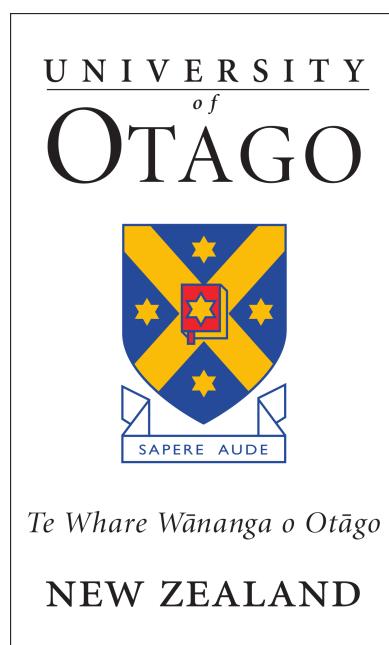


Selection and Metabolic Disease in the
Pacific

Murray Cadzow



a thesis submitted for the degree of
Doctor of Philosophy
at the University of Otago, Dunedin,
New Zealand.

Abstract

An example of a "signature of selection" in a population is a region of the genome that exhibits a reduction in genetic variability with a particular linkage disequilibrium pattern. This reduction in variation can arise when the phenotype of a neutral beneficial allele experiences a favourable change in environmental conditions. This results in an increased frequency of both the allele, and linked sites, within a population. Polynesian populations share a common genetic ancestry with East Asia, but little characterisation of genetic selection has been undertaken in Polynesian populations.

Serum urate has been associated with metabolic disorders such as obesity, type 2 diabetes, renal disease and metabolic syndrome. It is hypothesised that serum urate may have undergone positive selection in Polynesians due to some of the beneficial properties, such as its role as an anti-oxidant, or as an adjuvant for the innate immune system. New Zealand Polynesians have inherently elevated serum urate levels and increased rates of gout. This thesis presents the results of a genome-wide study of selection in Polynesian (and other) populations, focusing on testing the hypothesis that genomic loci containing genes involved in urate processing have undergone selection.

There was no evidence of wide-spread selection at genes associated with urate and gout, or related metabolic disorders, but there was evidence at some individual loci. Pathway analysis showed that of the significant pathways, there was a dominance of metabolic pathways that were enriched for genes with signatures of selection. Calcium related transport and signalling was a theme amongst loci that displayed signs of possible selection. Regions of the genome that were possibly selected in modern-day Polynesian populations also had similarities to those of modern-day East Asian populations.

This thesis has provided identification and characterisation of regions in the genome with possible evidence of genetic selection in Polynesian populations, that previously was not available. It has also provided insight into the role of genetic selection with respect to urate and metabolic disease.

Acknowledgements

It takes a village to undertake research is certainly an appropriate paraphrase to encapsulate the numerous people who have contributed to the research presented in this thesis who deserve to be acknowledged. Firstly, a large thank-you goes to my supervisors, Professor Tony Merriman, and Associate Professor Mik Black, you have had the most direct impact on this project in helping form and shape the direction and offering excellent advice. I am highly appreciative of the time you have spent mentoring, and the opportunities offered.

To my committee members, Professor Lisa Matisoo-Smith and Dr Phillip Wilcox, you were crucial in keeping me on track and providing additional guidance.

I would also like to thank the members of the Merriman Lab, you provided a wonderful friendly, family-like environment. Special mention needs to go to Marilyn, Ruth, and Mandy, the back-bone of the lab and good friends, without your knowledge, hard work and skill in the lab, this project would not have been possible. To the occupants of 'Red Room' - Tanya, Matt, Anna, and Riku, we have shared much banter, laughter, and joy, as well as the many 'I wonder...' curiosity moments. Thanks.

Acknowledgement and thanks go to the participants of the studies that were used in this thesis. Thanks also needs to go to NeSI, who provided the computational resources, and to the Virtual Institute of Statistical Genomics, who provided my scholarship stipend.

Thanks to my friends and family, your support has been highly appreciated. Ben, thank you for being there as a close friend and brother-in-law, for the chats and beers, and in which we could uniquely have a shared thesis writing experience together. Mum and dad, and Richard and Dianne, thank you for providing loving supportive spaces to come, crash and decompress.

To Isobel and Esther, my amazing daughters, you have brought so much joy and inspired me during the tough times. I hope I can inspire you, as you do for me.

And lastly, to my wife, Hana, it is finally done! You were critical in enabling me to embark on what seemed to be a far-fetched journey. A massive amount of gratitude is owed for your support. Your encouragement and love got me through. Without you, I could not have done this.

Contents

List of Tables	vii
List of Figures	x
Acronyms	xiii
1 Introduction	1
1.1 Settlement of Polynesia	1
1.2 Metabolic syndrome	2
1.2.1 Hyperuricaemia and gout	2
1.3 Genetic variation	3
1.4 Natural selection	4
1.4.1 Neutrality and genetic drift	4
1.4.2 Selection	5
1.5 Methods for detecting selection	7
1.5.1 Macroevolutionary methods	8
1.5.2 Haplotypic methods	9
1.5.3 Frequency spectrum methods	11
1.5.4 Composite methods	14
1.5.5 Power	15
1.5.6 Challenges	16
1.6 Improving GWAS	17
1.6.1 Selection and GWAS	18
1.6.2 Pathway analysis	19
1.7 Applications	19
1.7.1 Genome-wide selection	19
1.7.2 Selection of metabolic syndrome and urate	21
1.8 Research purpose and aims	22
2 Methods and data	23
2.1 Methods	23
2.1.1 Phasing	23
2.1.2 Selection Statistics	25
2.1.3 Disease associated gene lists	26

2.1.4	Principal component analysis	27
2.1.5	Admixture analysis	28
2.1.6	Heritability analysis	29
2.1.7	Pathway Analysis	31
2.2	Datasets	31
2.2.1	1000 Genomes Project Phase 3	31
2.2.2	Genetics of Gout in Aotearoa study	31
2.2.3	Selection dataset	34
2.2.4	UK Biobank	39
2.2.5	Reference datasets	40
3	Positive Selection in Polynesian Populations	41
3.1	Introduction	41
3.1.1	Positive selection	42
3.1.2	Selection in Polynesian populations	42
3.1.3	Selection in urate and gout	42
3.1.4	Objectives	44
3.2	Methods	44
3.2.1	Calculation of selection and neutrality statistics	44
3.2.2	Pathway enrichment analysis	45
3.3	Results	45
3.3.1	False discovery rate of windowed statistics	45
3.3.2	Comparison with prior publications	47
3.3.3	Selection in Polynesian populations - genome-wide analysis	48
3.3.4	Selection in disease-associated genes	64
3.4	Chapter Discussion	83
3.4.1	Identifying regions under selection in Polynesian populations	83
3.4.2	Selection of urate associated genes	84
3.4.3	Replication of candidate thrifty-genes	85
3.4.4	Selection in malaria associated loci in Polynesian populations	86
3.4.5	Study limitations	87
3.4.6	Conclusion	89
4	Clustering of Selection Statistics	91
4.1	Introduction	91
4.1.1	Signatures of selection	91
4.1.2	Settlement of Polynesia	92
4.1.3	Health disparities in Polynesian populations	92
4.1.4	Objectives	93
4.2	Methods	93
4.2.1	Data	93
4.2.2	Principal component analysis	93

4.2.3	Admixture analysis	93
4.2.4	Frequency spectrum	94
4.2.5	Extended haplotype homozygosity	94
4.2.6	Pathway enrichment analysis	95
4.2.7	Disease-associated genes	95
4.3	Results	95
4.3.1	Frequency spectrum	95
4.3.2	Hierarchical clustering of distribution extremes	103
4.3.3	Hierarchical clustering of selected haplotypic regions	122
4.3.4	Disease-associated gene clustering	127
4.3.5	Random draws	133
4.4	Chapter discussion	137
4.4.1	Use of selection and neutrality statistics to group populations	137
4.4.2	Potential shared selective histories for loci associated with urate and metabolic disease	139
4.4.3	Limitations	141
4.4.4	Conclusions	143
5	Selection and Association Studies	145
5.1	Genetic associations with Gout	145
5.1.1	Performance of gout definitions	145
5.1.2	Genetic associations for gout in Polynesian populations	146
5.1.3	Use of selection statistics to inform GWAS	146
5.1.4	Objectives	147
5.2	Methods	147
5.2.1	European Gout GWAS	147
5.2.2	Polynesian Gout GWAS	149
5.3	Results	149
5.3.1	UK Biobank	149
5.3.2	GWAS results	157
5.3.3	Comparison of haplotypic selection with gout GWAS	160
5.4	Chapter Discussion	166
5.4.1	Performance of gout definition	166
5.4.2	GWAS and selection	167
5.4.3	Limitations	168
5.4.4	Conclusions	169
6	Summary and Conclusions	171
6.1	Summary	171
6.1.1	Evidence of selection in Polynesian populations	171
6.1.2	Shared ancestry of selected loci	172
6.1.3	Incorporation of selection analyses into GWAS	173

6.2	Significance	173
6.3	Study limitations	174
6.4	Future directions	176
6.5	Conclusion	177
	References	177
	Appendices	223
A	Supplemental tables	223
A1	Chapter 3 tables	223
A1.1	Hider Regions	223
A1.2	Genes in 1st percentile for Polynesian populations for frequency based statistics	243
A1.3	Significant iHS and nSL markers in Polynesian populations	243
A1.4	Significant XP-EHH markers in Polynesian populations	243
A1.5	Inflammatory and auto-immune genes with significant results in selection statistics for Polynesian populations	243
A2	Chapter 4 Tables	252
A2.1	Admixture cross-validation error	252
A2.2	GWAS catalog studies and references table	253
A2.3	GWAS catalog disease gene lists	258
A2.4	Site-frequency spectrum based selection and neutrality statistics summary tables	267
A2.5	Polynesian windows for clustering of the extremes	272
A2.6	Clustered regions from significant markers for iHS and nSL	272
A2.7	Clustered median centered metabolic disease genes	273
B	Additional scripts	277
B1	GWAS catalog gene list creation	277
B2	SelectionTools Pipeline NeSI Scripts	281
B2.1	unimputed_selection_pipeline.sl	281
B2.2	unimputed_defaults_nesi-18-3-16.cfg	282
B2.3	run_selscan.sl	284
B2.4	run_nsl_selscan.sl	286
B2.5	run_xpehh.sl	287
B2.6	Extract results	288
C	Papers published during the course of this thesis	295
C1	Papers relating to this thesis	295
C2	Other papers	295

List of Tables

1.1	Summary of Selection Methods	8
2.1	Description of Populations used in the selection analysis	32
2.2	Clinical information for samples that were genotyped on the CoreExome platform by genetic ancestry.	34
2.3	Clinical information for New Zealand Populations used in the selection analysis	35
2.4	Reference Datasets	40
3.1	False discovery rate for windowed frequency spectrum based intra-population statistics by population.	46
3.2	Number of significant regions and genes in Polynesian populations for the intra population SFS statistics.	49
3.3	Significant pathway terms from Enrichr KEGG 2016 pathway enrichment analysis in Polynesian populations.	54
3.4	Number of significant windows/markers by selection or neutrality statistic by population, grouped by super population.	55
3.5	Number of significant SNPs that were in common between populations.	56
3.6	Number of genes with evidence of possible positive selection in super populations from intra-population tests from urate and co-morbidities GWAS associated loci.	67
3.7	Urate and gout associated loci that showed signs of possible selection in Polynesian populations.	68
3.8	Obesity-associated loci that showed signs of possible selection in Polynesian populations.	70
3.9	Type 2 diabetes-associated loci that showed signs of possible selection in Polynesian populations.	74
3.10	Kidney disease-associated loci that showed signs of possible selection in Polynesian populations.	76
3.11	Metabolic syndrome-associated genes that showed signs of possible selection in Polynesian populations	77
3.12	Neurological disease associated loci that showed signs of possible selection in Polynesian populations.	78
3.13	Linkage disequilibrium in EAS populations between markers in <i>DDC</i>	81
3.14	Markers with significant XP-EHH in Polynesian populations at <i>DDC</i> on chromosome 7. .	82

4.1	Proportion of populations clustered into their corresponding super population by selection and neutrality statistic.	105
4.2	Mean exclusivity of clusters for a given super population by selection and neutrality statistic.	106
4.3	Summary statistics for Tajima's D by super population.	107
4.4	Summary statistics for Fay and Wu's H by super population.	111
4.5	Summary statistics for Fu and Li's F by super population.	115
4.6	Summary statistics for Zeng's E by super population.	121
4.7	Proportion of individual populations assigned to their super population by hierarchical clustering across 10000 iterations of 2500 randomly drawn windows.	135
4.8	Percentage of draws from randomly selected genes for completely clustered super populations.	136
5.1	Clinical details of participants in the UK Biobank.	150
5.2	Number and prevalence of gout by gout definition.	152
5.3	The odds ratios for the 30 variants reported in Köttgen <i>et al.</i> (2013) for the different gout classification methods.	156
5.4	Top genome-wide significant SNPs by gout definition in the UK Biobank cohort. . . .	159
5.5	Clinical details for participants of the Polynesian gout GWAS.	160
5.6	SNPs that had $ iHS $ or $ nSL > 2$ from the GBR, 1000 Genomes Project population and a gout association $P > 2 \times 10^{-4}$ in the UK Biobank self-reported gout or ULT definition.	162
5.7	SNPs that had $ iHS $ or $ nSL > 2$ in the New Zealand Polynesian populations and a gout association $P > 2 \times 10^{-4}$ in the Polynesian GWAS.	165
S1	Regions from Hider et al 2013 that were replicated for selection.	224
S2	All genes that had a window in the 1st percentile and value < 0 from a Polynesian population	243
S3	Significant markers for iHS and nSL in Polynesian populations	243
S4	Genes that had a significant XP-EHH value in Polynesian populations.	243
S5	Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations.	243
S6	Admixture cross-validation error for different values of K.	253
S7	GWAS catalog studies used.	253
S8	Disease associated genes by category from the GWAS catalog.	258
S9	Tajima's D summary statistics by population	268
S10	Fay and Wu's H summary statistics by population	269
S11	Fu and Li's F summary statistics by population	270
S12	Zeng's E summary statistics by population	271
S13	Genes intersecting the windows of the 1st percentile Tajima's D in Polynesian populations.	272
S14	Genes intersecting the windows of the 99th percentile Tajima's D in Polynesian populations.	272
S15	Genes intersecting the windows of the 1st percentile Fay and Wu's H in Polynesian populations.	272

S16 Genes intersecting the windows of the 99th percentile Fay and Wu's <i>H</i> in Polynesian populations.	272
S17 Genes intersecting the windows of the 1st percentile Fu and Li's <i>F</i> in Polynesian populations.	272
S18 Genes intersecting the windows of the 99th percentile Fu and Li's <i>F</i> in Polynesian populations.	272
S19 Genes intersecting the windows of the 1st percentile Zeng's <i>E</i> in Polynesian populations.	272
S20 Genes intersecting the windows of the 99th percentile Zeng's <i>E</i> in Polynesian populations.	272
S21 Positions of regions created by clustering significant markers for iHS or nSL by population.	273
S22 Tajima's <i>D</i> for windows at gout-associated loci.	273
S23 Fay and Wu's <i>H</i> for windows at urate and gout-associated loci.	273
S24 Fu and Li's <i>F</i> for windows at urate and gout-associated loci.	273
S25 Zeng's <i>E</i> for windows at urate and gout-associated loci.	273
S26 Tajima's <i>D</i> for windows at obesity-associated loci.	273
S27 Fay and Wu's <i>H</i> for windows at obesity-associated loci.	273
S28 Fu and Li's <i>F</i> for windows at obesity-associated loci.	273
S29 Zeng's <i>E</i> for windows at obesity-associated loci.	273
S30 Tajima's <i>D</i> for windows at type 2 diabetes-associated loci.	274
S31 Fay and Wu's <i>H</i> for windows at type 2 diabetes-associated loci.	274
S32 Fu and Li's <i>F</i> for windows at type 2 diabetes-associated loci.	274
S33 Zeng's <i>E</i> for windows at type 2 diabetes-associated loci.	274
S34 Tajima's <i>D</i> for windows at kidney disease-associated loci.	274
S35 Fay and Wu's <i>H</i> for windows at kidney disease-associated loci.	274
S36 Fu and Li's <i>F</i> for windows at kidney disease-associated loci.	274
S37 Zeng's <i>E</i> for windows at kidney disease-associated loci.	274
S38 Tajima's <i>D</i> for windows at metabolic syndrome-associated loci.	275
S39 Fay and Wu's <i>H</i> for windows at metabolic syndrome-associated loci.	275
S40 Fu and Li's <i>F</i> for windows at metabolic syndrome-associated loci.	275
S41 Zeng's <i>E</i> for windows at metabolic syndrome-associated loci.	275

List of Figures

1.1	Diagram of hard selection sweep.	6
2.1	Principal components 2 and 4 for individuals used in the selection analysis.	35
3.1	Location of urate transporters of gut and kidney.	43
3.2	Schematic of <i>CREBRF</i> locus with flanking regions.	51
3.3	Upset plot of intra-population haplotypic and frequency spectrum-based evidence pooled by super population.	52
3.4	Upset plot of number of genes with evidence of possible positive selection in Polynesian populations.	53
3.5	Manhattan plot for significant markers in Cook Island Māori in New Zealand for both iHS and nSL.	58
3.6	Manhattan plot for significant markers in Māori in New Zealand for both iHS and nSL.	60
3.7	Manhattan plot for significant markers in Samoans in New Zealand for both iHS and nSL.	62
3.8	Manhattan plot for significant markers in Tongans in New Zealand for both iHS and nSL.	65
3.9	Genes associated with urate, gout, obesity, type 2 diabetes, kidney disease, and metabolic syndrome with evidence from site-frequency spectrum and haplotypic statistics in Polynesian populations.	66
4.1	Principal components 1 to 4 for all populations.	97
4.2	Principal components 1 and 2 for populations of the European super population.	98
4.3	Principal components 2 and 6 for the Polynesian populations.	98
4.4	Proportions of ancestral populations as inferred from ADMIXTURE	100
4.5	Hierarchical clustering of the populations using selection test statistics for each entire chromosome.	102
4.6	Hierarchical clustering on chromosomal FST	103
4.7	Plot of the distribution of Tajima's <i>D</i> by population	108
4.8	Hierarchical clustering of Tajima's <i>D</i> using the upper and lower 1% of the distribution.	109
4.9	Plot of the distribution of Fay and Wu's <i>H</i> by population.	112
4.10	Hierarchical clustering of Fay and Wu's <i>H</i> using the upper and lower 1% of the distribution.	113
4.11	Plot of the distribution of Fu and Li's <i>F</i> by population.	116
4.12	Hierarchical clustering of Fu and Li's <i>F</i> using the upper and lower 1% of the distribution.	117
4.13	Plot of the distribution of Zeng's <i>E</i> by population.	119

4.14	Hierarchical clustering of Zeng's <i>E</i> using the upper and lower 1% of the distribution.	120
4.15	Hierarchical clustering of iHS.	124
4.16	Hierarchical clustering of nSL.	126
4.17	Dendograms created from hierarchical clustering applied to windows from loci associated with urate and gout	128
4.18	Dendograms created from hierarchical clustering applied to windows from obesity-associated loci.	130
4.19	Dendograms created from hierarchical clustering applied to windows from type 2 diabetes-associated loci.	131
4.20	Dendograms created from hierarchical clustering applied to windows from kidney disease-associated loci.	132
4.21	Dendograms created from hierarchical clustering applied to windows from metabolic syndrome-associated loci.	134
5.1	Upset plot of number of samples and intersections for different gout classification criteria.	151
5.2	Odds ratios for rs2231142 and rs12498742 based on gout definitions.	153
5.3	Odds ratios for the 28 SNPs reported in Köttgen <i>et al.</i> (2013) based on gout definitions.	154
5.4	Overlap of markers that reached nominal genome-wide significance between the different gout classifications.	158
5.5	Manhattan plot for association with gout, adjusted for age, sex and BMI using the self-reported gout or ULT usage gout definition.	158
5.6	Manhattan plot for gout GWAS, adjusted for age and sex, in Polynesian ancestry individuals.	161
5.7	Locus zoom plot of Polynesian gout GWAS results for the region covering <i>IBSP</i> , <i>PKD2</i> , and <i>ABCG2</i>	164

Acronyms

Δ DAF	change in derived allele frequency
1KGP	1000 Genomes Project
ACB	African Caribbean in Barbados
ACR	American College of Rheumatology
AFR	African Super Population
AIC	Akaike information criterion
AMR	American Super Population
ARIC	Atherosclerosis Risk in the Community
ASW	Americans of African Ancestry in SW USA
BEB	Bengali from Bangladesh
BMI	body mass index
CDX	Chinese Dai in Xishuangbanna China
CEU	Utah Residents (CEPH) with Northern and Western Ancestry
CHB	Han Chinese in Beijing China
CHS	Southern Han Chinese
CI	confidence interval
CIM	Cook Island Māori in New Zealand
CLM	Colombians from Medellin Colombia
CVD	cardio vascular disease
DAF	derived allele frequency
EAS	East Asian Super Population
EHH	extended haplotype homozygosity
ESN	Esan in Nigeria
EUR	European Super Population
FDR	false discovery rate
FHS	Framingham Heart Study
FIN	Finnish in Finland
GBR	British in England and Scotland
GCTA	Genome-wide Complex Trait Analysis
GIH	Gujarati Indian from Houston Texas
GRM	genetic relationship matrix
GWAS	genome-wide association study
GWD	Gambian in Western Divisions in the Gambia
HWE	Hardy-Weinberg equilibrium
IBS	Iberian Population in Spain
ICD-10	International Classification of Diseases, Tenth Revision

iHH	integrated extended haplotype homozygosity
iHS	integrated haplotype homozygosity score
INDEL	insertion or deletion
ITU	Indian Telugu from the UK
JPT	Japanese in Tokyo Japan
KEGG	Kyoto Encyclopedia of Genes and Genomes
KHV	Kinh in Ho Chi Minh City Vietnam
LD	linkage disequilibrium
LWK	Luhya in Webuye Kenya
MAF	minor allele frequency
MSL	Mende in Sierra Leone
MXL	Mexican Ancestry from Los Angeles USA
nSL	number of segregating sites by length
NZ	New Zealand
NZC	Europeans in New Zealand
NZM	Māori in New Zealand
OR	odds ratio
PC	principal component
PCA	principal component analysis
PEL	Peruvians from Lima Peru
POL	Polynesian Super Population
PUR	Puerto Ricans from Puerto Rico
SAM	Samoans in New Zealand
SAS	South Asian Super Population
SD	standard deviation
SE	standard error
SFS	site frequency spectrum
SNP	single nucleotide polymorphism
TON	Tongans in New Zealand
TSI	Toscani in Italia
ULT	urate lowering therapy
VCF	variant call format
WTCCC	Wellcome Trust Case-Control Consortium
XP-EHH	cross-population extended haplotype homozygosity
ya	years ago
YRI	Yoruba in Ibadan Nigeria

Chapter 1

Introduction

This thesis investigates the role that positive genetic selection has played in causing the health disparities of metabolic disease in Polynesian populations.

