinson Disease from Mendelian Forms to Genetic Susceptibility: New Molecular Insights into the Neurodegeneration Process. *Cellular and molecular neurobiology*. 2018;38(6):1153-78.
143. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*. 2018;50(9):1219-24.
144. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*. 2015;31(9):1466-8.
145. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? *Genome medicine*. 2018;10(1):100.
146. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatric genetics*. 2018;28(5):85-9.
147. De La Vega FM, Bustamante CD. Polygenic risk scores: a biased prediction? *Genome medicine*. 2018;10(1):100-.
148. Vilhjalmsjon BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American journal of human genetics*. 2015;97(4):576-92.
149. Mackay TFC. The Genetic Architecture of Quantitative Traits. *Annual review of genetics*. 2001;35(1):303-39.
150. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics*. 2017;13(6):e1006836.
151. Whitfield JT, Pako WH, Collinge J, Alpers MP. Mortuary rites of the South Fore and kuru. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3721-4.
152. Thornton IWB, Cook S, Edwards JS, Harrison RD, Schipper C, Shanahan M, et al. Colonization of an island volcano, Long Island, Papua New Guinea, and an emergent island, Motmot, in its caldera lake. VII. Overview and discussion. *Journal of Biogeography*. 2001;28(11-12):1389-408.

153. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas of Human Admixture History. *Science (New York, NY)*. 2014;343(6172):747.
154. Halder I, Yang B-Z, Kranzler HR, Stein MB, Shriver MD, Gelernter J. Measurement of admixture proportions and description of admixture structure in different U.S. populations. *Hum Mutat*. 2009;30(9):1299-309.
155. Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. *Elife*. 2016;5:e15266.
156. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. *Science (New York, NY)*. 2014;343(6172):747-51.
157. Haddon AC. Smoking and Tobacco Pipes in New Guinea. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 1946;232(586):1-278.
158. Bourke R. Sweet potato in Papua New Guinea 2002.
159. The Genomes Project C, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. *Nature*. 2015;526:68.
160. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007;81(3):559-75.
161. Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, et al. Complex Patterns of Admixture across the Indonesian Archipelago. *Molecular biology and evolution*. 2017;34(10):2439-52.
162. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nature Methods*. 2011;9:179.
163. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS genetics*. 2012;8(1):e1002453.
164. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*. 2009;19(9):1655-64.
165. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. 2015;519(7543):309-14.
166. Perry GH, Foll M, Grenier J-C, Patin E, Nédélec Y, Pacis A, et al. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(35):E3596-E603.
167. Jewett EM, Steinrücken M, Song YS. The Effects of Population Size Histories on Estimates of Selection Coefficients from Time-Series Genetic Data. *Molecular biology and evolution*. 2016;33(11):3002-27.
168. Hunley K. Reassessment of global gene-language coevolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(7):1919-20.
169. Need AC, Kasperavičiūtė D, Cirulli ET, Goldstein DB. A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. *Genome Biology*. 2009;10(1):R7.
170. Lopez S, Thomas MG, van Dorp L, Ansari-Pour N, Stewart S, Jones AL, et al. The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection. *American journal of human genetics*. 2017;101(3):353-68.
171. Bradburd GS, Ralph PL, Coop GM. A Spatial Framework for Understanding Population Structure and Admixture. *PLoS genetics*. 2016;12(1):e1005703.
172. López S, van Dorp L, Hellenthal G. Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary bioinformatics online*. 2016;11(Suppl 2):57-68.
173. Musunuru K, Kathiresan S. HapMap and mapping genes for cardiovascular disease. *Circ Cardiovasc Genet*. 2008;1(1):66-71.
174. Ranciaro A, Campbell Michael C, Hirbo Jibril B, Ko W-Y, Froment A, Anagnostou P, et al. Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *The American Journal of Human Genetics*. 2014;94(4):496-510.

175. Karupiah G, Panchanathan V, Sakala IG, Chaudhri G. Genetic resistance to smallpox: lessons from mousepox. *Novartis Foundation symposium*. 2007;281:129-36; discussion 36-40, 208-9.
176. Tabajara de Oliveira Martins D, Rodrigues E, Casu L, Benitez G, Leonti M. The historical development of pharmacopoeias and the inclusion of exotic herbal drugs with a focus on Europe and Brazil. *Journal of ethnopharmacology*. 2019:111891.
177. Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Beck J, et al. Genetic susceptibility, evolution and the kuru epidemic. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3741-6.
178. Mead S, Stumpf MP, Whitfield J, Beck JA, Poulter M, Campbell T, et al. Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science (New York, NY)*. 2003;300(5619):640-3.
179. Hardy J, Scholz S, Evans W, Goldfarb L, Singleton A. Prion genotypes in Central America suggest selection for the V129 allele. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2006;141b(1):33-5.
180. Kreitman M, Di Rienzo A. Balancing claims for balancing selection. *Trends in genetics : TIG*. 2004;20(7):300-4.
181. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*. 2014;15:379.
182. King KC, Lively CM. Does genetic diversity limit disease spread in natural host populations? *Heredity (Edinb)*. 2012;109(4):199-203.
183. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature genetics*. 2019;51(2):277-84.
184. Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, et al. Early Neolithic genomes from the eastern Fertile Crescent. *Science (New York, NY)*. 2016;353(6298):499-503.
185. Albrechtsen A, Sand Korneliusen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol*. 2009;33(3):266-74.
186. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome research*. 2002;12(6):996-1006.
187. Soldevila M, Andres AM, Ramirez-Soriano A, Marques-Bonet T, Calafell F, Navarro A, et al. The prion protein gene in humans revisited: lessons from a worldwide resequencing study. *Genome research*. 2006;16(2):231-9.
188. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. *Nature*. 2016;538(7624):207-14.
189. Rougier H, Crevecoeur I, Beauval C, Posth C, Flas D, Wißing C, et al. Neandertal cannibalism and Neandertal bones used as tools in Northern Europe. *Scientific reports*. 2016;6:29005-.
190. Reich D. *Who we are and how we got here : ancient DNA and the new science of the human past* 2019.
191. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szecsenyi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. *Nature*. 2018;555(7695):197-203.
192. Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, et al. Lactase haplotype diversity in the Old World. *American journal of human genetics*. 2001;68(1):160-72.
193. Zhong M, Zhang Y, Lange K, Fan R. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. *Statistics and its interface*. 2011;4(1):51-63.
194. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nature reviews Genetics*. 2014;15(6):379-93.
195. Lindenbaum S. Review. Understanding kuru: the contribution of anthropology and medicine. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2008;363(1510):3715-20.