This chapter provides the background and rationale for the research presented in this thesis. It covers the settlement of the Pacific, and highlights the health disparities facing Polynesian populations and the relevance of genetics. It then introduces natural selection and the methodologies for detecting ‘signals of selection’, followed by how genetic selection may have played a role in creating the health disparities caused by metabolic disease in Polynesian populations.

1.1 Settlement of Polynesia

The Polynesian triangle is defined as the geographic region in the Pacific Ocean bounded by New Zealand (NZ) in the southwest, Hawaii in the north, and Easter Island in the east (Barcham *et al.*, 2009). The settlement history of this region begins with the Out of Africa migration approximately 50-100 thousand years ago (ya) (Nielsen *et al.*, 2017). This migration flowed up the Levant and Arabian peninsula (45-55 kya) and then split into a migration to South Asia, Indonesia, and Australia (50 kya, Kivisild *et al.* (1999); Quintana-Murci *et al.* (1999)), with a migration into Europe under way 45 kya. From South Asia, the migration continued into East Asia (20 kya, Groucutt *et al.* (2015)).

Near Oceania is the geographic region from New Guinea and the Bismarck Archipelago, through to the Solomon Islands in the east, with the remainder of the Pacific Islands east from the Reef/Santa Cruz group, southeast of the Solomon Island being Remote Oceania (Matisoo-Smith and Gosling, 2018). The settlement of Near Oceania occurred first (40 kya), and was followed by the settlement of Remote Oceania, which reached as far as Samoan and Tonga (Western Polynesia) as part of the Lapita expansion (3 kya, (Matisoo-Smith, 2015; Skoglund *et al.*, 2016)). After this settlement, came the peopling of the Eastern Polynesia (including the Cook Islands) (1-1.2 kya; Wilmshurst *et al.* (2011)) and then finally, the settlement of New Zealand by NZ Māori 800 ya (Duggan and Stoneking, 2014; Matisoo-Smith, 2015).

In the 1950's and 60's there was a rural to urban migration of the Polynesian populations, for mostly economic reasons. As part of this, many individuals and families relocated from Samoa, Tonga, and the Cook Islands to New Zealand in search of work, and now the populations in New Zealand largely out-number the populations of their islands of origin (Matisoo-Smith, 2012). Within New Zealand, the largest number of Polynesian people is found within the Auckland region (Barcham *et al.*, 2009).

1.2 Metabolic syndrome

The World Health Organisation has previously defined Metabolic Syndrome to be glucose intolerance, or diabetes mellitus, and or insulin resistance combined with at least two of the following: impaired glucose regulation or diabetes, insulin resistance, raised arterial pressure, raised plasma triglycerides or low high-density lipoproteins to cholesterol levels, central obesity, and microalbuminuria (Alberti and Zimmet, 1998). Hyperuricaemia can also be considered a component but is not required for the condition to be recognised. hyperuricaemia is defined as a serum urate ≥ 0.42 mmol/L in males and ≥ 0.36 mmol/L in females (Choi and Ford, 2007). There has been a slight revision and refinement in the criteria for a diagnosis of metabolic syndrome (Alberti *et al.*, 2009). The requirement for diabetes mellitus and or insulin resistance is no longer used, instead a requirement of having 3 of the 5 co-morbidities is used.

Obesity is defined as a body mass index (BMI) of >30 kg/m² and is a contributing risk factor for many other diseases such as type 2 diabetes and cardio vascular disease (CVD) (Haslam and James, 2005). Obesity has been a key focus for research as the prevalence continues to increase and is posing a major threat to health systems in many countries (Wang *et al.*, 2011c). With respect to Pacific people obesity is a major problem, with an obesity rate of 68%. New Zealand Māori have an obesity prevalence of 48% (New Zealand Ministry of Health, 2013). From 2012 to 2013 the prevalence of obesity increased by 6% in Pacific populations.

Type 2 diabetes is diagnosed by a fasting blood glucose of >7.0 mmol/L or 2-hr post blood glucose load of >11.1 mmol/L (Alberti and Zimmet, 1998). The risk of developing type 2 diabetes is strongly correlated with excess weight (Rana *et al.*, 2007). The prevalence of diabetes is 13% in Pacific people and 7% in NZ Māori (New Zealand Ministry of Health, 2013) compared to 5.5% in Europeans (Winnard *et al.*, 2013). In New Zealand, the rate of diabetes is twice that in Māori compared to non-Māori, and the rate is 3.6 fold higher in Pacific compared to non-Pacific ethnicities.

1.2.1 Hyperuricaemia and gout

Hyperuricaemia can be caused by diet, genetic predisposition or under-excretion of urate. The elevated uric acid levels in the blood contributes to the forming of urate crystals in the joints, triggering an inflammatory response by stimulating the synthesis and release of humoral and cellular inflammatory mediators (Choi *et al.*, 2005). There is a clear relationship between hyperuricaemia and the causal role in gout (Snaith, 2004). There is also an increased prevalence in gout in different ethnicities.

New Zealand Māori and Pacific ethnicities have a prevalence of 12% and 14%, respectively, in males, compared to 4% in both European and Asian males (Winnard *et al.*, 2012, 2013). While New Zealand Māori and Pacific people have an increased prevalence of hyperuricaemia, they also have a genetic predisposition to elevated urate and risk of gout (Hollis-Moffatt *et al.*, 2009; Phipps-Green *et al.*, 2010; Hollis-Moffatt *et al.*, 2011). The prevalence of metabolic syndrome in US gout patients is ~60%, or nearly 3 fold higher than US adults without gout (Choi and Ford, 2007).

1.3 Genetic variation

As with many complex diseases, such as gout, type 2 diabetes, obesity, kidney disease, and metabolic syndrome, there is an environmental component and a genetic component. Genetic variation at a single locus, or at many, influences the genetic risk of an individual for developing disease. Genetic variation refers to the differences in DNA sequence that occur between individuals of a species. Types of genetic variation include, single nucleotide polymorphism (SNP) which are single base changes (or polymorphisms), and insertion or deletions (INDELs). INDELs are either additional DNA, or DNA that had been lost, and can be short (a single base, up to approximately 100 base pairs) or long (kilobases to megabases) at which point that can also become structural variants. Other types of genetic variation includes copy number variation (alterations in the number of copies of a gene from what would be expected from the ploidy of the organism), and re-arrangements such as translocations - segments of DNA moving to different chromosomes, or inversions - segments of DNA that are excised, inverted, and reinserted at the same location on the chromosome. Within population based genetic studies, such as case-control association studies, the prominent type of genetic variation that is studied are SNPs.

Common methodologies of genotyping, to determine the genetic variation of an individual, are SNP arrays or sequencing. SNP arrays, such as the Illumina Human CoreExome bead-chip contain probes designed for specific genetic variants that bind DNA samples, and after a single base extension step with a labelled nucleotide, enable a genotype to be determined from a fluorescence sensitive scanner (Wang *et al.*, 1998; Gunderson *et al.*, 2005; Shen *et al.*, 2005). Sequencing technologies, such as the short read sequencing performed by the next-generation sequencers, take DNA and randomly fragment it, then determine the bases sequentially in the fragment to approximately 150 bp from the start to create a ‘read’ (Metzker, 2010). The reads are then aligned to a reference genome and at each base of the genome a genotype is called by using the distribution of bases from reads overlapping that position (Li *et al.*, 2008a; McKenna *et al.*, 2010). One of the key differences between SNP arrays and sequencing is that SNP arrays capture known variation, whereas sequencing allows for unknown genetic variants to be genotyped.

1.4 Natural selection

Charles Darwin (1909) in *The Origin of Species* declared: “[*The] preservation of favourable individual differences and variations, and the destruction of those which are injurious, I have called Natural Selection or Survival of the Fittest. Variations neither useful or injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions.*” This theory set the scene for research in population biology. Selection acts upon the phenotype and this pressure transfers through to act upon the genetic contribution of that phenotype. Most important to Darwin’s Theory was it created models for what natural selection should look like and the methods for being able to detect selective events on the genetics of organisms. A focus on selection in humans has added insight into past events since migration out of Africa (Soares *et al.*, 2012) and yielded possibilities of the influences on human disease.

1.4.1 Neutrality and genetic drift

The random sampling of alleles can cause an increase or decrease in allele frequency of a population leading to genetic drift. Wright (1931) proposed genetic drift as an evolutionary model. He put forward that by chance, in a finite population without selection acting, an allele could increase in frequency over generations and time, changing significantly from the original frequency values. Fisher believed that “Natural Selection depends on the succession of favourable chances”, and thought that genetic drift was insignificant compared to selection (Fisher, 1930). In Fisher’s model, genetic variation would disappear at an extremely slow rate in a population without selection, and a moderate rate of new mutations would be enough to maintain the population variability (Fisher, 1922). Feller (1951) unified the population models from both Wright and Fisher into what is now known as the Wright-Fisher model. The Wright-Fisher model in its simplest form treats a diploid population of size N as a haploid population of $2N$. In unifying the population models Feller noted that the assumption of constant population size contributes more than was previously thought, and this had been crucial in Wright’s model for gene frequency to satisfy a diffusion equation.

Adding further to the genetic drift side of the discussion was the neutral theory of molecular evolution. The neutral theory of molecular evolution put forward by Kimura (1979a) states that most evolutionary change is not due to selection but is due to random genetic drift. The term “drift” was used because the variants under neutrality are expected to convey no advantage, and rise and fall (or drift) randomly in prevalence over time. Random genetic drift plays a role in population structure by keeping the number of co-existing alleles down (Kimura, 1955). One of the problems with the neutral theory was how the diversity of species could be accounted for. The solution to this can be explained through a positive correlation between increased genetic diversity and an increase in recombination, hinting at a rise in mutation rate as frequency of genetic exchange increases (Hellmann *et al.*, 2003). This link provided an explanation of a neutral process leading to increased genetic diversity.

Crucial to both the neutral theory and the theory of natural selection is the estimation of the population

mutation rate, commonly referred to as θ . The Watterson estimator, $\theta = 2 \times \text{ploidy} \times N_e \times \mu$, represents the expected number of new mutations for a population per generation (Watterson, 1975), where μ is the mutation rate and N_e is the effective population size. Effective population size is the size of an idealised population with the same amount of inbreeding or random gene frequency drift as the population under consideration (Kimura and Crow, 1963). Other estimators of θ include: $\hat{\theta} = F_{ST}$ (Weir and Cockerham, 1984), and $\hat{\theta} = E(\pi)$, in the infinite sites model, where π is the average pair-wise nucleotide differences per site (Tajima, 1996). The infinite sites model assumes only one mutation occurs per site.

The neutral model allows for natural selection, but proposes that the majority of the genome is selectively neutral (Kimura, 1979b). The methodology for detecting selection utilises the neutral hypothesis for many of the methods mentioned later (section 1.5). These methods use a model for neutrality, and if the statistic reaches a level that is considered to be beyond that of the neutral model, the model is rejected. Typically, the neutral hypothesis is a composite hypothesis using assumptions for which it would be difficult to find a human population that fulfils them, such as: “the population is in equilibrium at constant size with no population subdivision or gene flow from other populations” (Nielsen *et al.*, 2005).

1.4.2 Selection

Possible types of selection that can act upon a population include balancing, background, or directional selection. Balancing selection maintains both alleles in a population and is often the case when the heterozygote has an advantage over the homozygote and can be referred to as “over-dominant selection”. Two examples of this are MHC class 1 (Hughes and Nei, 1988) and sickle-cell anaemia (Allison, 1956). With sickle-cell anaemia, individuals homozygous for normal haemoglobin alpha are susceptible to malaria, individuals that are homozygote for the ‘sickling’ mutation develop sickle-cell disease, and the heterozygote has a protective effect against malarial infection (Aidoo *et al.*, 2002). “Background selection” is a reduction of genetic diversity at a locus because of deleterious selection acting on a separate but linked locus (Charlesworth *et al.*, 1993). Background selection can be thought of as the opposite effect to genetic hitch-hiking. Genetic hitch-hiking occurs when a locus is under selection, linked sites on the same chromosome will increase in frequency, causing a change in frequency of the alleles that are not on the same chromosome (Maynard Smith and Haigh, 1974). Linkage disequilibrium (LD) can arise at sites that are under directional selection, even when the loci are unlinked (Felsenstein, 1965) but, recombination will act to reduce this LD.

Directional selection operates in two fashions, positive or negative. Positive selection is where an allele conveys an advantage in fitness that enhances reproductive success, resulting in the beneficial allele increasing in frequency within the population. An example of positive selection is the lactase gene in European populations which enables the absorption of lactose (Bersaglieri *et al.*, 2004). Negative selection occurs when the allele conveys a fitness disadvantage and ends with reduced reproductive success, decreasing the frequency of the allele in the population. The cystic fibrosis gene *CFTR*, is an example of negative selection, where in males splicing mutations can lead to sterility (Pagani *et al.*,

2005).

The way selection can affect allele frequencies can be thought of as a sweep: as an allele is selected, it increases in frequency, “sweeping” through the population. The intensity of this sweep can be classified as, hard, soft, and partial or incomplete. A hard sweep, also known as a classical sweep, is when a beneficial allele rapidly rises in frequency and through hitch-hiking the neutral variation is simultaneously reduced at linked sites, as shown in Figure 1.1 (Maynard Smith and Haigh, 1974; Hermisson and Pennings, 2005; Nielsen, 2005; Hermisson and Pennings, 2017).

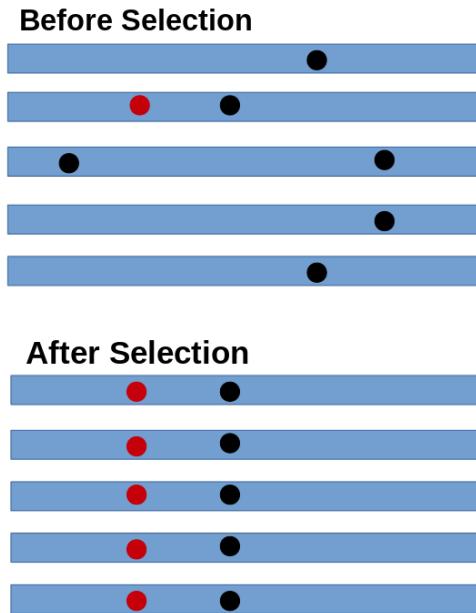


Figure 1.1: Representation of alleles before and after a hard sweep. Each rectangle represents a haplotype within a population. A new mutation is shown in red and neutral alleles are shown in black. After a hard sweep the new mutation becomes fixed in the population, along with the neutral allele that is linked to it.

A soft sweep occurs when an initially fitness-neutral allele evolving under drift experiences a beneficial environmental change and increases in frequency until fixation. The initial neutral nature of the allele prior to becoming beneficial means there is a range of surrounding haplotypes and there is not as much reduction in genetic diversity as happens with a hard sweep (Hermisson and Pennings, 2005; Schrider *et al.*, 2015; Hermisson and Pennings, 2017). The result of a soft sweep is multiple haplotypes of intermediate frequency around the selected site. If the selective environment for a population changes frequently enough, soft sweeps could be the main method of adaptation (Schrider *et al.*, 2015).

A partial or incomplete sweep occurs when part way through a sweep the selective pressure is changed/removed resulting in an increase in frequency for the allele that then returns to random genetic drift.

In order to determine if a selective event has occurred, one can attempt to find what are termed

“signatures of selection”. Population differentiation can sometimes be indicative of a selective event. If extreme levels of differentiation are observed at a locus between populations, this can be interpreted as evidence of positive selection. Regions of the genome that have a reduction in variability in a population can be the effect of a selective sweep. A third example of a signature of selection is deviation in the frequency spectrum for alleles beyond what would be considered possible under neutrality (Przeworski, 2002). A key signature of positive selection is the increase in proportion of high frequency variants (Fay and Wu, 2000).

Assuming a background of neutrality means that there needs to be a way to identify variants under selection as opposed to population demographic events, such as bottlenecks or expansions. In the case of Polynesian settlement, there is evidence for a founder effect which needs to be considered when interpreting selection results (Kayser *et al.*, 2006). It is for this reason that the unique features of a selective event need to be able to be identified from other possible features such as population demography or admixture. Population events and demography affect the entire genome, whereas selection acts on a localised region (Stajich and Hahn, 2005). To be able to recreate demography, the choice in the SNPs used in the calculation of neutrality and selection statistics is important as there is chance for a large ascertainment bias if SNPs are only selected from a small number of populations (Wall *et al.*, 2008). The SNPs used also need to be sampled away from genic areas.

1.5 Methods for detecting selection

Since 2005 there have been multiple reviews published (see Nielsen (2005), Sabeti *et al.* (2006), Utsunomiya *et al.* (2015), Vitti *et al.* (2013) and Haasl and Payseur (2016)) on methods of identifying selection, as there has been a renewed interest to examine selection at a genome wide scale due to increased availability of whole genome genotype data. Methods that have been developed to detect signatures of selection can be placed into two main categories, those that only work on a single population and those that compare between populations. There are also specific underlying genetic features that these methods can use, such as allele frequency (frequency spectrum methods) or differences in haplotype (linkage methods). Also, there are methods that combine different approaches (composite methods). A summary of selection methods is provided in Table 1.1.

The different methods for detecting selection work best over different time periods. Recent soft sweeps are able to be detected well with LD-based methods, but frequency spectrum tests struggle (Hermisson and Pennings, 2005). The opposite applies in the case of hard sweeps (Hermisson and Pennings, 2005). In humans, the method that is able to detect the most ancient selection, ranging to millions of years ago is the K_a/K_s ratio, however if a strong background of neutrality is present this can undermine the effectiveness of the test. Tajima’s D is able to detect selection events as old as 250,000 years ago in regions of low diversity containing an excess of rare alleles. Fay and Wu’s H is able to detect events less than 80,000 years ago based on derived alleles having arisen through mutation and then risen to a high frequency via hitch-hiking. F_{ST} , in the range of less than 50,000 to 70,000 years ago, picks up geographic separation of populations that have been subject to different physical and cultural environments based on the changes in allele frequency in one population versus another. Methods

Table 1.1: Summary of Selection Methods

Method	Detects selection within or between populations	Timeframe	Key reference
Macroevolutionary			
Ka/Ks	Within and Between	Ancient selection	Hughes and Nei (1988)
HKA	Within and Between		Hudson <i>et al.</i> (1987)
Haplotypic			
iHS	Within	< 30,000 years	Voight <i>et al.</i> (2006)
XPEHH	Between		Tang <i>et al.</i> (2007)
nSL	Within		Ferrer-Admetlla <i>et al.</i> (2014)
Frequency spectrum			
Tajima's <i>D</i>	Within	< 250,000 years	Tajima (1989)
Fay and Wu's <i>H</i>	Within	< 80,000 years	Fay and Wu (2000)
Fu and Li's <i>F</i>	Within		Fu and Li (1993)
Zeng's <i>E</i>	Within		Zeng <i>et al.</i> (2006)
F_{ST}	Between	< 50-70,000 years	Weir and Cockerham (1984)
ΔDAF	Between		Grossman <i>et al.</i> (2010)
SCCT	Within		Wang <i>et al.</i> (2014)
Composite			
CLR	Within		Kim and Stephan (2002)
XP-CLR	Between		Chen <i>et al.</i> (2010)
CMS	Within and Between		Grossman <i>et al.</i> (2010)
Meta-SS	Within and Between		Utsunomiya <i>et al.</i> (2015)
CSS	Within and Between		Randhawa <i>et al.</i> (2014)

relying on long haplotypes such as integrated haplotype homozygosity score (iHS) have the shortest range at less than 30,000 years ago for a long haplotype that has not had time to break down through recombination and contains a high frequency variant. Partial sweeps can also be detected down to a minor allele frequency (MAF) of approximately 10% (Sabeti *et al.*, 2006).

1.5.1 Macroevolutionary methods

1.5.1.1 K_a/K_s

The ratio of non-synonymous to synonymous mutations is called the K_a/K_s ratio (or d_N/d_S ratio as it is often referred to), and is used to infer functional impact from selection, since synonymous mutations are thought of as silent mutations with a neutral effect on the protein (Hughes and Nei, 1988). K_a is the number of non-synonymous differences between sequences normalised by the total number of non-synonymous sites in the sequence. K_s is likewise, but for synonymous differences normalised by the total number of synonymous sites in the sequence. Comparing the baseline rate of mutation (synonymous) to the non-synonymous rate gives an understanding of the tolerance of amino acid alternatives in the protein, with an excess of non-synonymous mutations indicating the novel protein structures are being favoured and the protein is being positively selected for. A K_a/K_s ratio greater than 1 is indicative of positive selection, while a value less than 1 is indicative of negative selection against deleterious mutations and for keeping the protein conserved.

1.5.1.2 HKA

The HKA test (Hudson *et al.*, 1987) compares between and within populations the polymorphism and divergence of two or more loci, and uses this to establish if a locus has been under selection, since variation within, and diversity between, species under neutrality should be based only on the mutation rate. The assumption that mutation rates at neutral loci are constant over time is important, because if the rate changes, it will affect the difference between the polymorphism and divergence. In practice the HKA test is powered for recent selection events, however, in practice it is difficult to find a putatively neutral locus (Zhai *et al.*, 2009).

1.5.2 Haplotypic methods

There are a few methods that are different derivations and applications of looking at haplotype decay around a core SNP. Sabeti *et al.* (2006) put forward the concept of using extended haplotype homozygosity (EHH). The haplotypes that had unusually long haplotype homozygosity and a high frequency in the population indicated that the polymorphism had risen in prominence faster than it would be expected under the neutral model and therefore under selection. The main concept is that a beneficial mutation would not solely be selected for, with the surrounding region initially “hitch-hiking” and therefore it would have a conserving effect. As time and generations progressed, this haplotype would decay due to recombination, and this haplotype decay can be used to assist in the ageing of the selection event. In order to use haplotypic methods, there is a requirement to establish the alleles that were inherited from the maternal side and those that were inherited from the paternal side constituting the two haplotypes (phasing).

1.5.2.1 Integrated haplotype homozygosity score

The iHS (Voight *et al.*, 2006; Szpiech and Hernandez, 2014) calculation starts with EHH, which is the probability that 2 chromosomes carrying a core haplotype are homogenous to the distance x . A value of 0 represents no homozygosity, meaning all haplotypes are different, while a value of 1 is complete homozygosity meaning all haplotypes are the same. The EHH value versus distance is integrated to find the area under EHH curve until EHH decays to 0.05. This integrated extended haplotype homozygosity (iHH) is calculated based on the ancestral or derived allele and annotated as iHH_A or iHH_D . The unstandardised iHS is calculated in equation (1.1), which is the `selscan` version (Szpiech and Hernandez, 2014). Standardisation of iHS is shown in equation (1.2), where the expectations and standard deviation of equation (1.1) are estimated from the empirical distribution of SNPs where the derived allele frequency (DAF) (see section 1.5.3.7) p matches the frequency at the core SNP (Voight *et al.*, 2006). Large positive values indicate long haplotypes carrying the derived allele, whereas negative values indicate haplotypes carrying the ancestral allele. The calculation in Voight *et al.* (2006) has iHH_A and iHH_D switched. The Szpiech and Hernandez (2014) iHS calculation is used in this thesis (equations (1.1) and (1.2)). IHS identifies variants under selection driven to intermediate frequencies.

$$iHS_{unstd} = \ln \frac{iHH_D}{iHH_A} \quad (1.1)$$

$$iHS_{std} = \ln \frac{iHH_D}{iHH_A} - \frac{E_p[\ln \frac{iHH_D}{iHH_A}]}{SD_p[\ln \frac{iHH_D}{iHH_A}]} \quad (1.2)$$

1.5.2.2 Cross population extended haplotype homozygosity

Cross-population extended haplotype homozygosity (XP-EHH), created by Tang *et al.* (2007), is also known as Rsb. XP-EHH was designed for use with SNP data rather than sequence. XP-EHH lacks power to detect intermediate frequency variants but is designed to detect variants that are near fixation or completely fixed. The EHH of SNP site (EHHS), integrates to sum the area under the decay by distance in EHH (iHH) which is then used as a log ratio between populations at a single site (equation (1.3)) and in a standardised form (equation (1.4)), with 1 and 2 indicating the population. Recombination rate is mostly conserved within populations, so provides an internal control in XP-EHH for effects of heterogeneous recombination rates. Extreme values of XP-EHH indicate a slower decay of EHH in one population compared to the other. XP-EHH relies on the breakdown of LD over time and has weak power to detect selective sweeps that were historical and ended thousands of generations ago (Chen *et al.*, 2010). F_{ST} is better than XP-EHH for differential selection between closely related populations because haplotype-based signals are mostly shared between geographically similar populations (Pickrell *et al.*, 2009).

$$XPEHH_{unstd} = \ln \frac{iHH_1}{iHH_2} \quad (1.3)$$

$$XPEHH_{std} = \frac{\ln \frac{iHH_1}{iHH_2} - E[\ln \frac{iHH_1}{iHH_2}]}{SD[\ln \frac{iHH_1}{iHH_2}]} \quad (1.4)$$

1.5.2.3 Number of segregating sites by length

Number of segregating sites by length is another single population haplotypic method that is extremely similar to iHS, however number of segregating sites by length (nSL) also measures the length of a segment of haplotype homozygosity between a pair of haplotypes in terms of number of mutations in the remaining haplotypes in the data set in the same region (Ferrer-Admetlla *et al.*, 2014). Demographic events affect nSL less than iHS. Under simulations for different demographic events, nSL had a smaller total difference in variation between the standard distribution of the test statistic and the distribution of the modelled demographic event (Ferrer-Admetlla *et al.*, 2014).

1.5.3 Frequency spectrum methods

The frequency spectrum methods are based on finding deviations in the frequency spectra of alleles that are outside of what would be expected under neutrality. These can involve an increase in particular types of frequencies such as low, intermediate, high variant frequencies, or a difference in allele frequencies between populations.

1.5.3.1 Tajima's D

Tajima's D (Tajima, 1989) counts the number of segregating sites of individuals in a population and finds the ratio between polymorphisms and pair-wise individual comparisons. Tajima's D is a test of the neutral hypothesis. The hypothesis is rejected when there is an excess of low frequency variants. A negative Tajima's D is indicative of positive selection or weak negative selection but a negative value can also be attributed to population events such as population expansion. Positive values of Tajima's D indicate an excess of intermediate frequency variants. An excess of intermediate frequency variants could also be attributed to balancing selection, population structure, or population bottlenecks (Kreitman, 2000; Nielsen, 2005; Nielsen *et al.*, 2007). The comparison of other polymorphisms, such as changes in SNPs versus changes in INDEL numbers can provide further evidence for either selection or a population event, since population events should affect both SNPs and INDELS in an equal fashion. The direction (positive, zero, or negative) of Tajima's D values after a population bottleneck depends on the strength and duration of the bottleneck and can generate similar Tajima's D values, and can be positive, negative or zero (Fay and Wu, 1999). After a population reduction or selective event, it is expected the population will have excess rare alleles (recent, low frequency) as the population recovers, therefore Tajima's D will be negative (Barton, 1998). Tajima's D is sensitive to sequencing errors that are called as SNPs because they appear equivalent to low frequency variants (Achaz, 2008).

1.5.3.2 Fu and Li's F

Fu and Li's F is the comparison of the number of derived singleton mutations and the mean pair-wise difference between sequences (Fu and Li, 1993). It can be calculated both with or without an out-group (F^*), with Fu and Li's F^* having the greatest power of the Fu and Li statistics (Fu and Li, 1993; Ramírez-Soriano *et al.*, 2008).

1.5.3.3 Fay and Wu's H

Fay and Wu's H is sensitive to SNPs rising to moderate to high frequency and uses an out-group (such as the ancestral reference sequence) to determine the derived/ancestral state of the allele. A positive value indicates a deficit in derived moderate to high frequency SNPs. A negative value indicates an excess of derived moderate to high frequency SNPs. Fay and Wu's H rejects neutrality when there is an excess of high but not low frequency variants. An increase in the proportion of high frequency variants compared to intermediate frequency variants is a unique signature of positive selection (Fay and Wu,

2000). Fay and Wu's H can be used in regions of low recombination to distinguish hitch-hiking from neutral or background selection where it is expected there will be the same level of intermediate to high frequency variants.

1.5.3.4 Zeng's E

In contrast to Fay and Wu's H and Tajima's D , which compare either high or low frequency variants with intermediate frequency variants, Zeng's E compares the low and high frequency variants (Zeng *et al.*, 2006). After a selective sweep, the lower frequency variants are expected to 'bounce back' faster than the higher frequency variants. The E test is not sensitive to population subdivision. A negative value for Zeng's E is indicative of a selective sweep, but a rapid population expansion can also produce a negative value.

1.5.3.5 F_{ST}

F_{ST} is a population differentiation metric that measures the differences in heterogeneity of a subpopulation compared to a global population. Initially created by Wright (1951), as an inbreeding coefficient for inbreeding as a whole in a population (equation (1.5)), F_{IT} is the inbreeding coefficient for the individual in the total population and F_{IS} is the inbreeding coefficient for inbreeding of the individual in the subpopulation. Both F_{IT} and F_{IS} compare the observed heterozygosity of the individual with the expected heterozygosity in either the total population (F_{IT}), or the subpopulation (F_{IS}). These fixation indices provide ways of summarising population structure. Selection will act to drive an allele to fixation, whereas immigration usually reduces the frequency. When the selective advantage is strong, selection will be the dominant force on the locus, however, when the selective advantage approaches zero, the allele will begin to act neutrally and disappear with a probability of one minus its frequency (Kimura, 1968). Cockerham (1969; 1973) equated θ (population genetic differentiation) with F_{ST} , extending F_{ST} from calculating population structure to population differentiation too. The fixation indices were further developed by Weir and Cockerham (1984) to add population size weightings to the calculation. F_{ST} values fall between 0 and 1, each representing the fixation of an allele. Large differences in F_{ST} values are indicative of population differentiation, suggesting directional selection at that locus, whereas small differences in F_{ST} indicate that the populations are similar. F_{ST} can have high variation in neighbouring loci under neutrality (Weir *et al.*, 2005). Unless populations are closely related, such as northwest and southeast Europeans (Price *et al.*, 2008), the noise will drown out the signal, and identification of genome-wide significance in F_{ST} will be difficult. Weir *et al.* (2005) suggested taking a sliding window approach to help improve the signal to noise ratio.

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}} \quad (1.5)$$

F_{ST} has been used in multiple studies as evidence for natural selection (Myles *et al.*, 2007; Pickrell *et al.*, 2009; Akey, 2012). Akey (2012) used F_{ST} at individual loci and compared this to the distributions for the genome, chromosome and gene levels. The use of the empirical distribution to compare to F_{ST}

for each SNP controlled for the effects of demography. F_{ST} as a method for detecting selection was developed further into d_i , which is a function of pair-wise F_{ST} between population i and the remaining populations (Akey *et al.*, 2010). Outside of the human context F_{ST} has been a very popular method of reporting both population structure and evidence of selection (Akey *et al.*, 2010; Hancock *et al.*, 2011; Qanbari *et al.*, 2011; Wei *et al.*, 2015). The population branching statistic is a F_{ST} based statistic that involves the use of an out-group to compare estimates in divergence time between populations (Yi *et al.*, 2010).

1.5.3.6 Selection by conditional coalescent tree

Selection by conditional coalescent tree (Wang *et al.* (2014)) detects recent positive selection by searching for an imbalance of genetic variants and conditions these on the allele frequencies of the candidate loci. Selection by conditional coalescent tree pinpoints the causal variant more accurately by using a method that is based on the conditional coalescent tree method from Wiuf and Donnelly (1999), where haplotypes are partitioned into two subgroups according to the allelic state at a particular locus. It is assumed one subgroup carries the derived allele at the locus and the other subgroup carries the ancestral allele. The coalescence of a sample of haplotypes, conditioning on the particular mutation, means each group should coalesce together individually before the two subgroups coalesce. This is a similar idea to that of EHH used by iHS, and the results of iHS and selection by conditional coalescent tree tend to be similar. The selection statistic (S) is based on comparing the lineage lengths of the two groups. Under neutrality S approaches θ , with significant departures from θ being considered evidence of positive selection. When natural selection has a strong hitch-hiking effect on a selectively favoured allele, selection by conditional coalescent tree (and iHS) is more powerful than Tajima's D , Fay and Wu's H , and measuring change in derived allele frequency (see 1.5.3.7). In most cases the power of selection by conditional coalescent tree is equivalent to iHS except when selection is very strong, in which case iHS has more power (Wang *et al.*, 2014). The power of selection by conditional coalescent tree can be improved by increasing the sample size. Selection by conditional coalescent tree performs moderately with population demographic events (such as bottlenecks) and out performs Tajima's D and iHS in false discovery rate, although it does not perform as well as Fay and Wu's H (Wang *et al.*, 2014).

1.5.3.7 Change in derived allele frequency

Similar to F_{ST} , change in derived allele frequency (ΔDAF) looks at population differentiation. The derived allele is determined through the use of an ancestral out-group such as chimpanzees when used in human populations. DAF itself is calculated by determining the ancestral and derived state for an allele and calculating the allele frequency with regard to the derived state. ΔDAF measures the absolute difference in the derived allele frequency between two populations (Grossman *et al.*, 2010) and has greater power to detect sweeps, both partial and complete, and outperforms XP-EHH for partial sweeps (Colonna *et al.*, 2014). ΔDAF is part of the composite of multiple sites statistic (Grossman *et al.*, 2010) but has also been used as a standalone metric in humans (Colonna *et al.*, 2014;

Gudbjartsson *et al.*, 2015) and in other species such as cattle (Randhawa *et al.*, 2014) to measure differentiation.

1.5.4 Composite methods

Composite methods combine multiple methods with the goal of increasing overall power and/or spatial resolution for detecting selection. The combination of methods also gives greater resilience against demographic events.

1.5.4.1 Composite likelihood ratio

The composite likelihood ratio (CLR), created by Kim and Stephan (2002), is a test for local hitch-hiking selection along a recombining chromosome. It calculates a null distribution of variation from neutral coalescent simulations involving recombination, and uses this to calculate the probability of observing a threshold ratio of segregating variants. The Nielsen *et al.* (2005) version of composite likelihood ratio is similar to Kim and Stephan's test but creates the null distribution from the data itself, instead of using the neutral population model. A key consideration for the use of composite likelihood ratio is that it is very sensitive to SNP ascertainment bias (Chen *et al.*, 2010).

1.5.4.2 Cross population composite likelihoods ratio

The cross-population composite likelihoods ratio (XP-CLR) is an extension of the composite likelihood ratio test developed by Nielsen *et al.* (2009) that scans for multi-locus allele frequency differentiation between two populations and then tests if these frequency differences are too recent to be consistent with neutrality based on the LD in the region (Chen *et al.*, 2010). A down-weighting of SNPs that are in perfect or high LD is performed because they essentially provide the same information. By doing this it is hoped that false positives are reduced, since the correlation of marginal likelihood terms in the composite likelihood function is often overlooked and leads to false positives. Normalised cross-population composite likelihoods ratio scores should be resistant to variation in demographic events, however a conservative approach of using rank-order of scores across the genome is usually taken. Unlike single population composite likelihood ratio and similar to XP-EHH, cross-population composite likelihoods ratio is resistant to SNP ascertainment bias because population differentiation is not affected by ascertainment bias. There is also a strong correlation between genes identified with cross-population composite likelihoods ratio and XP-EHH. cross-population composite likelihoods ratio is also considered an extension of F_{ST} and is beginning to be used as a replacement for investigating selection in animal studies (e.g., Cattle – Lee *et al.* (2014); Goats - Benjelloun *et al.* (2015))

1.5.4.3 Composite of multiple signals

Composite of multiple signals (CMS) is a combination metric composed of iHS, XP-EHH, F_{ST} , ΔDAF , and ΔiHH . ΔiHH is used to compare the actual length of the haplotype. Composite of multiple signals

was created to take advantage of the combined power gained from using all metrics together to be able to detect the source of the selective signal better than any test individually could (Grossman *et al.*, 2010). In composite of multiple signals, it was found that XP-EHH and F_{ST} contributed mostly to spatially locating the causal variant whereas iHS, ΔDAF, and ΔiHH did poorly in spatially locating the selective signal but performed well at identifying the precise causal variant.

1.5.4.4 Meta-analysis of multiple tests

Meta-analysis of multiple tests (meta-SS) (Utsunomiya *et al.*, 2013, 2015) is only able to combine selection statistics that generate P-values. It uses an adapted version of a weighted Stouffer method (Stouffer *et al.*, 1949; Zaykin, 2011) to combine Z-transformed P-values. For each marker from each test *i* the respective P-value is transformed into a Z score by $Z_i = -\phi^{-1}(1 - \pi)$, where ϕ is the normal cumulative density function. Each test is weighted (w_i) to 1/n where n is the number of comparisons. The combined statistic of *k* tests, for each SNP in each population is defined in equation (1.6).

$$metaSS = \frac{\sum_{i=1}^k \omega_i Z_i}{\sqrt{\sum_{i=1}^k \omega_i^2}} \quad (1.6)$$

1.5.4.5 Composite of selection signals

Composite of selection signals (CSS) is similar to meta-analysis of multiple tests, in that it is a method for combining the results from multiple selection tests but is not limited to tests that produce P-values (Randhawa *et al.*, 2014). Composite of selection signals currently uses F_{ST}, XP-EHH, and ΔDAF or change in selected allele frequency, and combines them by taking the test statistic for each method across each SNP(1, ..., n) then obtaining the rank of each statistic across each SNP(1 ... n) and converting these into fractional ranks so they lie between 0 and 1, that is, 1/(n+1) to n/(n+1). By doing this the magnitudes of the original statistics are not used, thus increasing the robustness. The fractional ranks are converted to Z-statistics and the average z-values are calculated at each SNP position. P-values are obtained from the distribution of the means with the -log10(p) being the composite of selection signals value. The composite of selection signals is plotted against genomic position with an excess in the number of values at a position showing a common signal between the multiple test statistics.

1.5.5 Power

Power to detect selected sites can be very limited if the site is not an outlier from empirical methods, such as simulations or using permutations of the data, as they would not ‘stand out’ (Teshima *et al.*, 2006). A key consideration when selecting a particular method for detecting selection is the limitations the method has on power to detect a selective event, such as starting and final allele frequency of the

variant after a sweep. Tajima's D is powered for low and intermediate frequency variants (Simonsen *et al.*, 1995) whereas Fay and Wu's H is powered for intermediate and high frequency variants. Some methods such as iHS, are good for detection of middle frequency variants. XP-EHH, on the other hand is powered at the fixation ends of the spectrum. XP-EHH, because it is correlated with LD, lacks power to detect ancient selective events (Chen *et al.*, 2010). Sample size is another factor to consider. IHS has modest reduction of power when reducing samples until approximately 40 samples of single chromosomes. XP-EHH can maintain power down to about 20 samples of single chromosomes, so long as the reference population is a fixed size. The grouping of similar genetic populations may be an option to increase power with low sample sizes (Pickrell *et al.*, 2009)). For F_{ST} , power to detect selection should be sufficient if sample size is $> 1/F_{ST}$ (Bhatia *et al.*, 2011, 2013)

Power of a statistic also needs to be considered for the type of selective sweep that may have occurred, such is the case with nSL, which has similar power to detect selection to iHS, but performs better for larger starting allele frequencies. It also outperforms iHS in power in both hard and soft sweep scenarios (Ferrer-Admetlla *et al.*, 2014). When it comes to picking up demographic events, haplotype tests are useful at detecting recent and more moderate bottlenecks, frequency spectrum tests (such as Fay and Wu's H / Tajima's D) have the best power for detecting moderately ancestral severe bottlenecks (Depaulis *et al.*, 2003), whereas Zeng's E and Fu and Li's F have the most power during population expansion (Zeng *et al.*, 2006; Ramírez-Soriano *et al.*, 2008). Assessment of power to detect selection has previously been done through simulation or the use of well established loci as positive controls (Voight *et al.*, 2006; Zhai *et al.*, 2009; Cadzow *et al.*, 2016). When simulating, it important to generate scenarios that include different demographic events as well as selection events. Simulated models are generally done in one of three ways, as a coalescent model, or a forward-in-time model, or by resampling (Yuan *et al.*, 2012). The coalescent model (and most commonly used) works backwards-in-time to find the most recent common ancestor, and then permutes through models to arrive at the current state (Kingman, 1982; Yuan *et al.*, 2012). The forward-in-time simulations will take a starting ancestral population and then apply a population model, iterating through subsequent generations (Peng and Amos, 2010; Yuan *et al.*, 2012). Another approach can be to take the observed data and shuffle and resample it to generate empirical distributions. From each of these methods false positive and false negative rates can be calculated.

1.5.6 Challenges

When scanning genome-wide in a population (such as in a genome-wide association study (GWAS), or genome-wide selection scan), the variability in the genome also has population demographic events that have acted upon it, such as population bottlenecks, migration, or growth. Mutation and recombination are also acting on the genome. Methods based on comparing non-synonymous to synonymous variants are relatively unaffected, but methods based on allele frequency are susceptible to detecting population demographic events, along with mutation and recombination rate changes (Nielsen *et al.*, 2009). A meta population is a population which can be subdivided into many different sub-populations among which there is some pattern of migration, extinction, and recolonization (Wakeley and Aliacar, 2001). Bottlenecks with a meta-population model can lead to a high frequency of derived alleles, greater than

would be expected to arise through a neutral model (Jensen *et al.*, 2005) and as previously mentioned by Fay and Wu (2000) about high frequency derived alleles being a unique signature of selection, it is not actually unique (Przeworski, 2002).

SNP ascertainment bias is caused by the non-random sampling of SNPs on array chips. This is a problem that affects the selection detection methods to different extents, with frequency spectrum methods being the most susceptible. Tajima's D is biased upwards due to the SNP discovery process having ascertainment biases, which leads to an excess of intermediate frequency alleles in the sample (Kelley *et al.*, 2006). Tests that rely on long haplotypes are less susceptible to ascertainment bias (Sabeti *et al.*, 2006) but because long haplotypes can be quickly broken down through recombination, they are only useful for short time periods. SNP ascertainment bias can be overcome with the use of whole genome sequencing (Albrechtzen *et al.*, 2010).

Many of these statistics were developed to investigate selection at a particular locus, at a time when applying the statistics across a genome was not possible. Up until recently, genome scans have been performed using SNP genotypes identified through a SNP discovery process, which means they are not random samples – which would include non SNP genotypes (such as INDELs) and create an ascertainment bias (Nielsen *et al.*, 2005)

Identification of exactly what is deemed as being “significantly selected” has issues. The current practice is to take outlier loci, usually from an empirical distribution of the selection statistic and report these as significant. This is not necessarily indicative of selection, as cut off levels are created subjectively rather than derived from the model (Qanbari *et al.*, 2012). Using an empirical approach will result in many false positives, however this is considered an acceptable approach for selecting candidate genes so long as it is realised that the false positive rate is high (Kelley *et al.*, 2006).

In order to use haplotypic methods, (re)construction of haplotypes is required. Currently this involves the need for the phasing of haplotypes from genomic data. Phasing is the process of determining which alleles are found together on the same physical chromosomes. There are a number of ways that this can be done, such as through a pedigree, through the use of large population data, or using the reads from next generation sequencing. Phasing through pedigree information is utilised in most agricultural species, whereas for human populations the population genotypes are largely from unrelated individuals. Phasing of unrelated individuals currently relies on probabilistic methods (Browning and Browning, 2009; Delaneau *et al.*, 2012, 2013) in order to make the best guess at each individual haplotype. Newer methods have been developed to use the reads from next generation sequencing and are able to trace a haplotype across multiple overlapping reads and be used to complement other probabilistic phasing methods, however these read-backed methods are dependent on the depth of coverage and read quality (Delaneau *et al.*, 2014).

1.6 Improving GWAS

GWAS is a method to scan genomes, usually SNP genotype data, for association with a phenotype. Association is usually tested between a trait and a single SNP using either linear regression (continuous

variable) or logistic regression (binomial variable), with confounders included where possible as covariates to the models. These association tests are repeated for each SNP across the genome (for a comprehensive explanation of GWAS methods see Bush and Moore (2012)). Commonly an additive model is used for the SNP. The strength of GWAS is that prior knowledge of loci is not required and can be used as a tool for hypothesis generation. GWAS, because of the large number of association tests, require large cohorts of samples to achieve sufficient power for statistical significance. As the effect size for a variant decreases, larger sample sizes are also needed, as power is proportional to the square of the effect size (Hemani *et al.*, 2013; Korte and Farlow, 2013). Limitations in recruiting sufficient numbers of participants for GWAS therefore requires new ways of increasing power to detect smaller effects without needing to increase sample size, as outlined in this section. The goal of being able to prioritise GWAS results by using additional information, such as incorporating LD or selection statistics by using a weighting system to increase the power of a GWAS, has been suggested (Roeder *et al.*, 2006; Ayodo *et al.*, 2007). When LD between observed SNPs and causal variants is high, such as in the case of selection, greater power in GWAS can be achieved by focusing on searching for non-additive variance (Hemani *et al.*, 2013).

The basis of GWAS studies is to identify SNPs that are associated with differences in the mean of a trait. These studies do not necessarily identify the causal SNP but instead identify a marker for the causal variant, usually in LD. Fisher (1919) showed that total genetic variance could be split into components of additive, dominance and epistatic effects. The additive component has been shown to contribute the most to the overall genetic variance, with epistatic variance contributing little in an outbred population (Hill *et al.*, 2008). Complex traits gain their complexity from being the sum of interactions from many small effect loci, or the product of non-additive interactions between loci and the environment, making it challenging to find the exact contribution of a locus to a trait (Fu *et al.*, 2013). These small, true effect loci can end up non-significant due to the noise in the data but might be retrieved using methods that find non-additive effects.