196. Liu X, Ong RT, Pillai EN, Elzein AM, Small KS, Clark TG, et al. Detecting and characterizing genomic signatures of positive selection in global populations. *American journal of human genetics*. 2013;92(6):866-81.
197. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution*. 2014;31(10):2824-7.
198. Synofzik M, Smets K, Mallaret M, Di Bella D, Gallenmüller C, Baets J, et al. SYNE1 ataxia is a common recessive ataxia with major non-cerebellar features: a large multi-centre study. *Brain : a journal of neurology*. 2016;139(5):1378-93.
199. Edvardson S, Cinnamon Y, Jalas C, Shaag A, Maayan C, Axelrod FB, et al. Hereditary sensory autonomic neuropathy caused by a mutation in dystonin. 2012;71(4):569-72.
200. Mather CA, Mooney SD, Salipante SJ, Scroggins S, Wu D, Pritchard CC, et al. CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genet Med*. 2016;18(12):1269-75.
201. Sakharkar MK, Chow VT, Kanguane P. Distributions of exons and introns in the human genome. *In silico biology*. 2004;4(4):387-93.
202. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*. 2015;528(7583):499-503.
203. Hwang JY, Aromolaran KA, Zukin RS. The emerging field of epigenetics in neurodegeneration and neuroprotection. *Nature reviews Neuroscience*. 2017;18(6):347-61.
204. Berson A, Nativio R, Berger SL, Bonini NM. Epigenetic Regulation in Neurodegenerative Diseases. *Trends in neurosciences*. 2018;41(9):587-98.
205. Balding DJ. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*. 2006;7(10):781-91.
206. Baranzini SE. The era of GWAS is over - Commentary. *Mult Scler*. 2018;24(3):260-1.
207. Verweij KJ, Mosing MA, Zietsch BP, Medland SE. Estimating heritability from twin studies. *Methods in molecular biology (Clifton, NJ)*. 2012;850:151-70.
208. Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. *Nature genetics*. 2017;49(7):986-92.
209. Schelzke G, Kretzschmar HA, Zerr I. Clinical aspects of common genetic Creutzfeldt-Jakob disease. *European journal of epidemiology*. 2012;27(2):147-9.
210. Manjrekar J. Epigenetic inheritance, prions and evolution. *Journal of genetics*. 2017;96(3):445-56.
211. Alzualde A, Moreno F, Martinez-Lage P, Ferrer I, Gorostidi A, Otaegui D, et al. Somatic mosaicism in a case of apparently sporadic Creutzfeldt-Jakob disease carrying a de novo D178N mutation in the PRNP gene. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*. 2010;153b(7):1283-91.
212. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nature genetics*. 2016;48(10):1284-7.
213. Hu B, Shen N, Li JJ, Kang H, Hong J, Fletcher J, et al. Genome-wide association study reveals sex-specific genetic architecture of facial attractiveness. *PLoS genetics*. 2019;15(4):e1007973.
214. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*. 2018:355057.
215. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017;95:1.22.1-1..3.
216. Wulf M-A, Senatore A, Aguzzi A. The biological function of the cellular prion protein: an update. *BMC Biology*. 2017;15(1):34.
217. Castellani RJ, Perry G, Smith MA. Prion disease and Alzheimer's disease: pathogenic overlap. *Acta neurobiologiae experimentalis*. 2004;64(1):11-7.
218. Zhou J, Liu B. Alzheimer's disease and prion protein. *Intractable Rare Dis Res*. 2013;2(2):35-44.

219. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American journal of human genetics*. 2015;97(4):576-92.
220. Gaffney D, Summerhayes GR, Ford A, Scott JM, Denham T, Field J, et al. Earliest Pottery on New Guinea Mainland Reveals Austronesian Influences in Highland Environments 3000 Years Ago. *PLoS one*. 2015;10(9):e0134497.
221. Hellenthal G, Falush D, Myers S, Reich D, Busby GB, Lipson M, et al. The Kalash Genetic Isolate? The Evidence for Recent Admixture. *American journal of human genetics*. 2016;98(2):396-7.
222. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jorgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science (New York, NY)*. 2015;349(6254):1343-7.
223. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. *Briefings in functional genomics*. 2014;13(5):371-7.
224. Klitzman R. *The trembling mountain : a personal account of Kuru, Cannibals, and Mad Cow Disease*. 2007.
225. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science (New York, NY)*. 2016;352(6282):235-9.
226. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*. 2015;16:359.
227. Stern DL. The genetic causes of convergent evolution. *Nature Reviews Genetics*. 2013;14:751.
228. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. *Science (New York, NY)*. 2016;354(6313):760-4.