Explained heritability is defined as the proportion of heritability (or phenotypic variance) explained by a collection of genetic variants, with the “missing” portion referring to the unaccounted genetic variance that has been estimated through family studies. When the ratio of explained heritability to total heritability is < 1 , the unexplained component is defined as missing heritability (Zuk *et al.*, 2012). The use of methods that capture additional loci with non-additive effects should be able to explain more of missing heritability.

1.6.1 Selection and GWAS

The significance of GWAS results is greatly increased when selection statistics and association studies are combined, this has been applied in a GWAS for malarial resistance variants (Ayodo *et al.*, 2007). A smaller sample size can be used if there is evidence of positive selection because the beneficial allele will rise in prevalence and be in high LD with other variants (due to genetic hitch-hiking) which can act as proxy markers (Karlsson *et al.*, 2014). The converse would apply for negative selection. Complex disease traits compared to Mendelian disease associated traits have a bias towards larger mouse-human

K_a/K_s ratios, by comparing the K_a/K_s ratios for ortholog genes between mouse and human, which suggests that evidence of positive selection could be utilised in identifying variation associated with complex disease (Thomas and Kejariwal, 2004).

1.6.2 Pathway analysis

A biological pathway is a dependency graph detailing the genes or proteins involved in a series of biochemical reactions, such as a metabolic pathway like the Krebs cycle. There are curated databases of many of the biological pathways used by living organisms, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017) or Reactome (Fabregat *et al.*, 2018). Pathways have been hypothesised to be involved in the genic adaptation of populations since altering a gene in a biochemical pathway is likely to alter different genes in the same pathway (Bigham *et al.*, 2010). The use of pathway analysis rather than focusing on individual SNPs is another way to increase the power of a GWAS (Jia *et al.*, 2011). This increase in power is largely obtained through a reduction in dimensionality in ways that attempt to remove noise from the underlying biological signal. There are however limitations to the effectiveness of using such an approach, such as the selection of pathway database, density of SNPs per gene, or gene length and boundaries. The pathway database is important because pathways can differ based on curator, age of the database - with older databases not incorporating current knowledge, or differing focuses. SNPs per gene, and gene length or boundaries all affect the pathway analysis as smaller genes or SNP density reduces the observations for the gene compared to large genes or genes with high SNP density. Pathways that involve many large genes also have a higher chance of a SNP having an extreme statistic value under a null model, because overall, they are more likely to have more SNPs than pathways containing small length genes.

1.7 Applications

1.7.1 Genome-wide selection

Genome-wide selection in humans has been investigated by Sabeti *et al.* (2006), Voight *et al.* (2006), Hancock *et al.* (2008), Pickrell *et al.* (2009), the 1000 Genomes Project Consortium (1000 Genomes Project Consortium, 2010, 2012, 2015), Grossman *et al.* (2010), Colonna *et al.* (2014), and Mallick *et al.* (2016). The population datasets used involved the HapMap Project (The International Hapmap Consortium, 2005), the Human Genome Diversity Project (Cann *et al.*, 2002; Rosenberg *et al.*, 2002) and the 1000 Genomes Project (populations are defined in section 2.2; 1000 Genomes Project Consortium (2010)). The 1000 Genomes Project Consortium (1000 Genomes Project Consortium, 2010, 2012, 2015) reported genome wide diversification between main population groups based on F_{ST} and ΔDAF . Genome wide mean F_{ST} was found to only be a maximum of 8%, indicating overall little genetic differentiation between populations, however several thousand SNPs had large F_{ST} indicating local population adaptation (1000 Genomes Project Consortium, 2010). The analysis also found 139 non-synonymous SNPs with large frequency differences between populations. Investigation of the 1000

Genomes Project (1KGP) data set was also performed by Colonna *et al.* (2014). Using Δ DAF, it was found that sites of high population differentiation (Δ DAF > 0.7 between and > 0.25 within continents) clustered together, the same was found for sites of low differentiation. Genes containing highly differentiated sites were likely to also have additional evidence of positive selection in a third. There was also a strong association of highly differentiated sites with genes and gene regulatory elements.

Metabolic syndrome and gout have associations and interactions with genes in metabolic pathways and the immune system (Osborn and Olefsky, 2012). The categories of genes that have displayed evidence of selection from the previously mentioned genome-wide selection scans include: metabolism of sugars (Voight *et al.*, 2006; Tang *et al.*, 2007), drug metabolism (Tang *et al.*, 2007), immune system (Tang *et al.*, 2007; Grossman *et al.*, 2010), and climate adaptation (Hancock *et al.*, 2008). Genes that play a role in adaptation to climate extremes are likely to play a role in metabolic disorders that make up metabolic syndrome (Hancock *et al.*, 2008). In contrast, variation in metabolic related genes can lead to elite phenotypes in athletes, showing that metabolic genes also impact on fitness (Ahmetov *et al.*, 2009).

1.7.1.1 Selection in Polynesia

The inclusion of Oceanic populations in multi-population studies has had a focus on genetic structure, measured via F_{ST} for differentiation to investigate population ancestry, rather than selection (Friedlaender *et al.*, 2008; Tennesen and Akey, 2011). The migratory history of Polynesian populations suggests that the points of differentiation are likely the events possibly occurring since the out of Africa dispersal \sim 60,000 years ago (Soares *et al.*, 2012) with settlement of Near Oceania \sim 40,000 years ago, and Remote Oceania \sim 3,100 years ago (Matisoo-Smith and Robins, 2004) and the Polynesian Triangle (bounded by New Zealand, Easter Island, and Hawaii) \sim 1,000-1,200 years ago (Wilmshurst *et al.*, 2011). The time frame for these events means that nearly all the previously mentioned methods would still be within their powered time frames to detect selective events. The haplotypic methods would possibly not be able to detect selective events as old as the Africa dispersal and may be at the fringe of the time period for the Near Oceania settlement.

Large population comparison studies have used Oceanic populations, but analyses were limited to Melanesian and Papuan populations because the cohort used was the Human Genome Diversity Project (Cann *et al.*, 2002) which did not include Polynesian populations. The Melanesian and Papuan populations are often used to infer the genetics of Polynesian populations. This extrapolation does not take into account the current ancestry predictions of Polynesians populations being more closely related to Asian/Taiwanese Aboriginal populations than Melanesian groups, or the presence of European admixture (Friedlaender *et al.*, 2008). Genomic DNA and mitochondrial DNA suggest Polynesians have East Asian ancestry, with substantial male Melanesian admixture (Kayser *et al.*, 2008). There has been limited analysis of selection in Polynesian populations beyond Kimura *et al.* (2008), where 24 Tongan samples in conjunction with other populations were analysed using both F_{ST} and a modified EHH test. Genomic regions with evidence of having been selected in the Tongan samples were potential

candidates for increased fat, muscle and bone masses (Kimura *et al.*, 2008).

A few candidate-genes for selection which also have been associated with obesity and type 2 diabetes in Polynesian populations have been investigated, such as the *PPARGC1A* (Myles *et al.*, 2011) and *CREBRF* (Minster *et al.*, 2016) genes. The “Thrifty-gene hypothesis” (Neel, 1962) has been the reasoning behind why these loci may have undergone selection. The hypothesis was originally conceived to attempt to explain the difference in prevalence of type 2 diabetes, that is, given that there should be a strong selective pressure against it, there must be some genetic advantage. This hypothesis has been challenged in the Polynesian context as fitting a particular narrative (Gosling *et al.*, 2014, 2015; Cadzow *et al.*, 2016) with an alternative hypothesis being a link between the metabolic diseases and the innate immune system that is under selective pressure via an infectious agent, such as malaria, leading to disease susceptibility in these populations (Gosling *et al.*, 2015).

1.7.2 Selection of metabolic syndrome and urate

The genetic influence of urate and gout is found in the under-excretion of serum urate. There are 28 loci that have been identified as having a role in hyperuricaemia and gout, of which, two (*ABCG2* and *SLC2A9*) are responsible 3.4% of the variability of urate (Köttgen *et al.*, 2013). *SLC2A9* encodes a urate transporter expressed on both the apical and baso-lateral membranes in the proximal tubule of the kidney and reabsorbs urate (Mandal and Mount, 2015). *ABCG2* encodes an ATP powered urate transporter and is responsible for urate excretion mostly in the gut (Maliepaard *et al.*, 2001; Mandal and Mount, 2015). Twenty-six other loci associated with serum urate explain a further 3.6% of the variance (Köttgen *et al.*, 2013). The heritability of serum uric acid ranges from 40 - 70% (Yang *et al.*, 2005; Nath *et al.*, 2007). This difference in explained variability and total heritability leaves a large component yet to be explained. This missing heritability could be reduced through identifying additional variance attributed to non-additive effects.

There have been multiple genetic loci associated with type 2 diabetes (Billings and Florez, 2010) and BMI (Locke *et al.*, 2015a), however, Chen *et al.* (2010) found no evidence for enrichment in selection for BMI, but found significant enrichment for type 2 diabetes. The genes associated with type 2 diabetes have been found to be differentiated by population, however the individual SNPs are often not (Pickrell *et al.*, 2009). Koh *et al.* (2014) identified six SNPs in or near nine genes that had been reported as type 2 diabetes associated and overlapped with evidence of positive selection in East Asian populations. Common variants that have been found in European populations associating with obesity do not replicate in the Samoan population, although this is likely due to power (Karns *et al.*, 2012). The non-replication of European variants in Polynesians and the high prevalence of obesity in Polynesian populations suggest there is likely to be some population specific variants that are yet to be identified, similar to the Polynesian specific variant in *CREBRF* (rs373863828), associated with obesity and protection for type 2 diabetes, identified in the Samoan population (Minster *et al.*, 2016). Rare variants tend to be recent and therefore geographically restricted (1000 Genomes Project Consortium, 2012).

The benefits of urate are in line with the categories found to be enriched by selection in genome wide

scans. Urate is an anti-oxidant in the blood, acting to protect not just erythrocytes but also T and B lymphocytes and macrophages (Ames *et al.*, 1981). Urate has also been hypothesised to have had a survival advantage during the Miocene period, when a series of mutations in urate oxidase occurred, leading to elevated urate. Hyperuricaemia enables blood pressure to be maintained under low salt dietary conditions (Watanabe *et al.*, 2002). Serum urate is also a potent anti-oxidant which accounts for >50% of the anti-oxidant activity of the blood (Glantzounis *et al.*, 2005; Parmar, 2009). Later in life elevated serum urate is associated with increased cognitive function (Euser *et al.*, 2009). Urate crystals activate the NLRP3 inflammasome which plays an important role in systemic infection and sepsis (Opitz *et al.*, 2009). The role of urate in these situations shows that while hyperuricaemia is now considered as contributing to disease (such as gout), in the past it may have influenced longevity and survival, therefore increasing reproductive success and being a phenotype that has undergone selection.

An elevated prevalence of metabolic syndrome co-morbidities and hyperuricaemia against a unique genetic background of Polynesian populations provides an important opportunity to study the role and effects that selection might have played on these metabolic conditions.

1.8 Research purpose and aims

This thesis seeks to understand selection in the context of Polynesian populations, and the effect it may have played in the metabolic disease burden that is present in these populations. There are three research chapters, each with their own aims. In Chapter 3: “Positive Selection in Polynesian Populations”, the aims were to discover and characterise regions that have undergone selection in Polynesian populations. In Chapter 4: “Clustering of Selection Statistics”, the aims were to investigate potential shared ancestry regions that are relevant to urate and metabolic disease. And in Chapter 5: “Selection and Association Studies”, the aims were to investigate the use of gout definition in genetic association studies, and the use of selection statistics in conjunction with GWAS results.

It is hoped that the investigation of genetic selection in Polynesian populations will further understanding of the cause of health disparities due to metabolic disease in these populations.

Chapter 2

Methods and data

This chapter describes the methods and datasets used in this thesis. Section 3.2 covers the methods used for the preparation of the datasets, such as haplotype phasing, principal component analysis (PCA), and file format conversions, as well as the methods used for the analysis of the datasets, such as the calculation of selection statistics. Section 2.2 contains a description of the two main data sources that were used and how they were transformed for the analysis.

2.1 Methods

All genomic resources and co-ordinates reported in this project use the human genome reference build GRCh37, unless otherwise specified.

2.1.1 Phasing

Haplotype phasing is a method of establishing the parental origin of haplotypes. It looks for the co-location of alleles to form haplotypes. In diploids such as humans, there is a maternal and a paternal haplotype at any given position in the genome. Selection statistics such as iHS require phased data in order to calculate the degradation of haplotypes by recombination. Phasing can be performed through the use of trios (parents and offspring), or it can be done through probabilistic phasing using large population haplotype reference panels (such as created from the 1KGP, see section 2.2.1 and Table 2.4) and probabilistically determining the most likely haplotype(s) for an individual given a set of markers.

Phasing was performed using SHAPEIT2 v2.r837 (Delaneau *et al.*, 2013) using the 1KGP reference haplotype panel (Table 2.4). There were two steps involved for phasing of genotypes. The first was to check the markers in each data set were matched with the markers in the reference haplotypes. Markers were marked for exclusion by this step for reasons such as the marker not being in the reference panel, or if the marker had different allele types to that in the reference panel. An example of this step is provided in the following code.

```

# Check alignment of markers against reference haplotypes
# ? = chromosome
shapeit2 \
-check \
-M genetic_map_chr?_combined_b37.txt \
--input-vcf coreExome_norm.chr?.vcf.gz \
--input-ref 1000GP_Phase3_chr?.hap \
1000GP_Phase3_chr?.legend \
1000GP_Phase3.sample \
--output-log coreExome_norm.chr?.checked \
-T 12

```

The second step was the phasing of genotypes into their haplotypes by using the 1KGP phased haplotype reference and a genetic map that was provided as part of the reference files. Markers that were marked for exclusion from the previous step were excluded. The most likely pair of haplotypes were outputted. The haplotypes were then converted from the haps format¹ into a phased variant call format (VCF)². These two steps are shown in the following code.

```

# Run phasing against reference haplotypes
# Exclude misaligned markers
# Use 8 threads
# ? = chromosome
shapeit2 \
-M genetic_map_chr?_combined_b37.txt \
--input-vcf coreExome_norm.chr?.vcf.gz \
--input-ref 1000GP_Phase3_chr?.hap \
1000GP_Phase3_chr?.legend \
1000GP_Phase3.sample \
--output-max coreExome_norm.chr?.phased \
--exclude-snp coreExome_norm.chr?.checked.snp.strand.exclude \
-T 8

# Convert haps to vcf
shapeit2 \
-convert \
--input-haps coreExome_norm.chr?.phased \
--output-vcf coreExome_norm.chr?.phased.vcf

```

¹https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#formats

²<https://samtools.github.io/hts-specs/VCFv4.2.pdf>

2.1.2 Selection Statistics

2.1.2.1 PopGenome

The PopGenome package (v2.2.3, Pfeifer *et al.* (2014)) for R was used for the calculation of Tajima’s D , Fay and Wu’s H , Fu and Li’s F , and Zeng’s E . Popgenome calculates the D^* and F^* version from Fu and Li (1993) which do not require an outgroup that should be a closely related population or species. The selection and neutrality statistics were calculated using a sliding window approach with a window size of 100 kb and a slide of 10 kb. Windows that had fewer than four segregating sites were filtered out, and window co-ordinates were altered to be +/- 5 kb from the centre of the window. The empirical distribution for each statistic, for each population, was used to create a significance threshold. The lower threshold was $\min(1^{\text{st}} \text{ percentile}, 0)$. The upper significance threshold was $\max(99^{\text{th}} \text{ percentile}, 0)$. The greater or less than zero condition was to enable the interpretation of the selection and neutrality statistic results, as zero is the point at which these statistics indicate an excess or deficit of a particular allele frequency category. F_{ST} was also calculated in the same sliding window setup, and pair-wise between the Polynesian populations and the other populations. Negative F_{ST} values were set to 0 because F_{ST} is biologically bounded to [0,1]. Negative values occur when the variation within a population is larger than between populations.

The ancestral allele was used as annotated in the 1000 Genomes Phase 3 VCF file where possible, for use as the out-group population. A new sample identified as ‘Ancestor’ was included in the phased VCF, and for each marker the genotype for this sample was set as the homozygote for either the reference or alternate allele, dependent on the ancestral allele matching, otherwise the Ancestor genotype was set to missing. When the phased VCF files were loaded into PopGenome, populations were identified using panel files which were white-space delimited files with sample id, population, and super population as the columns. The out-group population for Fay and Wu’s H was set to be that of the Ancestor sample so that the ancestral allele would be used in the calculation.

2.1.2.2 SelectionTools

SelectionTools v1.1 (Cadzow *et al.*, 2014) was used to generate the haplotypic selection statistics of iHS, nSL, and XP-EHH. The combined population phased VCF file was split into individual populations. Within each population, markers were filtered for a MAF of > 0.01 and a Hardy-Weinberg Equilibrium (HWE) exact test of $P > 10^{-6}$ (Wigginton *et al.*, 2005). Markers were converted into “ancestral” or “derived” by comparing alleles to the *homo sapiens* ancestral fasta (see Table 2.4) with alleles that matched the ancestor set to 0, and non-missing, non-matching alleles set to 1.

iHS, nSL, and XP-EHH normalisation with frequency bins of 0.05 was done using *norm* as part of selscan v1.1.0b (Szpiech and Hernandez, 2014). Markers with an $|i\text{HS}|$, $|n\text{SL}|$ or $|XP\text{-EHH}|$ value > 2.6 met the threshold for significance, which was equivalent to approximately 1% of the most extreme values.

Significant markers were also clustered into genomic regions using the DBSCAN package v1.1.1 (Hahsler

and Piekenbrock, 2017) in R, where nearby SNPs were assigned the same group identifier. The search radius used was 200 kb and the minimum number of points for a cluster was one. Cluster regions were created by taking both the minimum and maximum position for each group identifier, population, and chromosome.

2.1.3 Disease associated gene lists

In order to create lists of genes that were associated with urate, gout, obesity, type 2 diabetes, metabolic syndrome, and kidney disease, disease trait entries were downloaded from the GWAS catalog³ (MacArthur *et al.*, 2017). The file was a rectangular format, with rows for each study result, and columns such as reference paper, disease trait, association statistics, and mapped genes. Entries were then filtered for $P < 5 \times 10^{-8}$. From the filtered results, the kidney disease gene list was created by filtering the “disease trait” column to select rows with the keywords “kidney” or “renal”. Entries were then removed that had the keywords “transplant”, “carcinoma”, “Type”, “stones”, “gout”, “related”, or “Diabetic kidney disease”. For the gout, urate, obesity, type 2 diabetes, and metabolic syndrome gene lists the following keywords were used on the “disease trait” column to select rows of the P-value filtered data: “metabolic syndrome”, “obesity”, “diabetes”, “urate”, “gout”, “body mass”, and “lipid traits”. This subset was then filtered to remove rows containing these keywords in the “disease trait” column: “child”, “erectile”, “lean”, “autoantibodies”, “gestational”, “cancer”, “psychopharmacol”, “metformin”, “metformin”, “obstructive”, “interaction”, “asthmatics”, “omega”, “pain”, “cataracts”, “time”, “bilirubin”, “chain”, “thyroid”, “zhi”, “Type 1”, or “cystic”. From this filtered data, keywords were used to select rows relating to each trait of interest from the “disease trait” column. The keywords “urate” and “gout” were used for gout and urate, “obesity” and “body mass” used for obesity, “diabetes” used for type 2 diabetes, and “syndrome” used for metabolic syndrome. The keywords “kidney” and “renal” were used for kidney disease, with entries removed that had keywords of “transplant”, “carcinoma”, “type”, “stones”, “gout”, “related”, and “Diabetic kidney disease”.

Additional gene lists were created for malaria, auto-immune and auto-inflammatory diseases, and neurological diseases. The keyword “malaria” was used to filter the disease trait column to create the malaria gene list. The auto-immune and auto-inflammatory gene list was based on diseases that were listed in Table 2 of Zhang *et al.* (2013). The following keywords were filtered for in the disease trait column: “Crohn’s disease”, ‘Celiac disease’, “Ulcerative colitis”, “Inflammatory bowel disease”, “Type 1 diabetes”, “Rheumatoid arthritis”, “Multiple sclerosis”, “Psoriasis”, “Systemic lupus erythematosus”, “Primary biliary cirrhosis”, and “Vitiligo”. Finally, a list of genes associated with neurological diseases was created by filtering the disease trait column for the keywords “Parkinson’s disease” and “Alzheimer’s disease”. All keyword matching was case-insensitive. The code used for the creation of the gene lists is found in Appendix B1. A list of the traits for each category can be found in Table S7 and a list of the genes and what category they are from is in Table S8.

³gwas_catalog_v1.0.1-associations_e89_r2017-06-19.tsv <https://www.ebi.ac.uk/gwas/> accessed 19 June 2017

2.1.4 Principal component analysis

Principal component analysis is a statistical dimension reduction technique that transforms potentially correlated variables into a linear and non-correlated set of variables. In a genetic context PCA is used to reduce variation at many thousands of markers into a handful of components that represent the majority of the variation of the data (Patterson *et al.*, 2006). The components are ordered such that the first principal component (PC) captures the most variation, with each subsequent component capturing less. These components often, but not necessarily, represent population differences and population substructure.

PCA was used to identify the genetic ancestry of the samples from the Genetics of Gout in Aotearoa study to be used in the selection analysis (section 2.2.3). It was also used in the clustering analysis to identify the genetic groupings for the populations (section 4.3.1.1). To calculate the principal components of the genetic data, all populations and chromosomes were combined into a single VCF file with BCFtools v1.3.1, and then the independent markers were identified via Plink v1.9b4.9, using a sliding window to remove markers that had an inter-marker LD $R^2 > 0.2$, with windows of 50 kb and a slide of 5 markers. The first 10 principle components were calculated using smartPCA v13050 from Eigensoft v6.0.1 (Price *et al.*, 2006). The following code was used to accomplish these steps.

```
#combine the chromosomes
bcftools concat \
    -O z \
    -o -o NZ_1KGP_allchr.vcf.gz \
    --threads 10 $(ls NZ_1KGP.chr*gz | sort -n -t'.' -k2)

#find the independent markers
plink1.9b4.9 --vcf NZ_1KGP_allchr.vcf.gz \
    --maf 0.1 \
    --indep-pairwise 50 5 0.2 \
    --out NZ_1KGP_allchr

# create an empty affection file that is required for Plink to use the --make-pheno
# which in turn is required for the creation of the ped file just the way
# SmartPCA wants it
touch cases.txt
plink1.9b4.9 --vcf NZ_1KGP_allchr.vcf.gz \
    --extract NZ_1KGP_allchr.prune.in \
    --recode \
    --out NZ_1KGP_allchr_indep \
    --make-pheno cases.txt '*'

# create the eigenstrat file
echo -e "genotype: NZ_1KGP_allchr_indep.ped\nsnpname: \\"
```

```

NZ_1KGP_allchr_indep.map\nindivname: \
NZ_1KGP_allchr_indep.ped\noutputformat: \
EIGENSTRAT\ngenotypeoutname: \
NZ_1KGP_allchr_indep.eigenstratgeno\nsnputname: \
NZ_1KGP_allchr_indep.snp\nindivoutname: \
NZ_1KGP_allchr_indep.ind\nfamilynames: \
NO" > par.PED.EIGENSTRAT

# calculate the principle components
convertf -p par.PED.EIGENSTRAT > eigen.log
smartpca.perl \
-i NZ_1KGP_allchr_indep.eigenstratgeno \
-a NZ_1KGP_allchr_indep.snp \
-b NZ_1KGP_allchr_indep.ind \
-o NZ_1KGP_allchr_indep_eigen.pca \
-p NZ_1KGP_allchr_indep_eigen \
-e NZ_1KGP_allchr_indep_eigen.eval \
-l NZ_1KGP_allchr_indep_eigen.log \
-m 0

```

2.1.5 Admixture analysis

Admixture arises when two populations that were previously separate, begin to interbreed, changing the allele frequencies in the new population (Pritchard *et al.*, 2000). Genetic admixture analysis estimates the proportions and the variant frequencies of the ancestral populations that contributed to the admixture. This is often done to account for population structure in GWAS studies. Admixture analysis was performed to estimate the number of ancestral populations that contributed to the populations used in the selection analysis. It was also used because the Polynesian populations have varying degrees of admixture (Wollstein *et al.*, 2010).

Admixture analysis was performed using ADMIXTURE v1.3.0 (Alexander and Novembre, 2009). VCF files for all autosomes were concatenated and the independent markers were selected by a moving window of 50 kbp sliding by 10 markers and removing markers with a marker LD $R^2 > 0.1$. This was done using Plink v1.9b4.9. Following this, cross validation was performed to find the best value of K (number of ancestral populations), for values of K from 1 to 15. The default of 5-fold cross-validation was used, whereby the data were partitioned into five groups, with four used to train the model, and the fifth to test the model. This was repeated five times, using a different combination of the partitions each iteration. The following code was used to calculate the cross-validation errors.

```

# Combine the chromosomes
bcftools concat \
-0 z \

```

```

-o NZ_1KGP_allchr.vcf.gz \
--threads 10 $(ls NZ_1KGP.chr*gz | sort -n -t'.' -k2)

# find the independent snps
plink1.9b4.9 \
--vcf NZ_1KGP_allchr.vcf.gz \
--indep-pairwise 50 10 0.1 \
--out NZ_1KGP_allchr_admix

# extract the independent snps
plink1.9b4.9 \
--vcf NZ_1KGP_allchr.vcf.gz \
--make-bed --extract NZ_1KGP_allchr_admix.prune.in \
--out NZ_1KGP_allchr_admix

# do the cross-validation for the admixture components for K 1-15
for K in $(seq 1 15)
do
    admixture -s 123456 --cv NZ_1KGP_allchr_admix.bed $K -j20 | tee log${K}.out
done

grep CV log* | cut -d':' -f1,3 | tr -d ':' | sed 's/log\\|\\.\out//g' > CV_error.txt

```

2.1.6 Heritability analysis

Heritability analysis is a method for determining what proportion of phenotypic variation can be attributed to environmental effects (non-heritable) and genetic effects (heritable) (Visscher *et al.*, 2008). There are two measures of heritability, the first is known as broad-sense heritability (H^2) and is the ratio of the total genetic variance to the total phenotypic variance. The second measure is narrow-sense heritability (h^2) and is the ratio of the genetic variance attributed to an additive genetic model to the total genetic variance.

The Genome-wide Complex Trait Analysis (GCTA) v1.26.0 software (Yang *et al.*, 2011a) was used to calculate the proportion of genetic heritability explained for gout. First the UK Biobank genetic data (section 2.2.4) were subsetted to only contain gout cases and controls and the genome partitioned into chromosomes. A genetic relationship matrix (GRM) was created for each chromosome. The genetic variance explained was calculated for each chromosome using restricted maximum likelihood analysis and a general population prevalence for gout of 2%. The following code was used to calculate the GRM and partition the genetic heritability by chromosome.

```

# load fam into R
# sample(10000, fam[AFF == controls]) -> controls.txt
# cat condition.fam /
# awk '{if($6 == 2){print $1"\t"$2}}' / cat - controls.txt > ids_keep.txt

# Subset samples
for i in $(seq 1 22)
do
    plink2 --bed $ukbio_path/chr${i}impv1.bed \
        --bim $ubkbio_path/chr${i}impv1.bim_1kg_marker \
        --fam $ukbio_path/chrallimpv1.fam_allgout_allcontrols \
        --hwe 0.000001 \
        --maf 0.01 \
        --keep ids_keep.txt \
        --make-bed \
        --out gout_chr${i} \
done

# Calculate the genetic relationship matrix by chromosome
for i in $(seq 1 22)
do
    gcta1.26.0 \
        --bfile gout_chr${i} \
        --chr ${i} \
        --make-grm-bin \
        --out ukbio_10ksample_grm_chr${i} \
        --thread-num 16 > grm_${i}.log \
done

# Calculate the genetic variance explained using a
# general population gout prevalence of 2%
for i in $(seq 1 22)
do
    gcta1.26.0 \
        --grm ukbio_10ksample_grm_chr${i} \
        --pheno phenos.txt \
        --prevalence 0.02 \
        --out var_chr${i} \
        --reml \

```

```
--chr ${i} \
--thread-num 16 \
done
```

2.1.7 Pathway Analysis

Gene set analysis was performed by inputting gene lists into Enrichr⁴ (Chen *et al.*, 2013; Kuleshov *et al.*, 2016). The pathway enrichment results were based on the KEGG 2016 table, which is a long-standing database, curated by a small group of experts (Kanehisa *et al.*, 2017). A significance threshold for a pathway was set at $P < 0.05$ after Benjamini-Hochberg adjustment (Benjamini and Hochberg, 1995) for multiple hypothesis testing, as provided by Enrichr.

2.2 Datasets

The following sections describe the genomic datasets used in this thesis.

2.2.1 1000 Genomes Project Phase 3

The 1000 Genomes Project was an international consortium that was established in 2007 to provide a comprehensive record of human genetic variation (Siva, 2008). The project consisted of three main data phases. A pilot phase that whole-genome sequenced 179 individuals from four populations at low coverage (2-4x), along with high coverage sequencing for two trios (mother, father, and child), and exon targeted sequencing for 697 individuals from seven populations (1000 Genomes Project Consortium, 2010). The second main data phase provided sequencing data for 1092 individuals from 14 populations. This sequencing data set was a combination of low coverage whole-genome and exon sequencing (1000 Genomes Project Consortium, 2012). The third main phase (Phase 3) was a dataset consisting of low coverage whole-genome sequencing, deep exome sequencing, and dense SNP array genotyping for 2504 individuals from 26 populations (1000 Genomes Project Consortium, 2015). The 1000 Genomes data set used in this thesis was the Phase 3 release⁵.

2.2.2 Genetics of Gout in Aotearoa study

The genetics of Gout in Aotearoa study is a case-control cohort for gout, with recruitment mainly from the Auckland, Wellington, Christchurch and Dunedin regions of New Zealand. Participants were asked to fill in a questionnaire regarding demographic information, clinical information, and gout-relevant food consumption at the time of recruitment. Gout cases fulfilled the American College of Rheumatology (ACR) criteria (Wallace *et al.*, 1977). Controls self-reported no history of gout at the time of recruitment. All individuals gave written informed consent, and ethical approval was obtained

⁴<http://amp.pharm.mssm.edu/Enrichr/>

⁵<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/> accessed 20 March 2017

Table 2.1: Description of Populations used in the selection analysis (n = 3004).

Population	Description	n	Genotype Platform
African Super Population (AFR) n = 661			
ACB	African Caribbeans in Barbados	96	Illumina HiSeq
ASW	Americans of African Ancestry in SW USA	61	Illumina HiSeq
ESN	Esan in Nigeria	99	Illumina HiSeq
GWD	Gambian in Western Divisions in the Gambia	113	Illumina HiSeq
LWK	Luhya in Webuye Kenya	99	Illumina HiSeq
MSL	Mende in Sierra Leone	85	Illumina HiSeq
YRI	Yoruba in Ibadan Nigeria	108	Illumina HiSeq
Admixed American Super Population (AMR) n = 347			
CLM	Colombians from Medellin Colombia	94	Illumina HiSeq
MXL	Mexican Ancestry from Los Angeles USA	64	Illumina HiSeq
PEL	Peruvians from Lima Peru	85	Illumina HiSeq
PUR	Puerto Ricans from Puerto Rico	104	Illumina HiSeq
East Asian Super Population (EAS) n = 504			
CDX	Chinese Dai in Xishuangbanna China	93	Illumina HiSeq
CHB	Han Chinese in Beijing China	103	Illumina HiSeq
CHS	Southern Han Chinese	105	Illumina HiSeq
JPT	Japanese in Tokyo Japan	104	Illumina HiSeq
KHV	Kinh in Ho Chi Minh City Vietnam	99	Illumina HiSeq
European Super Population (EUR) n = 603			
CEU	Utah Residents (CEPH) with Northern and Western Ancestry	99	Illumina HiSeq
FIN	Finnish in Finland	99	Illumina HiSeq
GBR	British in England and Scotland	91	Illumina HiSeq
IBS	Iberian Population in Spain	107	Illumina HiSeq
TSI	Toscani in Italia	107	Illumina HiSeq
NZC*	Europeans in New Zealand	100	CoreExome_v24
Polynesian Super Population (POL)* n = 400			
CIM*	Cook Island Maori in New Zealand	100	CoreExome_v24
NZM*	Maori in New Zealand	100	CoreExome_v24
SAM*	Samoans in New Zealand	100	CoreExome_v24
TON*	Tongans in New Zealand	100	CoreExome_v24
South Asian Super Population (SAS) n = 489			
BEB	Bengali from Bangladesh	86	Illumina HiSeq
GIH	Gujarati Indian from Houston Texas	103	Illumina HiSeq
ITU	Indian Telugu from the UK	102	Illumina HiSeq
PJL	Punjabi from Lahore Pakistan	96	Illumina HiSeq
STU	Sri Lankan Tamil from the UK	102	Illumina HiSeq

* NZC, NZM, TON, SAM, and CIM are from the 'Genetics of Aotearoa' study. All others are from the 1000 Genomes Project.

from the New Zealand multi-region ethics committee (MEC/105/10/130). Individuals from this study were genotyped on Illumina CoreExome v24 SNP arrays using DNA extracted from blood samples provided at the time of recruitment. Table 2.2 provides the overall clinical characteristics for this cohort.

A subset of the individuals from this study were used to form four populations of Polynesian ancestry, two for East Polynesia (Cook Island Māori in New Zealand (CIM) and Māori in New Zealand (NZM)), and two for West Polynesia (Samoans in New Zealand (SAM) and Tongans in New Zealand (TON)), and one population of European ancestry (Europeans in New Zealand (NZC)). These sub-populations formed part of the selection analysis dataset described in section 2.2.3).

2.2.2.1 CoreExome

The Infinium CoreExome-24 bead-chip is a genotyping platform available from Illumina and is comprised of a core set of 551,839 markers. A subset of the individuals in the Genetics of Gout in Aotearoa study were genotyped on this platform at the University of Queensland (Centre for Clinical Genomics). Genotype quality control was performed by Dr Tanya Major (Merriman Lab) following the protocol from Guo *et al.* (2014) and the Illumina GenomeStudio best practice guidelines⁶. Illumina's GenomeStudio v2011.1 genotyping module v1.9.4 was used for the initial calling of genotypes. Samples were exported from GenomeStudio that had a call rate > 98%, markers were removed if the call-rate was < 95%. Individuals who had not reported their sex were assigned their genetic sex where possible. Individuals were removed where genetic and reported sex did not match. The cohort was checked for genotype consistency between duplicated markers, with duplicates subsequently removed.

Markers were subsetted to only include biallelic SNPs, with a final marker number of 305214 and density of 9142 bp/marker. Relatedness between individuals was assessed through inheritance by state using Plink v1.9b3.32 and a pedigree of families created by Dr Tanya Major (Merriman lab) to identify family groups. Duplicates and first-degree relations were excluded and single individuals were randomly selected from family groups identified through identity by state analysis, which determined the proportions of individual genomes that were identical in a pair-wise manner. In order to account for population specific genetic differences, self-report of grandparent ethnicity was used to infer the genetic ancestral populations from clusters of individuals after plotting different PCs from the PCA on the genetic markers (section 2.1.4). Individuals were removed from further analysis where there was disagreement between self-reported grandparent ethnicity and the inferred genetic ancestry. Hardy-Weinberg equilibrium of markers was checked for European, East Polynesian, and West Polynesian populations using a HWE exact test (Wigginton *et al.*, 2005) in Plink v1.9b3.32, and variants were removed if they had a Bonferroni multiple testing corrected P < 0.05.

⁶https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf accessed 18 January 2016

Table 2.2: Clinical information for samples that were genotyped on the CoreExome platform by genetic ancestry.

	European		Eastern Polynesian		Western Polynesian	
	Control	Gout	Control	Gout	Control	Gout
Fatty liver, n (%)	10 (0.88)	27 (2.45)	2 (0.23)	4 (0.66)	1 (0.26)	6 (1.34)
Kidney disease, n (%)	32 (2.82)	240 (21.82)	26 (3.01)	133 (21.8)	11 (2.85)	83 (18.53)
Mean Diastolic BP (SD)	130.3 (17.4)	139.4 (45.7)	130.6 (19.3)	135.9 (20.4)	130 (18.4)	132.6 (17.6)
Diabetes, n (%)	54 (4.77)	159 (14.45)	100 (11.59)	159 (26.07)	53 (13.73)	99 (22.1)
n	1133	1100	863	610	386	448
Heart disease, n (%)	114 (10.06)	387 (35.18)	96 (11.12)	225 (36.89)	22 (5.7)	90 (20.09)
Mean age, years (SD)	48.79 (18.36)	63.54 (13.05)	44.17 (15.46)	56.12 (12.65)	38.54 (14.45)	48.82 (12.58)
Mean BMI (SD)	27.16 (5.47)	30.33 (5.4)	31.9 (7.84)	35.42 (8.17)	34 (6.7)	36.64 (7.32)
Mean Systolic BP (SD)	76.9 (10.9)	79.3 (11.5)	81.1 (13.8)	83.9 (13.6)	79.8 (15.2)	82.3 (11.7)
Sex, n male (% male)	606 (53.49)	927 (84.27)	333 (38.59)	466 (76.39)	207 (53.63)	406 (90.62)

BP = blood pressure (mmHg). BMI = body mass index (kg/m^2).

2.2.3 Selection dataset

2.2.3.1 Sample selection

A subset of the individuals were chosen to create representative sample Polynesian populations of similar size to the other populations in the 1KGP to be used in a comparison for the selection analyses. Principal components were calculated for all individuals genotyped on the CoreExome SNP array using SmartPCA (EIGENSOFT v6.0.1, see section 2.1.4) using 2858 ancestry informative markers (Guo *et al.* (2014) supplementary material). The first 10 eigenvectors were outputted, with no outlier removal or population size limit. Individuals were removed who did not match between self-reported ethnicity of grandparents and their genetic ancestry, or if they had self-reported all grandparent ethnicities as unknown.

Principal component 2 was identified as providing the best separation of European ancestry from Polynesian ancestry, and PC 4 had the best separation of East Polynesian from West Polynesian ancestry. Individuals were then filtered for European or Polynesian ancestry. Individuals reporting four grandparents of European, New Zealand Māori, Cook Island Māori, Samoan, and Tongan ancestry were used to define the threshold values, as they were most likely to represent the ancestral populations. The mean and SD for PC 2 and PC 4 were calculated for the individuals who had self-reported four grandparents of the same ethnicity. Thresholds were then applied using the four grandparent population mean ± 2 SD for PC 2 and PC 4 for each corresponding self-report population group. Samples were filtered so that only those that lay within the thresholds remained. Population samples of 100 individuals were created by using genetic ancestry for the following populations: Māori in New Zealand, Cook Island Māori in New Zealand, Samoans in New Zealand, Tongans in New Zealand, and Europeans in New Zealand, by randomly sampling from within each ancestry group (Figure 2.1). To create a population sample that resembled the general population for gout prevalence, individuals were prioritised based on gout affection to reach population specific prevalences based on Winnard *et al.* (2013) of 2.3% for NZC, 7.7% for NZM, and 8.6% for CIM, SAM, and TON. Final sample population prevalence of 2.0% (NZC), 7.0% (NZM), 45.0% (CIM), 12.0% (SAM), and 54.0% (TON) were obtained for gout (Table 2.3).

Table 2.3: Clinical information for New Zealand Populations used in the selection analysis

Population	n	Mean Age yrs (SD)	Sex (% Male)	Mean BMI kg/m ² (SD)	Mean Waist cm (SD)	Diabetes (%)	Gout (%)	Kidney Disease (%)	Heart Problems (%)
CIM	100	53.77 (16.16)	53	34.64 (7.44)	108.76 (15.49)	31	45	10	24
NZC	100	54.76 (16.38)	71	28.28 (6.29)	95.38 (13.65)	6	2	7	14
NZM	100	48.17 (14.19)	43	33.86 (8.22)	106.46 (16.07)	16	7	3	18
SAM	100	40.71 (13.23)	66	34.95 (6.71)	106.78 (12.75)	13	12	4	2
TON	100	42.43 (15.03)	84	36.02 (6.81)	112.57 (13.77)	17	54	11	8

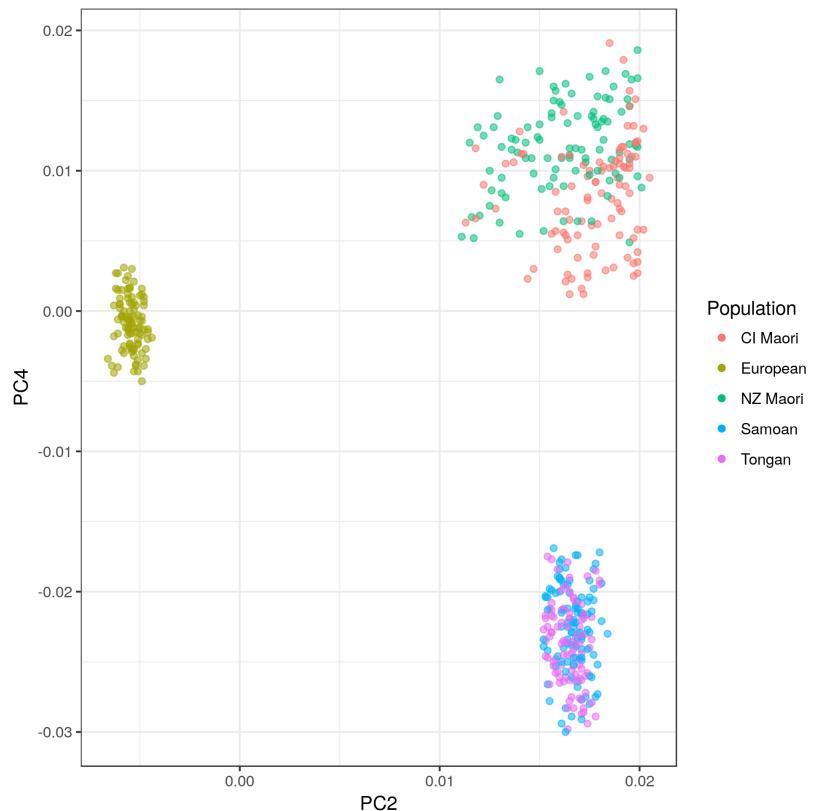


Figure 2.1: Principal components 2 and 4 for individuals used in the selection analysis.

2.2.3.2 Filter individuals and markers

To create a data set from which to calculate selection statistics, the subset of the Genetics of Aotearoa individuals genotyped on the Illumina CoreExome chip were combined with all of the individuals in the 1KGP phase 3 release. This was done in order to have a common set of markers between all populations, making the populations comparable. Before the data sets could be combined there were a series of steps that were performed. The first step (code following) was to keep only the individuals who had been selected through PCA and had a genotyping rate (percentage of all markers with genotypes) of at least 95%. Due to software requirements, markers were filtered to remove INDELs so that only SNPs remained.

```
parallel '
plink2 \
--bfile src_data/QC1_7-plus_correctAff \
--recode vcf \
--out QC1_7-plus_correctAff.chr{} \
--chr {} \
--remove src_data/QC1_7-BlanketExclusions.txt \
--keep coreExome_selection_keep_ids.txt \
--allow-no-sex \
--snps-only no-DI \
--geno 0.95
' :: $(seq 1 22)
```

The second step was to ensure that all the CoreExome markers were normalised against the hs37d5 human reference. This was to make sure the reference (REF) and alternative (ALT) alleles would match with the 1KGP phase 3 for a successful merge after phasing and subject selection. SHAPEIT2 was then used to find SNP alleles that disagreed with the 1KGP haplotype reference and phase the CoreExome markers. This step (code following) also reduced markers to biallelic positions and removed duplicate positions.

```
parallel '
bcftools norm \
-N \
--rm-dup any \
--check-ref s \
-f hs37d5.fa \
-O v QC1_7-plus_correctAff.chr{}.vcf | \
bcftools view \
-m 2 \
-M 2 \
-O z \
-o QC1_7-plus_correctAff_norm.chr{}.vcf.gz \
'
```

```
' :::: $(seq 1 22)
```

After normalisation, the VCF files were phased using the 1000 Genomes Project phase 3 reference haplotypes as described in section 2.1.1.

2.2.3.3 Merge genotypes

In order to efficiently merge the data sets, the intersection of the markers from the 1KGP dataset and the CoreExome dataset was found. This was done by extracting the marker positions and alleles from the CoreExome data and then matching these to the marker positions extracted from the 1KGP phase 3. The 1KGP phase 3 markers were also filtered to remove non-biallelic SNPs. The following code demonstrates these steps.

```
# extract markers from core exome
parallel 'zgrep \
-v "^#" QC1_7-plus_correctAff_norm.chr{}.phased.vcf.gz |\
cut -f1,2,4,5 > coreExome_chr{}_biallelic_markers.txt
' :::: $(seq 1 22)

# extract 1000 Genomes markers
parallel 'bcftools view \
-O v \
-m 2 \
-M 2 \
-v snps \
-o - \
ALL.chr{}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |\
grep -v "^#" |\
cut -f 1,2,4,5 > 1kgp_chr{}_biallelic_markers.txt
' :::: $(seq 1 22)
```

An R script (below) was created in order to merge the two marker lists based on chromosome position and the reference and alternate alleles.

```
# create list of markers present in CoreExome data
ce_markers <- data.frame()
tmp_list <- list()
for(i in 1:22){
  tmp_list[[i]] <- read.table(file = paste0(
    'NZ_coreExome/coreExome_chr',
    i, '_biallelic_markers.txt'), header=FALSE)
```

```

}

ce_markers <- do.call(rbind, tmp_list)

# create list of markers present in 1KGP data
kg_markers <- data.frame()
tmp_list2 <- list()
for(i in 1:22){
  tmp_list2[[i]] <- read.table(file= paste0(
    '1kgp_chr',i,'_biallelic_markers.txt'),
    header=FALSE)
  # find the markers in common between CoreExome and 1KGP
  tmp_list2[[i]] <- tmp_list2[[i]][ tmp_list2[[i]][,2] %in% tmp_list[[i]][,2],]
}
kg_markers <- do.call(rbind, tmp_list2)
ce_kg <- merge(ce_markers, kg_markers, by = c("V1","V2","V3","V4"))

# write out the common markers
write.table(file = 'ce_1kg_matched_markers.txt', ce_kg[,c(1,2)],
            row.names = FALSE, col.names=FALSE, quote=FALSE)

```

Once the combined marker list was created, both the CoreExome and 1KGP data sets had markers filtered to match this consensus set. The code for marker filtering of the VCF files and subsequent merge is shown below.

```

# filter 1000 Genomes markers
parallel '
bcftools view
-R ../ce_1kg_matched_markers.txt \
-O z \
-m 2 \
-M 2 \
-v snps \
-o ALL.chr{}.phase3_shapeit2_mvncall_integrated_v5a.20130502.matched.vcf.gz \
ALL.chr{}.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
' :: $(seq 1 22)

# filter coreExome markers
parallel '
bcftools view \
-R ../ce_1kg_matched_markers.txt \
-O z \
-m 2 \

```

```

-M 2 \
-v snps \
-o NZ_coreExome.chr{}.norm.phased.matched.vcf.gz \
QC1_7-plus_correctAff_norm.chr{}.phased.vcf.gz
' :::: $(seq 1 22)

# merge 1kgp and nz coreexome
parallel
'bcftools merge \
-O z \
-o NZ_1KGP.chr{}.phased.vcf.gz \
ALL.chr{}.phase3_shapeit2_mvncall_integrated_v5a.20130502.matched.vcf.gz \
NZ_coreExome/NZ_coreExome.chr{}.norm.phased.matched.vcf.gz
' :::: $(seq 1 22)

```

Sample identifiers were then updated to match the panel file that described the population and super population an individual belonged to. The following code applies this step to each merged VCF file.

```

# Combine FID and IID of CoreExome samples so they match the panel file
for i in $(seq 1 22)
do
zcat NZ_1KGP.chr$i.phased.vcf.gz |\
head -1000 |\
grep '^#CHROM' |\
cut -f10- |\
tr '\t' '\n' |\
awk -F "_" '{if(NF ==1 ) {print $1 "\t" $1}else{print $1 "_"$2"\t" $2}}' |\
bcftools reheader \
-s /dev/stdin/ \
-o NZ_1KGP.chr$i.phased.sample_updated.vcf.gz \
NZ_1KGP.chr$i.phased.vcf.gz
done

```

2.2.4 UK Biobank

The UK Biobank is a collection of 500,000 individuals, mostly of European ancestry from around the United Kingdom, and aged between 40 and 69. It consists of genetic, health, and lifestyle information. The interim release dataset (approval number 12611) was downloaded in November 2015 and contained genotypes for 152,249 individuals. Gout affection was determined by self-report, or self-reported use of urate lowering therapy, or an International Classification of Diseases, Tenth Revision (ICD-10) code for gout (M10, including sub-codes). The dataset was also filtered to individuals with a self-reported ethnic background of British, Irish, or ‘any other white background’. Individuals were removed who

had a mismatch between self-reported sex and genetic sex, or whose samples failed genotype quality control. The remaining individuals were eligible for inclusion in the GWAS performed in section 5.2.1.

2.2.5 Reference datasets

Table 2.4 provides the name of the dataset, the date it was accessed, and the URL from which the dataset was downloaded. These datasets were incorporated into many of the analysis steps.

Table 2.4: Reference Datasets

Name	Date	URL
1000 Genomes Project Phase 3 release 5	20 Mar 2017	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
1000 Genomes Phase 3 phased haplotypes	14 Dec 2014	https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
Human Genome Reference FASTA	5 Aug 2014	ftp://ftp.1000genomes.ebi.ac.uk:21/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
Human Ancestral Allele FASTA	6 Aug 2014	ftp://ftp.ensembl.org/pub/release-66/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e66.tar.bz

Chapter 3

Positive Selection in Polynesian Populations

In this chapter I investigate regions of the genome that exhibit signatures of selection in Polynesian populations, with a particular focus on loci associated with serum urate levels and metabolic conditions.

3.1 Introduction

At the turn of the millennium the draft reference human genome was released (Lander *et al.*, 2001). Following from this, there were a number of large projects to characterise human variation and diversity. The first project was the HapMap project (The International Hapmap Consortium, 2005) and was based on single nucleotide polymorphism (SNP) array data. This was followed by the 1000 Genomes Project (1000 Genomes Project Consortium, 2010, 2012, 2015) which was the first large project to be based on whole-genome (re)sequence data. Alongside these variation projects were a few projects that aimed to focus on the genomic diversity of humans, such as the Human Genome Diversity Project (Cann *et al.*, 2002; Rosenberg *et al.*, 2002), and later the Simons Genome Diversity Project (Mallick *et al.*, 2016).

Over the last decade there has been a steady flow of studies on genome-wide selection (Sabeti *et al.*, 2006; Voight *et al.*, 2006; Tang *et al.*, 2007; Coop *et al.*, 2009; Pickrell *et al.*, 2009; Grossman *et al.*, 2010). The main population samples for this have been part of the 1000 Genomes Project, which sampled over 2500 individuals for full genome (re)sequencing from five ‘super’ population groups (1000 Genomes Project Consortium, 2015). The five super populations cover each of the geographical regions of Africa, Europe, East Asia, South Asia, and the Americas. This dataset, while capturing a large representation of the human population variation, does not have representation for the populations of Polynesia.

3.1.1 Positive selection

3.1.2 Selection in Polynesian populations

Projects to capture the genetic diversity of humans have sampled populations worldwide, but outside the populations sampled by the 1000 Genomes Project, the population sample sizes have been small (Cann *et al.*, 2002; Rosenberg *et al.*, 2002; Mallick *et al.*, 2016). Past studies that have included Oceanic populations have had minimal representation of Polynesian populations, for example, Kimura *et al.* (2008) only had 24 Tongan individuals. The latest project for human diversity, the Simons Human Diversity Project, did include New Zealand Māori but they were only represented by a single individual and did not feature in many of the analyses (Mallick *et al.*, 2016).

More recently there have been studies that have looked at selection in Polynesian populations but with a focus on a few candidate loci, not at a genome-wide scale (Myles *et al.*, 2011; Cadzow *et al.*, 2016; Minster *et al.*, 2016). The main feature of these studies has been the investigation of “thrifty genes”.

3.1.2.1 Thrifty gene hypothesis

The “Thrifty genotype hypothesis” was originally hypothesised by Neel (1962), whereby it was posed there must be a genetic advantage to a disease, given what should be a strong selection pressure against it, for it to continue to have a high prevalence - referring to diabetes. In the Polynesian context, this hypothesis has been invoked in an attempt to explain the higher prevalence of metabolic related diseases such as type 2 diabetes and obesity (Myles *et al.*, 2011; Minster *et al.*, 2016). However, there are several refutations of the thrifty genotype for Polynesian populations as fitting a particular narrative (Gosling *et al.*, 2014, 2015; Cadzow *et al.*, 2016). An alternative to the ‘thrifty-gene’ hypothesis is that the link between metabolic diseases and the innate immune system, susceptibility of populations to metabolic disease might be due to selective pressure from infectious disease, such as malaria (Gosling *et al.*, 2015). Uricase, the enzyme for metabolising urate, was lost through multiple mutations - a reduction in promoter activity followed by two further loss of function mutations and occurred before the divergence of humans from the great apes (Kratzer *et al.*, 2014). The evolutionary history of uricase however has also been described as “the original thrifty-gene” (Kratzer *et al.*, 2014).

3.1.3 Selection in urate and gout

The benefits of urate include its ability to increase blood pressure and it is thought to have helped early hominins to stand upright (Watanabe *et al.*, 2002) and increase blood flow to the brain. Serum urate is also a potent anti-oxidant which accounts for >50% of the anti-oxidant activity of the blood (Glantzounis *et al.*, 2005; Parmar, 2009). Mono-sodium urate crystals that form in conditions of elevated urate are a potent adjuvant for the immune system and enhances the immune response (Ames *et al.*, 1981; Opitz *et al.*, 2009). Uric acid has also been identified as being an important part of the malarial response, contributing to cytokine secretion and response of dendritic and T-cells (Gallego-Delgado *et al.*, 2014).

An increased level of urate in the blood is called hyperuricaemia and is causal of gout (Choi *et al.*, 2005). The prevalence of hyperuricaemia in Polynesian populations is higher than many other populations (Gosling *et al.*, 2014), and co-morbid with it are higher prevalences of the co-morbidities of gout, diabetes, obesity, dyslipidaemia, kidney disease, and cardio vascular disease (CVD) (Winnard *et al.*, 2013). In New Zealand, gout prevalence is ~3.1%, however, prevalence in Māori and Pacific peoples is 6-8% with 40% of individuals with gout also having diabetes or CVD (Winnard *et al.*, 2013). Gout is an immune response to the monosodium urate crystals that are formed from the precipitation of urate from the blood, with hyperuricaemia being a prerequisite for the formation of these crystals (Merriman and Dalbeth, 2011).

Hyperuricaemia has two main components: overproduction of urate, and under-excretion. Overproduction is influenced by diet, such as high purine or fructose content, or increased cell turnover as observed in cancer, or under genetic control. Under-excretion is the main cause of hyperuricaemia, caused by either increased reabsorption or decreased secretion. The kidneys are the organ where the majority of secretion and excretion of urate is performed, with about 90% of urate being reabsorbed (Kutzing and Firestein, 2008). In the proximal tubule, on the apical membrane, the urate transporters of URAT1 (*SLC22A12*), OAT4 (*SLC22A11*), and GLUT9 (*SLC2A9*) reabsorb urate, whereas ABCC4, NPT4 (*SLC17A3*), and NPT1 (*SLC17A1*) are responsible for secretion (Figure 3.1) (Mandal and Mount, 2015). On the basolateral membrane, OAT1-3 (*SLC22A6-8*), increase intracellular uric acid, but GLUT9 moves uric acid from intracellular to extracellular (Mandal and Mount, 2015). The ATP powered transporter ABCG2 is responsible for excretion of urate in the gut and there are SNPs that lead to a reduction of function in ABCG2 in all populations (Figure 3.1) (Phipps-Green *et al.*, 2010; Cleophas *et al.*, 2017). Together the genes *SLC2A9* and *ABCG2* account for 3.4% of the genetic variability of urate (Köttgen *et al.*, 2013).

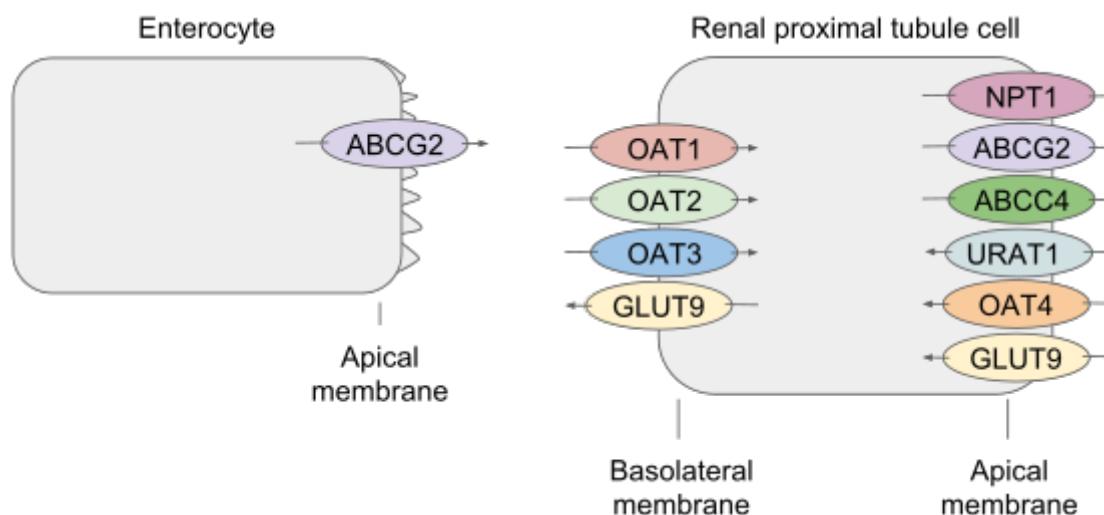


Figure 3.1: Diagram representing the urate transporters in the gut (enterocyte) and kidney (renal proximal tubule cell). Direction of urate movement is indicated by arrows. Adapted from Major *et al.* (2018).

From previous studies there is prior evidence of selection at urate associated loci (Grossman *et al.*, 2013;

Zhang *et al.*, 2013; Ramos, 2017). In the review from Ramos (2017) on natural selection in rheumatic disease, one of the main loci for gout, *SLC2A9*, in the Yoruba in Ibadan Nigeria (YRI) population was reported as having been selected. The region of *BCAS3* and *NACA2* also had evidence, originally found by Grossman *et al.* (2013) in the Asian population from the HapMap project data. Zhang *et al.* (2013) had evidence for the urate-associated loci of *ATXN2* (rs653178) and *RREB1* (rs675209) having undergone selection when using a F_{ST} -based method. The co-morbidities of hyperuricaemia also show signs of selection, with loci associated with type 2 diabetes having been reported as having undergone selection (Voight *et al.*, 2006; Pickrell *et al.*, 2009; Grossman *et al.*, 2010, 2013). Most recently a Samoan specific variant in *CREBRF* that is protective for type 2 diabetes but causal for obesity has also been identified as having evidence of selection in the Samoan population (Minster *et al.*, 2016).

3.1.4 Objectives

The objectives of this chapter are to:

- Investigate regions that are under selection in Polynesian populations.
- Investigate regions previously associated with urate, type 2 diabetes, obesity, kidney disease, and metabolic syndrome for signatures of selection.

3.2 Methods

The choice of individuals, and preliminary data preparation for the selection data set is detailed in section 2.2.3. It consisted of combining 500 individuals (100 per population for each of Māori in New Zealand (NZM), Cook Island Māori in New Zealand (CIM), Samoans in New Zealand (SAM), Tongans in New Zealand (TON), and Europeans in New Zealand (NZC)) from the Genetics of Gout in Aotearoa study, who had been genotyped on the Illumina CoreExome v24 SNP array with 2504 individuals from the 1000 Genomes Project.

3.2.1 Calculation of selection and neutrality statistics

The R package PopGenome v2.2.3 (Pfeifer *et al.*, 2014) was used to calculate the site-frequency spectrum statistics Tajima's D , Fay and Wu's H , Fu and Li's F , Zeng's E , and F_{ST} over non-overlapping windows of 10 kb (see section 2.1.2.1 for further details). Except for F_{ST} , the empirical distribution for each population was used to create an upper and a lower significance threshold. The lower threshold was the 1st percentile and less than zero. The upper significance threshold was the 99th percentile and greater than zero. Contiguous regions of score depression or elevation were created using a rolling window approach, whereby a condition of 15 or more windows out of a minimum of 20 consecutive 10 kb windows were required to have met a significance threshold in the same tail of the distribution. These represent extended regions of the genome and can be indicative of a selective sweep (Carlson *et al.*, 2005).

The integrated haplotype homozygosity score (iHS), number of segregating sites by length (nSL), and cross-population extended haplotype homozygosity (XP-EHH) were calculated and normalised using selscan v1.1.0b (Szpiech and Hernandez, 2014) as part of the selectionTools 1.1 (Cadzow *et al.*, 2014), pipeline as detailed in section 2.1.2.2. The default setting was used which was a maximum gap extension of 200 kb. Significance was defined as $|value| > 2.6$ which was equivalent to the most extreme ~1% of values after normalisation using frequency bins of 0.05.

3.2.2 Pathway enrichment analysis

Pathway enrichment analysis was performed by taking genes that intersected the 1st percentile windows for the site frequency spectrum-based statistics that also had a value < 0, or genes that had a significant marker for iHS or nSL and inputting them into Enrichr¹ (Chen *et al.*, 2013; Kuleshov *et al.*, 2016) and exporting the KEGG 2016 pathway table.

3.3 Results

Results will be split into two themes, the first will be regions of the genome that show evidence of selection in the Polynesian populations. This theme will initially cover the results of the Polynesian populations as a combined group, and then results that were specific to individual populations. The second theme will look specifically at selection of urate and associated loci in Polynesian populations. For reporting, there will be a focus on the 1st percentile/lower tail for the intra-population statistics as this is the distribution end that corresponds with positive selection.

3.3.1 False discovery rate of windowed statistics

In order to control for false positives, the empirical false discovery rate (FDR) was calculated for Tajima's *D*, Fu and Li's *F*, Fay and Wu's *H*, and Zeng's *E*. To calculate the FDR, 5000 permutations of the genotype data were performed, each randomly shuffling the genotypes of markers by marker position (i.e. shuffling rows only). This process re-defined the marker composition of each window, requiring new selection statistics to be calculated for each window. This procedure therefore, generated a new distribution for values of each selection statistic (per permutation) under the null hypothesis of no association between marker genotype and genomic position. A similar method was used by Teshima and colleagues (2006). For each permutation, for each statistic, the values for quantile 'bins' were calculated for both tails of the distribution. The 'bins' ranged from 0.01 to 0.1% in 0.01 increments, 0.1 to 1% in 0.1% increments, 1 to 5% in 1% increments, and 10%. The median and 95% confidence intervals were calculated for the values of the quantile 'bins' across the permutations (i.e., the median, 2.5th and 97.5th percentiles for the values of the 1st percentile bin from all permutations). The enrichment for each quantile bin was then calculated by comparing the number of windows that were observed in the data compared to the number expected from the permutations. FDR was then

¹<http://amp.pharm.mssm.edu/Enrichr/> accessed 22 November 2017

Table 3.1: False discovery rate for windowed frequency spectrum based intra-population statistics by population.

Population	Lower tail		Upper tail	
	Percentile	FDR (95% CI)	Percentile	FDR (95% CI)
Fay and Wu's H				
CIM	0.06	0.059 (0.051, 0.071)	99.4	0.571 (0.532, 0.609)
NZM	0.07	0.066 (0.057, 0.076)	99.4	0.585 (0.538, 0.625)
SAM	0.06	0.056 (0.049, 0.065)	99.4	0.570 (0.524, 0.615)
TON	0.05	0.050 (0.043, 0.059)	99.4	0.589 (0.546, 0.626)
Fu and Li's F				
CIM	0.20	0.181 (0.161, 0.206)	99.6	0.348 (0.305, 0.389)
NZM	0.20	0.170 (0.152, 0.193)	99.6	0.371 (0.327, 0.423)
SAM	0.20	0.184 (0.167, 0.205)	99.5	0.436 (0.386, 0.484)
TON	0.20	0.191 (0.175, 0.220)	99.5	0.490 (0.436, 0.554)
Tajima's D				
CIM	0.03	0.028 (0.023, 0.034)	99.8	0.165 (0.147, 0.189)
NZM	0.04	0.037 (0.032, 0.045)	99.8	0.174 (0.156, 0.194)
SAM	0.02	0.019 (0.016, 0.024)	99.8	0.191 (0.169, 0.219)
TON	0.03	0.026 (0.022, 0.031)	99.8	0.188 (0.165, 0.213)
Zeng's E				
CIM	2.00	1.000 (1.000, 1.000)	99.5	0.436 (0.405, 0.472)
NZM	2.00	1.000 (1.000, 1.000)	99.5	0.434 (0.401, 0.470)
SAM	2.00	1.000 (1.000, 1.000)	99.6	0.389 (0.361, 0.422)
TON	2.00	1.000 (1.000, 1.000)	99.6	0.389 (0.362, 0.423)

Percentile is the more conservative FDR quantile bin of the permuted data that was equivalent to either the 1st (lower tail) or 99th (upper tail) percentile in the observed data.

estimated by using the value for the 1st or 99th percentile and finding the nearest conservative quantile bin using the median, that was either larger for the lower tail, or smaller for the upper tail and taking the inverse of enrichment for the bin (Table 3.1). As a result, the estimate will be conservative compared to the true FDR but will still maintain FDR control.

The FDR for Zeng's E was 1.0 for all Polynesian populations in the lower end of the distribution and the 1st percentile of the data fell within the 2nd percentile bin from the permutations. This means that the 1st percentile threshold from the observed Zeng's E fell between the 1st and 2nd percentile of the null distribution and was entirely within the 95% confidence interval. Whereas, for Tajima's D and Fay and Wu's H the maximum quantile bin for the Polynesian populations in the lower tail was 0.07% (NZM), with the maximum upper confidence interval being 0.076 (NZM). For the purpose of reporting on individual statistics, a threshold of 0.1 was applied, to reduce the expected false positive rate to be less than 1 in 10 on average. Fu and Li's F and Zeng's E both had an FDR above this threshold for the Polynesian populations and will be reported on only as additional evidence when they overlap with other results or were in a region of contiguously depressed score (section 3.2.1).

3.3.2 Comparison with prior publications

The selection and neutrality statistics that were in the 1st percentile and < 0 were compared to regions and genes identified in Hider *et al.* (2013) for East Asian Super Population (EAS) and Jonnalagadda *et al.* (2017) for South Asian Super Population (SAS) as these papers had used a similar, but not identical methodology to identify regions under selection. Hider *et al.* (2013) investigated selection in the East Asian populations (Han Chinese in Beijing China (CHB), Southern Han Chinese (CHS), and Japanese in Tokyo Japan (JPT)) of the 1000 Genomes Project in the Phase 1 release. For this they used a non-overlapping window size of 25 kb and excluded windows with less than 10 segregating sites and then looked at the top 1% of results. They reported regions that were significant for Tajima's D , Fay and Wu's H , Fu and Li's F , and iHS. Comparing the regions from Hider *et al.* (2013) with regions that met the significance threshold for this project there were 118 of 206 regions for Tajima's D , 15 of 151 for Fay and Wu's H , 83 of 255 for Fu and Li's F , and 5 of 85 for iHS (Table S1).

Of the genes reported in Jonnalagadda *et al.* (2017), there were 2 of 5 for Tajima's D (*DST* and *SLC24A5*), 0 of 5 for Fay and Wu's H , and 2 of 5 for iHS (*TYR* and *ADAM17*) that were also significant in this project. Jonnalagadda *et al.* (2017) used the same window set up as Hider *et al.* (2013), but used the 1000 Genomes Project Phase 3 release for the Gujarati Indian from Houston Texas (GIH) and Indian Telugu from the UK (ITU) populations.

In conjunction with the studies from Hider *et al.* (2013) and Jonnalagadda *et al.* (2017), Ramos (2017) looked at evidence of selection in rheumatic disease, including gout, using the iHS results from Voight *et al.* (2006) from the HapMap phase II dataset. There were three populations used which were Asian, European, and African. When comparing the loci reported in Ramos (2017) with the iHS results, here the equivalent populations of CHB and JPT were used to represent the Asian population, Utah Residents (CEPH) with Northern and Western Ancestry (CEU) for European, and YRI for African. Where the population was unknown in Ramos (2017), all of the specified populations (CEU, CHB,

JPT, and YRI), from the 1000 Genomes Project individuals were used. Loci reported for European had 47.8% overlap with CEU. The Asian reported loci matched 31.6% with CHB and JPT, and for African, 23.1% of the reported loci were also found in YRI. For the loci reported with an unknown population, 65.0% were found to match in any of CEU, YRI, CHB, or JPT.

To supplement the Ramos (2017) iHS regions, additional genes were included from Voight *et al.* (2006) Table 1 to include genes that showed evidence of selection across different combinations of the populations and were not specific to rheumatic disease. The populations covered by these genes were CEU, CHB, CHS, JPT, and YRI. The genes included were: *LCT* - specific for CEU; *SLC44A5* - specific for CHB, CHS, and JPT; *NCOA1*, *ADCY3* and *SYT1* - specific for YRI; *SNTG1* - specific for CEU, CHB, CHS, JPT, and YRI; and *SPAG4* - specific for CEU and YRI. All genes except *SPAG4* were positive for significant iHS markers in the corresponding populations they were originally reported in.

The differences in regions being found as significant can be partially attributed to the differing window sizes, and the difference between sequence and chip-based genotyping for Hider *et al.* (2013) and Jonnalagadda *et al.* (2017), and difference in markers for Ramos (2017).

3.3.3 Selection in Polynesian populations - genome-wide analysis

In this subsection results will be presented first for those that were in common between all Polynesian populations and the sub-groups of East and West Polynesia. Following this, the results for individual Polynesian populations will be presented. A complete table of all genes that had results meeting the various thresholds for the various statistics used can be found in the Appendix Tables S2 (site frequency spectrum (SFS)-based statistics), S3 (iHS and nSL), and S4 (XP-EHH).

There were 465 genes that were associated with urate and metabolic diseases from the genome-wide association study (GWAS) catalog (see Table S7 for references), with 152 having at least a single window from the lower tail, or SNP meeting a threshold for a single statistic in the Polynesian populations. Genes reported here were highlighted based on three criteria: a significant haplotypic result in multiple Polynesian populations, or having multiple statistics meeting the significance threshold in at least a single Polynesian population, or the locus only having the main significant result in Polynesian populations, with minimal signal in other populations, that is, multiple significant markers in Polynesian populations and one or two in other populations.

Each Polynesian population had a similar number of windows in the 1st percentile, except for the Western Polynesian populations with Zeng's *E* where they had up to ~50% less (Table 3.2). The Eastern Polynesian populations tended to have more independent non-consecutive regions and number of genes that were intersected by windows than the Western Polynesian populations. However, this did not translate into a higher number of genes that were significant in only a single population. There were similar numbers of genes that intersected the windows that met the significance threshold in either distribution tail between the Polynesian populations of CIM, NZM, SAM, and TON.

Table 3.2: Number of significant regions and genes in Polynesian populations for the intra population SFS statistics.

Pop	Windows	Lower Tail			Windows	Upper Tail		
		Regions (n)	Genes (n)	Pop. Genes (n)		Regions (n)	Genes (n)	Pop. Genes (n)
Tajima's D								
CIM	2420	590	644	40	2420	963	466	66
NZM	2429	577	713	37	2429	954	436	42
SAM	2396	559	537	50	2396	961	479	75
TON	2380	561	508	48	2380	972	512	72
Fay and Wu's H								
CIM	2420	807	460	25	2420	895	553	57
NZM	2429	770	443	20	2429	916	573	53
SAM	2396	766	470	23	2396	895	565	36
TON	2380	764	467	41	2380	879	575	43
Fu and Li's F								
CIM	2419	655	805	72	2420	818	386	32
NZM	2429	624	966	92	2429	797	379	21
SAM	2396	620	671	79	2396	812	458	47
TON	2380	669	643	96	2380	782	467	50
Zeng's E								
CIM	2143	810	754	36	2418	959	435	24
NZM	2335	865	894	43	2428	953	431	24
SAM	1429	573	508	28	2396	895	448	28
TON	1140	476	405	11	2378	897	467	43

Regions is the number of independent non-consecutive regions the significant windows form. Genes is the number of unique genes that the significant windows intersected. Pop. Genes is the number of genes that had significant windows intersect in only that particular population.

3.3.3.1 Polynesian super population genome-wide selection analysis

Out of all the genes that met the thresholds, the Polynesian super population had the fewest genes unique to a super population (Figure 3.3). The “upset plot” shows the number of genes that were in common between super populations that had both intra-populational haplotypic and SFS evidence. There were 111 genes that only had evidence of selection in Polynesian populations, and 11 genes that had evidence in all super populations.

From the 1st percentile, 77 genes had at least one marker that was significant in the haplotypic tests, and one window that intersected the same gene that was in the 1st percentile for a frequency-based statistic and were shared between all the Polynesian populations (Figure 3.4). The number of genes in common between the Eastern Polynesian populations that had both intra-populational haplotypic and SFS support was 385. There were about 30% more genes for the same criteria in the Western Polynesians, with 490 genes. Between the individual Polynesian populations, approximately one third of the genes that met a threshold in a population were not in common with the other Polynesian populations. The NZM population had the most genes (395) that were not shared with any of the other Polynesian populations (Figure 3.4).

3.3.3.1.1 Previously reported Polynesian “thrifty-genes”

There was no signal of selection for the previously reported ‘thrifty genes’ of *PPARGC1A* and *CREBRF* (Myles *et al.*, 2011; Minster *et al.*, 2016). There were no markers or windows that met the significance thresholds for any of the neutrality and selection statistics. This was consistent with the results of Cadzow *et al.* (2016) for *PPARGC1A*. The two identified SNPs for body mass index (BMI) in Samoans

at *CREBRF*, rs12513649 and rs373863828 (Minster *et al.*, 2016) were both absent from the CoreExome SNP array. While there was no selection signal from iHS or nSL at *CREBRF* (chr5:172483355-172566291), downstream at chr5:172,800,000-173,000,000 there were four markers for SAM and seven for TON that were significant for iHS (Figure 3.2). Rs373863828 is specific to Polynesian populations, and therefore, important for detecting selection at *CREBRF*. At the same location for nSL there were six significant markers for SAM and seven for TON. The populations from the EAS super population had fewer significant markers for both iHS and nSL, with most of the populations having less than three, although Chinese Dai in Xishuangbanna China (CDX) had five significant markers for iHS and seven for nSL. The significant markers at this location in SAM were consistent with Minster *et al.* (2016), were intergenic, and intersected the long intergenic non-coding RNA, *CTB-164N12.1*.

3.3.3.1.2 Genes with possible selection in Polynesian populations

A complete list of all of the genes that had windows in the 1st percentile and values < 0 for the Polynesian populations can be found in table S2. A complete list of genes that had significant iHS, nSL, or XP-EHH in Polynesian populations can be found in Appendix Tables S3, and S4.

There were 10 genes (*ANO1-AS1*, *CCDC180*, *COL6A3*, *ESPNL*, *GRIP1*, *LINC00661*, *LINC01006*, *NPAT*, *NPFFR1*, and *PPA1*) for which all Polynesian populations and no other super populations had at least one SNP with a significant iHS marker, and there were 659 genes that at least one Polynesian population and no other super populations had selection signals for. The Eastern Polynesian populations had 284 genes, similarly, the Western Polynesian populations had 296 genes. This compared to the significant nSL markers where there were 4 genes (*GRIP*, *LINC00661*, *PITPNC1*, and *SDC2*), that had at least one significant marker and were found in only all four Polynesian populations. Within only Eastern Polynesian populations, there were 255 genes that had a significant marker, and within the Western Polynesian populations, there were 244 genes only significant in that group, for nSL. There were two genes in common that had at least one SNP in only the four Polynesian populations, when the intersection of Polynesian specific genes for both iHS and nSL were compared. The genes were *GRIP1* (Glutamate Receptor Interacting Protein 1) and *LINC00661* (Long Intergenic Non-Protein Coding RNA 661).

3.3.3.1.3 Pathway enrichment of genome-wide selection results

Pathway enrichment analysis was performed for each intra-population statistic separately, by taking all the genes that met the threshold for the haplotypic-based statistics, or met the lower tail of the distribution threshold (SFS-based statistics). Pathway terms from the pathway enrichment analysis from the Enrichr KEGG 2016 table showed there was minimal overlap in the terms that were significant between the Polynesian populations (Table 3.3). There were more pathways in common between the Eastern Polynesian populations (10) than the Western Polynesian populations (1). The individual population results are described in each population's subsection (subsections 3.3.3.2.1, 3.3.3.3.1, 3.3.3.4.1, and 3.3.3.5.1).

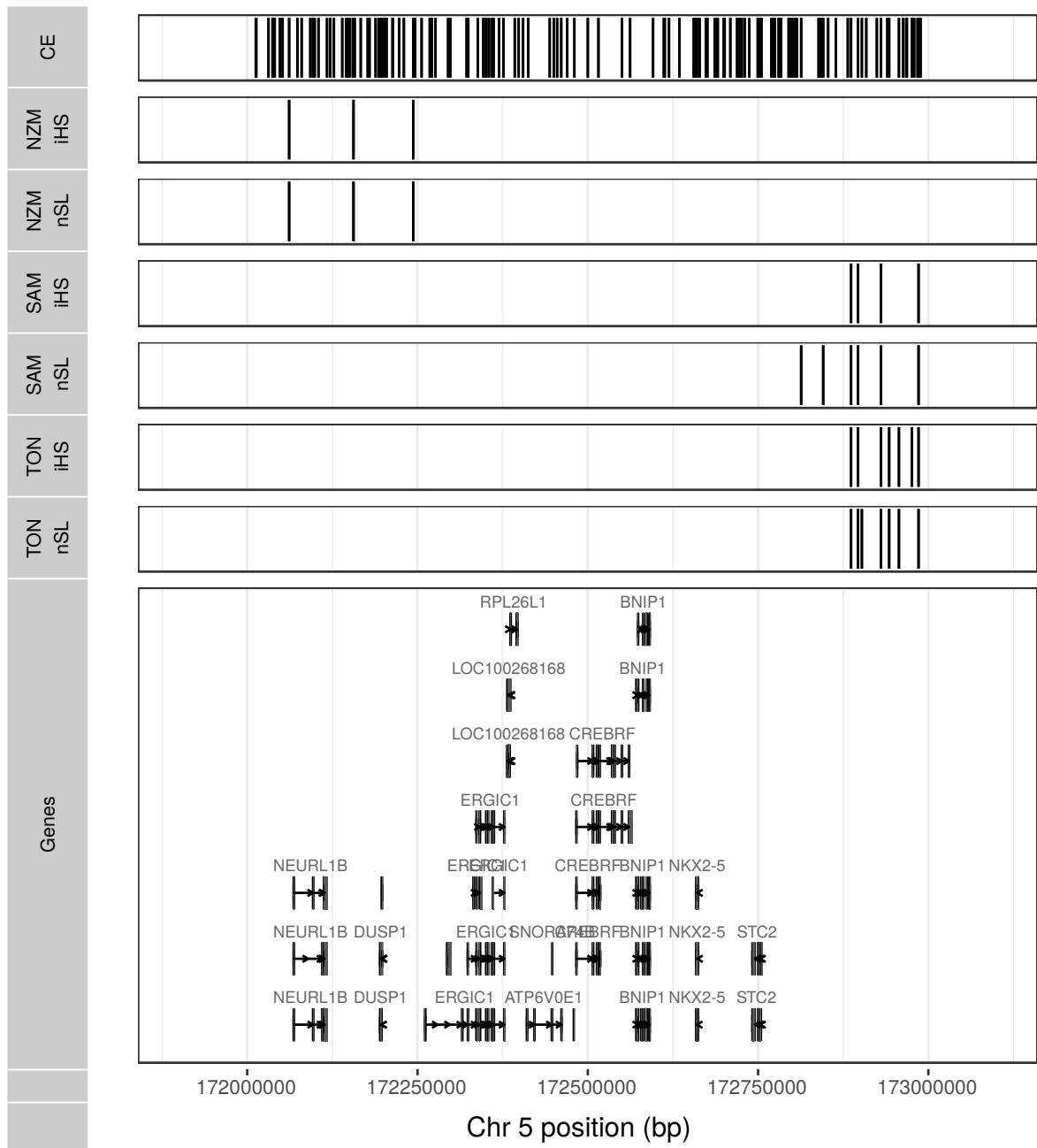


Figure 3.2: Schematic of the *CREBRF* locus with flanking regions, at chr5:172,600,000-173,000,000. Positions of the markers included on the Core-Exome SNP array are indicated by vertical lines (top). Positions for the significant iHS and nSL markers are shown for NZM, SAM, and TON. There were no significant markers for CIM. Exons for genes are indicated by rectangles, and direction of transcription indicated by arrow heads (bottom).

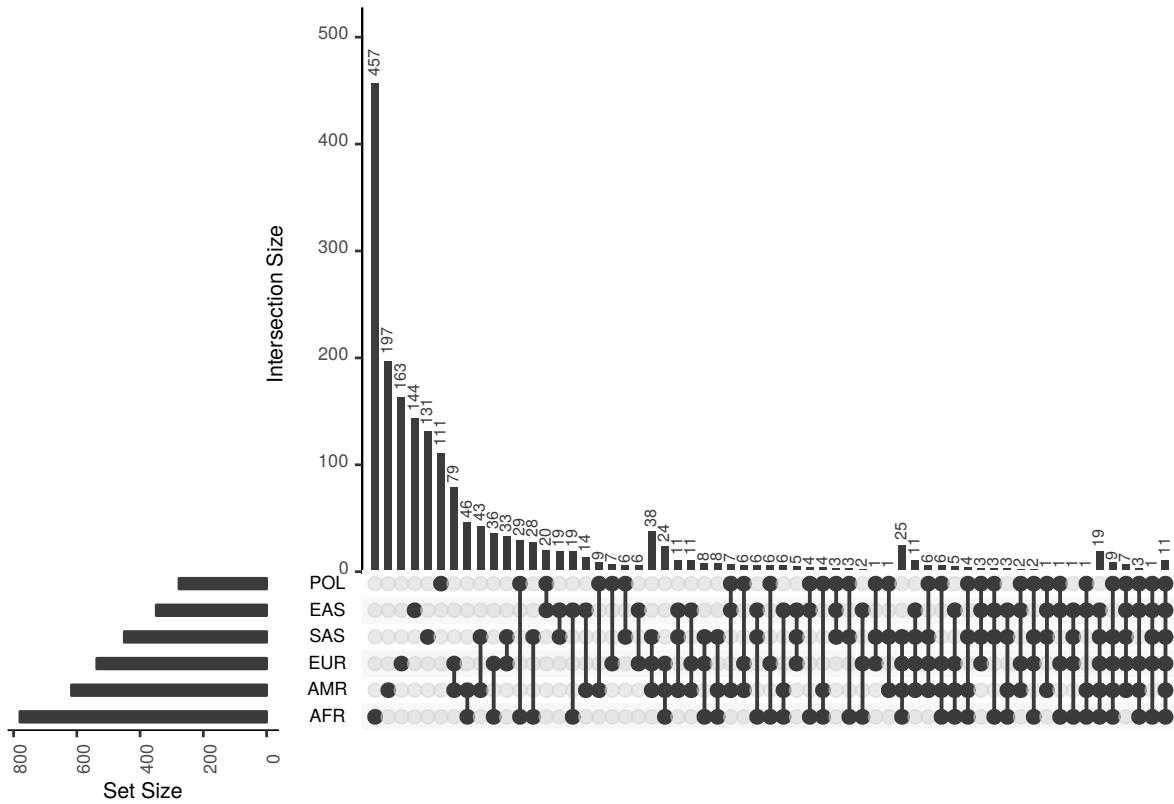


Figure 3.3: Upset plot showing the intersections of genes that had both intra-population haplotypic and frequency spectrum-evidence at the individual population level and were then pooled into their super population group. The left histogram shows the total number of genes with both haplotypic and frequency spectrum-based evidence for each super population. The top histogram is the number of genes that were in common between the super populations. The dots indicate the particular set of super populations for the intersection and are ordered by number of intersecting sets.

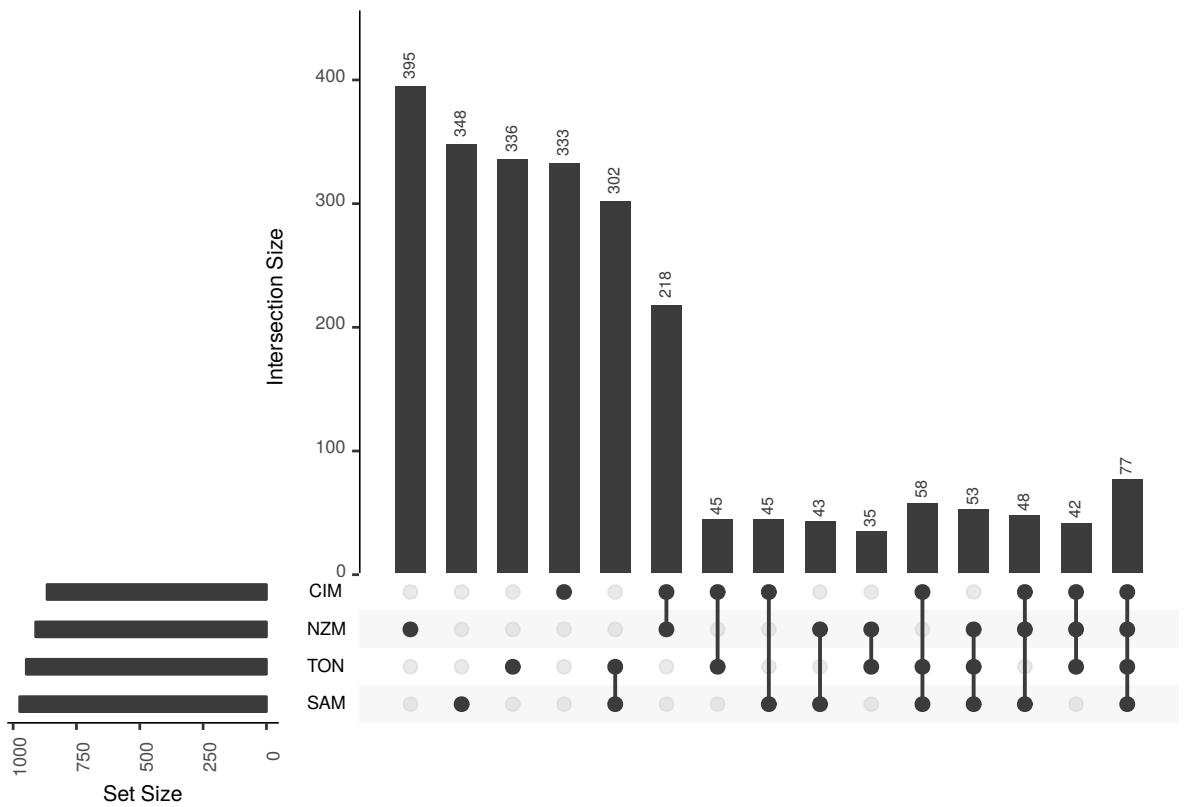


Figure 3.4: Upset plot showing the number of genes that have evidence from iHS or nSL, and met the lower threshold from any of Tajima's D , Fay and Wu's H , Fu and Li's F , or Zeng's E , in the four Polynesian populations. The left histogram shows the total number of genes with both haplotypic and frequency spectrum-based evidence for each Polynesian population. The top histogram is the number of genes that were in common between the Polynesian populations. The dots indicate the particular set of populations for the intersecting sets.

Table 3.3: Pathway terms that were significant after multiple-testing adjustment in Polynesian populations in Enrichr KEGG 2016 pathway enrichment analysis.

Pathway Term	Population			TON
	CIM	NZM	SAM	
ABC transporters	nSL (12/148) P = 0.010	nSL (13/148) P = 0.005	iHS (9/44) P = 0.013	iHS (9/44) P = 0.006, nSL (6/44) P = 0.036
Adrenergic signaling in cardiomyocytes	nSL (9/74) P = 0.007	nSL (8/81) P = 0.010	nSL (8/74) P = 0.033	nSL (8/74) P = 0.033
Aldosterone synthesis and secretion		nSL (9/74) P = 0.005		
Atrial fibrillation right ventricular cardiomyopathy (ARVC)	Tajima's D (11/78) P = 0.009	Tajima's D (11/78) P = 0.021		
Axon guidance	nSL (12/180) P = 0.032	nSL (14/180) P = 0.005		
Bacterial invasion of epithelial cells	nSL (16/167) P = 5.877 × 10 ⁻⁴	nSL (7/78) P = 0.028		
Calcium signaling pathway		nSL (13/167) P = 0.007		
Cardiac muscle contraction		nSL (10/111) P = 0.007		
cGMP-PKG signaling pathway		nSL (10/95) P = 0.005		
Cholinergic synapse		nSL (9/90) P = 0.007		
Circadian entrainment	nSL (9/90) P = 0.011	nSL (9/90) P = 0.043		
Dilated cardiomyopathy		nSL (9/88) P = 0.007		
ECM-receptor interaction		nSL (13/114) P = 8.374 × 10 ⁻⁴		
Endocrine and other factor-regulated calcium reabsorption		nSL (7/91) P = 0.045		
Gap junction		nSL (7/83) P = 0.037		
Glutamatergic synapse		nSL (10/98) P = 0.005		
GnRH signaling pathway		nSL (7/85) P = 0.040		
Hypertrophic cardiomyopathy (HCM)	nSL (9/83) P = 0.009	nSL (6/66) P = 0.043		
Inflammatory mediator regulation of TRP channels		nSL (7/91) P = 0.045		
Insulin secretion		nSL (9/66) P = 0.021		
Long-term potentiation	nSL (12/158) P = 0.013	nSL (12/158) P = 0.009		
Morphine addiction	nSL (11/122) P = 0.009	nSL (9/96) P = 0.009		
Olfactory transduction		nSL (9/122) P = 0.028		
Oxytocin signaling pathway		nSL (9/90) P = 0.007		
Pancreatic secretion		nSL (12/211) P = 0.043		
Platelet activation		nSL (6/56) P = 0.024		
Protein digestion and absorption		nSL (7/64) P = 0.011		
Rap1 signaling pathway		nSL (9/101) P = 0.010		
Regulation of actin cytoskeleton		nSL (7/89) P = 0.043		
Regulation of lipolysis in adipocytes		nSL (9/112) P = 0.018		
Renin secretion		nSL (7/71) P = 0.018		
Retrograde endocannabinoid signaling		nSL (5/48) P = 0.043		
Salivary secretion		nSL (10/120) P = 0.010		
Serotonergic synapse		nSL (10/120) P = 0.010		
Thyroid hormone synthesis				
Type II diabetes mellitus				
Vascular smooth muscle contraction				

Statistic the pathway enrichment was significant for is provided. Numbers in parentheses are number of genes present and the total number of genes in the pathway. P values are adjusted for multiple-testing.

Table 3.4: Number of significant windows/markers by selection or neutrality statistic by population, grouped by super population.

Population	Lower Tail				Upper Tail				iHS SNPs	nSL SNPs
	Fay and Wu's <i>H</i>	Fu and Li's <i>F</i>	Tajima's <i>D</i>	Zeng's <i>E</i>	Fay and Wu's <i>H</i>	Fu and Li's <i>F</i>	Tajima's <i>D</i>	Zeng's <i>E</i>		
AFR										
ACB	2452	2345	1230	2438	2453	2453	2452	2452	3590	2730
ASW	2452	2243	1355	2442	2453	2453	2451	2451	3555	2704
ESN	2424	875	780	2419	2424	2424	2423	2423	3478	2681
GWD	2435	1267	902	2435	2435	2435	2435	2435	3504	2607
LWK	2438	1210	861	2436	2438	2438	2438	2438	3788	2802
MSL	2424	1026	871	2424	2424	2424	2424	2424	3651	2629
YRI	2424	891	749	2425	2425	2426	2426	2426	3576	2785
AMR										
CLM	2462	2463	2463	2221	2462	2463	2463	2463	3168	2490
MXL	2451	2451	2449	2442	2451	2451	2451	2445	3043	2313
PEL	2448	2448	2448	2447	2448	2448	2448	2448	2786	1953
PUR	2465	2463	2465	2289	2465	2465	2465	2462	3110	2585
EAS										
CDX	2387	2387	2387	1013	2387	2387	2387	2384	2558	1949
CHB	2395	2395	2396	1058	2396	2396	2395	2395	2637	1918
CHS	2387	2387	2387	952	2387	2387	2387	2385	2468	1786
JPT	2377	1848	2378	788	2377	2379	2379	2377	2566	2032
KHV	2397	2398	2398	1077	2398	2398	2397	2398	2794	2025
EUR										
CEU	2445	2445	2445	2365	2445	2445	2443	2445	2770	2267
FIN	2436	2436	2436	1814	2436	2436	2436	2436	2613	2205
GBR	2442	2441	2441	2277	2440	2442	2440	2440	2722	2130
IBS	2456	2456	2456	2314	2456	2456	2455	2454	2864	2185
NZC	2452	2453	2453	2448	2453	2453	2453	2453	2636	2145
TSI	2448	2449	2449	2098	2449	2449	2449	2449	2832	2185
POL										
CIM	2420	2419	2420	2143	2420	2420	2418	2350	1628	
NZM	2429	2429	2429	2335	2429	2429	2428	2428	2345	1442
SAM	2396	2396	2396	1429	2396	2396	2396	2396	2487	1851
TON	2380	2380	2380	1140	2380	2380	2380	2378	2378	1866
SAS										
BEB	2433	2436	2436	1542	2436	2436	2436	2436	2825	2218
GIH	2436	2436	2436	1425	2436	2436	2435	2435	2756	2194
ITU	2433	2420	2433	1337	2433	2433	2432	2430	2738	2230
PJL	2438	2438	2438	1597	2438	2438	2437	2438	2853	2222
STU	2434	2414	2434	1287	2434	2434	2434	2434	2809	2332

3.3.3.1.4 Regions of haplotypic selection in Polynesian populations - genome-wide

The Polynesian populations had the least number of significant SNPs for both iHS and nSL (Table 3.4). For the four Polynesian populations, the mean number of significant SNPs for iHS was 2911.3 (SD 423.4) and 2228.7 (SD 348.4) for nSL (Table 3.4). The Polynesian populations had a mean of 2390 SNPs with an iHS value that was significant, compared to 1697 for nSL. The minimum number of SNPs that had a significant iHS was 2345 and was from the NZM population. For nSL this was also NZM with 1442 SNPs. The maximum number of SNPs was for Luhya in Webuye Kenya (LWK) with 3788 for iHS and LWK with 2802 SNPs. Between significant SNPs for iHS and nSL there was a mean of 1188.8 SNPs in common for the Polynesian Super Population (POL) populations, 1391.2 for EAS, 1728.9 for African Super Population (AFR), 1466.0 for European Super Population (EUR), 1562.0 for American Super Population (AMR), and 1498.4 from SAS.

On average there were 89.5 SNPs in common between the Polynesian populations and the other populations for iHS and 45.3 SNPs for nSL. Whereas, there was a higher number of significant SNPs in common between the Polynesian populations with a mean of 557 for iHS and 375.8 for nSL (Table 3.5). The Eastern/Western Polynesian split was also evident with there being a two to four-fold difference in the number of SNPs in common between the Eastern and Western Polynesian populations for both iHS and nSL. There were 24 genes that had at least 1 marker significant across at least 20 populations for all Polynesian populations for XP-EHH. The Eastern Polynesian populations had 28 genes, and the Western Polynesian populations had 49 genes.

Table 3.5: Number of significant SNPs that were in common between populations.

Population	iHS				nSL			
	CIM	NZM	SAM	TON	CIM	NZM	SAM	TON
AFR								
ACB	69	75	85	68	24	21	29	27
ASW	57	85	92	68	25	36	24	30
ESN	47	57	81	77	15	18	36	38
GWD	84	80	70	61	19	15	16	17
LWK	81	89	95	80	22	21	18	17
MSL	57	56	141	104	24	20	25	29
YRI	67	61	83	77	23	25	30	29
AMR								
CLM	80	84	113	129	35	45	59	68
MXL	91	87	104	92	56	43	60	64
PEL	72	101	97	82	37	40	63	62
PUR	59	58	99	76	27	26	33	38
EAS								
CDX	80	89	154	140	71	65	127	146
CHB	115	138	189	168	74	73	98	129
CHS	98	106	215	204	76	77	109	138
JPT	113	94	123	135	90	63	86	129
KHV	91	112	248	231	56	63	118	136
EUR								
CEU	62	46	73	75	30	28	26	34
FIN	71	50	99	91	32	23	38	42
GBR	71	55	79	71	31	26	26	26
IBS	50	43	64	71	27	21	27	35
NZC	62	44	66	67	26	22	20	22
TSI	54	43	75	73	23	14	24	24
POL								
CIM	2670	863	330	330	1796	542	208	212
NZM	863	2671	284	260	542	1634	178	192
SAM	330	284	2939	1275	208	178	2131	923
TON	330	260	1275	2802	212	192	923	2158
SAS								
BEB	84	63	93	85	46	32	45	56
GIH	72	48	94	103	23	26	44	54
ITU	86	71	84	94	37	38	55	53
PJL	98	63	97	94	41	33	53	46
STU	100	90	104	79	30	41	77	59

3.3.3.1.5 F_{ST}

F_{ST} was calculated in sliding windows of 10 kb across the genome, pair-wise between the Polynesian populations and all other populations. Similar to the results for the F_{ST} calculated on entire chromosomes (section 4.3.1.4), the mean F_{ST} using windows had the Polynesian populations most differentiated from the AFR populations. The range of mean F_{ST} was from 0.148 to 0.203. The Polynesian populations were least differentiated from the EAS populations with a range of F_{ST} means from 0.051 to 0.076. Between the Polynesian populations the mean F_{ST} ranged from 0.004, between the Western Polynesian populations of SAM and TON, to 0.029 between NZM and SAM. The Eastern Polynesian populations had a marginally higher mean F_{ST} than the Western Polynesian populations at 0.006. The largest maximum F_{ST} of 0.888, across four windows, was between TON and Mende in Sierra Leone (MSL) at chr17:62455002-62495001. This same region was the maximum F_{ST} for the CIM and NZM, and the second largest for SAM, all with MSL. The maximum F_{ST} between the Polynesian populations was between NZM and TON, with a value of 0.346 at chr10:110685002-110695001. The smallest maximum F_{ST} of the Polynesian populations was between SAM and TON, with a value of 0.081 at chr1:3965002-3975001.

3.3.3.2 Cook Island Māori genome-wide selection analysis

3.3.3.2.1 Pathway enrichment analysis of genome-wide selected loci in CIM

Pathway gene-set enrichment analysis on gene lists for each statistic that met the significance thresholds was done using Enrichr. There was a total of 10 pathways that after multiple testing correction had significant P values (Table 3.3). Nine pathways were significant for the gene list from nSL and had a heart and lung focus based on calcium ion movement in the form of genes for calcium transporters and voltage-gated calcium channels. There were five genes related to calcium that were in at least five of the pathway terms: *CACNA1D*, *ADCY9*, *SLC8A1*, *CACNA2D2*, *CACNA2D3*, with all of the 15 markers except 3, being in favour of the ancestral allele. There was a single pathway that was significant from the genes from the 1st percentile of Tajima's *D* which was "bacterial invasion of epithelial cells". This could be due to selective pressure for bacterial resistance.

3.3.3.2.2 Contiguous regions of score depression in CIM

Looking at regions of contiguous depressed score that also had either a significant iHS or nSL marker in them, there were a total of four regions that met this criteria. For Tajima's *D*, there were two regions. The first, a 290 kb region at chr2:24215002-24505001 intersected *FAM228B* and had a single significant marker for both iHS and nSL (rs13035774). The second, was 470 kb in length at chr11:68105002-68575001, and included *LRP5*, which had a single significant marker (rs634008) for iHS. Fu and Li's *F* also had a region that covered *LRP5* but at chr11:68085002-68355001 and was 270 kb long. For Fu and Li's *F*, there was a 250 kb region at chr16:3545002-3795001 that intersected *DNASE1* with a single marker (rs13926) for iHS and nSL, and also *TRAP1* with two markers (rs13926 and rs1639150) each for iHS and nSL. Fay and Wu's *H* had a single 250 kb region that intersected *MYT1L* at chr2:2015002-2265001 and had four significant markers for iHS (rs4571084, rs13404264, rs11888121, and rs13382326) and four for nSL (rs4571084, rs12470297, rs11888121, and rs13382326), outside this region but surrounding and still within the gene.

3.3.3.2.3 Genome-wide selection from haplotypic statistics in CIM.

The significant iHS and nSL values after conversion to a P-value are shown by position across the genome as a Manhattan plot in Figure 3.5. The conversion of the iHS or nSL value to a P value used $1 - P(|Z|)$, where Z was the iHS or nSL value, due to the similarity of the iHS and nSL to a Z-score. The most extreme markers for iHS (by genomic position, $-\log_{10}(P) > 4.8$) were chr2: rs11683451, rs2890456, and rs6755308; chr4: rs17060079; chr8: rs2319924; chr11: rs7931930 and rs11212617; chr20: rs647518; chr21: rs28559700. The most extreme markers for nSL (by genomic position, $-\log_{10}(P) > 4.8$) were chr4: rs1491411; chr11: rs543215 and rs1963626; chr12: rs7977414. Within the most extreme 100 iHS values, there were 37 genes. For nSL there was 45 genes. Eighteen of these genes were in common, with *C11orf65* and *CNTN4* having at least five significant SNPs each. The gene with the most significant SNPs out of any population was *CNTN4* in CIM with 27 markers having a significant iHS score. The only other population with more than two significant SNPs was NZM with 14. There were four genes represented in the top 100 markers that were associated with obesity or type 2 diabetes, these were: *ADCY9*, *ARL15*, *PSMD6*, and *ZFAND6*. In total there were 124 genes that were only significant in CIM for iHS, and 102 for nSL. There were 57 genes that had at least 20 populations with significant markers for XP-EHH in only CIM from the Polynesian populations. *ARL15* was associated with

obesity and type 2 diabetes (Mahajan *et al.*, 2014; Shungin *et al.*, 2015) and *RREB1* was associated with type 2 diabetes and urate (Köttgen *et al.*, 2013; Mahajan *et al.*, 2014).

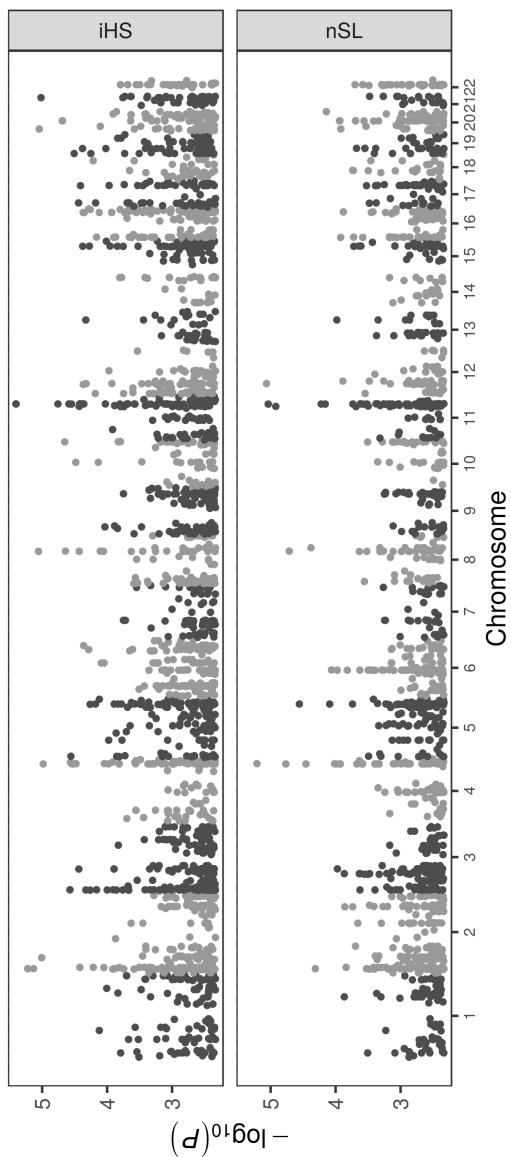


Figure 3.5: Manhattan plot for significant markers in CIM for both iHS and nSL.

3.3.3.3 New Zealand Māori genome-wide selection analysis

3.3.3.3.1 Pathway enrichment analysis of genome-wide selected loci in NZM

The pathway gene-set enrichment analysis from Enrichr on the KEGG 2016 pathways had 32 pathways that were significant after multiple testing correction, this was the largest number of pathways out of the Polynesian populations (Table 3.3). On average only 9.3% of a pathway was represented by the genes. Thirty-one pathways were from the significant genes for nSL, with many in common with CIM.

Calcium transporters and voltage gated calcium channels featured in many pathways that were in common. Compared to the CIM pathway results, there were more hormone-based pathways that were significant, such as renin secretion, oxytocin signalling pathway, adrenergic signalling in cardiomyocytes, thyroid hormone synthesis, aldosterone synthesis and secretion, insulin secretion, and GnRH signalling pathway. Furthermore, diabetes-related pathways were significant such as type 2 diabetes mellitus, insulin secretion, and pancreatic secretion pathways. The genes that were shared in the most pathways included two adenylate cylases (*ADCY8* and *ADCY9*), two phospholipase C genes (*PLCB1* and *PLCB4*), and calcium voltage-gated channel subunits (*CACNA1C* and *CACNA1D*). Similar to CIM, there was a single pathway that was significant from the Tajima's *D* list and that too was for bacterial invasion of epithelial cells.

3.3.3.3.2 Selection from haplotypic statistics in NZM - genome-wide

The significant markers from iHS and nSL were converted to P-values and plotted on a Manhattan plot (Figure 3.6). The most extreme markers for iHS (by position, $-\log_{10}(P) > 4.8$) were: chr2 - rs2710684; chr3 - rs74823804 and rs2280162; chr8 - rs11987519; chr9 - rs2780246; chr10 - rs4075326 and rs11017145; chr16 - rs28564718 and rs3743759. For nSL, the markers were: chr4 - rs1491411; chr8 - rs11987519; chr12 - rs7977414. From the haplotypic tests for selection (iHS and nSL) there were 122 genes that were only significant in NZM for iHS, and 93 for nSL. Looking at the genes represented by the most extreme 100 iHS and nSL markers there were 30 genes for iHS, and 40 genes for nSL. There were 12 genes in common between the two sets (*CTNNA3*, *WWOX*, *KCNS3*, *NRXN1*, *ADAM29*, *VSNL1*, *BCCIP*, *BICD1*, *DHX32*, *SLC35F2*, *STAU2*, and *TENM3*). *CTNNA3* had five markers in the top 100 for nSL (rs2441727, rs12220315, rs1911341, rs2660024, and rs10997250), and was also the gene for NZM that had the most overall significant markers for both iHS and nSL. There were 40 genes that had at least 20 populations with significant markers for XP-EHH in only NZM from the Polynesian populations, only one (*PAX5*) of which was associated with obesity (Melka *et al.*, 2012). There were eight genes, *LYST*, *PHIP*, *CTNNA3*, *RAD51AP2*, *SUPT3H*, *EPB41L4A*, *RAD51B*, and *VSNL1*, that had at least one significant marker for iHS or nSL, and also had a window in the 1st percentile from three or more statistics from Tajima's *D*, Fay and Wu's *H*, Fu and Li's *F*, and Zeng's *E*.

3.3.3.3.3 Contiguous regions of score depression in NZM

The regions of contiguous depressed score from the 1st percentile of windows that also contained markers with a significant iHS or nSL value covered three regions of the genome for NZM. Two of the regions were for Tajima's *D*, the first was 270kb at chr2:24215002-24485001, and overlapped the region in CIM. It had a single significant marker for iHS in each of *MFSD2B/FKBP1B* (rs10185680), and *FAM228B* (rs10197527). The second Tajima's *D* region was 260 kb at chr4:106555002-106815001 and had a single significant marker for nSL in *INTS12* (rs2553453). A 200 kb region for Fu and Li's *F* at chr2:179395002-179595001 had four significant markers in *TTN-AS1* for iHS (rs3731752, rs2278196, rs72648270, and rs3813243).

No contiguous regions of score depression intersected with genes associated with urate, kidney disease, or type 2 diabetes. There were 2 regions that intersected obesity-associated loci; for Tajima's *D*

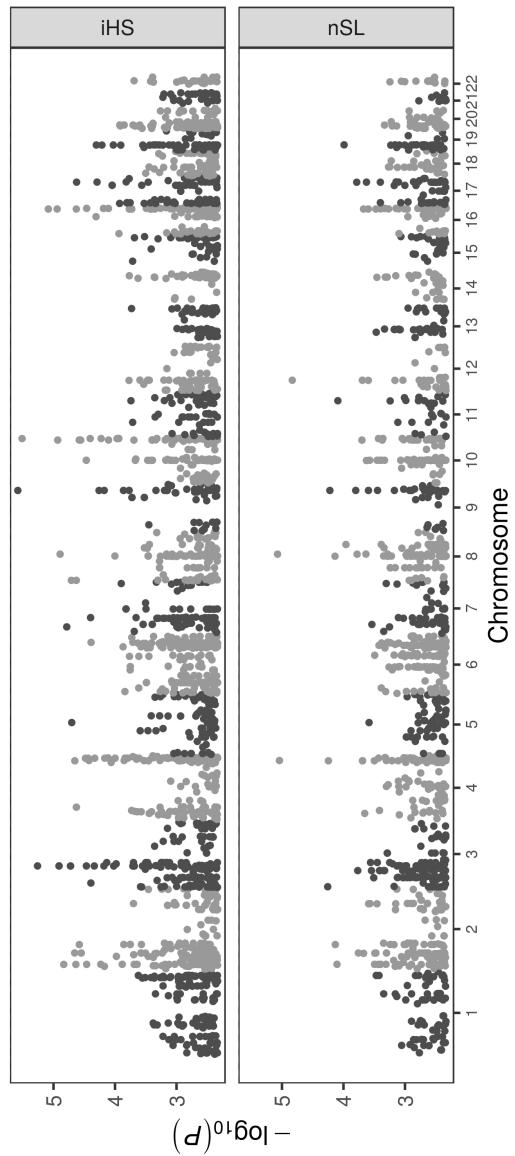


Figure 3.6: Manhattan plot for significant markers in NZM for both iHS and nSL.

a 320 kb region at chr16:67145002-67465001 contained *KCTD19*, and for Fu and Li's *F*, 220 kb at chr16:3555002-3775001 contained *NLRC3*. *EDC4* which is associated with metabolic syndrome (Kristiansson *et al.*, 2012), had slightly differing regions that covered it. Tajima's *D* was 330 kb spanning chr16:67775002-68105001, and Fu and Li's *F* was 320 kb spanning chr16:67775002-68095001.

3.3.3.4 Samoan genome-wide selection analysis

3.3.3.4.1 Pathway enrichment analysis of genome-wide selected loci in SAM

Only two pathways were significant for the gene-set pathway enrichment analysis using Enrichr (Table 3.3). The pathway term “ABC transporters” was significant for gene list from iHS (adjusted P = 0.0128) with 9 of 44 genes. The genes were *ABCC4*, *ABCA5*, *ABCC8*, *ABCC5*, *ABCB5*, *TAP2*, *ABCA9*, *ABCB8*, and *ABCA8*. The second pathway that was significant was “long-term potentiation” and that was from the nSL gene list (adjusted P = 0.0213) with 9 of 66 genes. The genes were *PPP3CA*, *GNAQ*, *RPS6KA1*, *PRKCA*, *CACNA1C*, *PLCB1*, *PLCB2*, *CAMK2G*, and *RAPGEF3*.

3.3.3.4.2 Selection from haplotypic statistics in SAM - genome-wide

The significant markers for iHS and nSL are shown by position across the genome as a Manhattan plot after conversion to P-values in Figure 3.7. The most extreme markers for iHS (by position, $-\log_{10}(P)$) were: chr1 - rs10495181; chr4 - rs12331849; chr11 - rs6578634; chr13 - rs4941616; chr16 - rs12596728. For nSL they were: chr17 - rs11654176 and rs1860316. From the haplotypic tests for selection (iHS and nSL) there were 110 genes that were only significant in SAM for iHS, and 90 for nSL. The genes with the most significant markers for iHS were *CDH23* and *CNTN5* with 12 significant markers each. The gene with the most significant markers for nSL was *CDH23* with 16 markers. Looking at the most extreme 100 markers for nSL there were 35 genes represented, *SNX29* had the most with three markers (rs350277, rs7201595, and rs12931604). For iHS, the gene with the most markers in the most extreme 100 was *C11orf65*, with three (rs425538, rs7931930, and rs11212617). There was a total of 39 genes represented. Between the top markers for both iHS and nSL there were 10 genes that were in common (*SNX29*, *LINC00693*, *BANK1*, *DLC1*, *DUSP13*, *HBE1*, *HBG2*, *MYCBP2*, *NLRP1*, and *SAMD8*).

There were ten genes that had windows from at least three of the frequency-based statistics and also had at least a single significant marker for either iHS or nSL. The genes were *DISP1* (iHS: rs2789931 and rs2789954), *PARD3B* (iHS and nSL: rs13000345), *C4orf45* (iHS and nSL: rs11722868), *NEK1* (iHS: rs4235024), *CNTNAP2* (iHS: rs2620441, rs2249958, and rs17170777; nSL: rs2249958, rs17170777, and rs10255956), *CTNNA3* (iHS: rs1948946; nSL rs4297361, rs1948946, rs2764813, rs2394324, and rs10823054), *CEP112* (iHS: rs11652795; nSL rs1373074), *CNTNAP5* (nSL: rs314710 and rs2602647), *SUPT3H* (nSL: rs9472376), and *DGKI* (nSL: rs12056089). XP-EHH had 19 genes that had at least 20 populations with significant markers in only SAM from the Polynesian populations but none had been associated with urate or related co-morbidities.

3.3.3.4.3 Contiguous regions of score depression in SAM

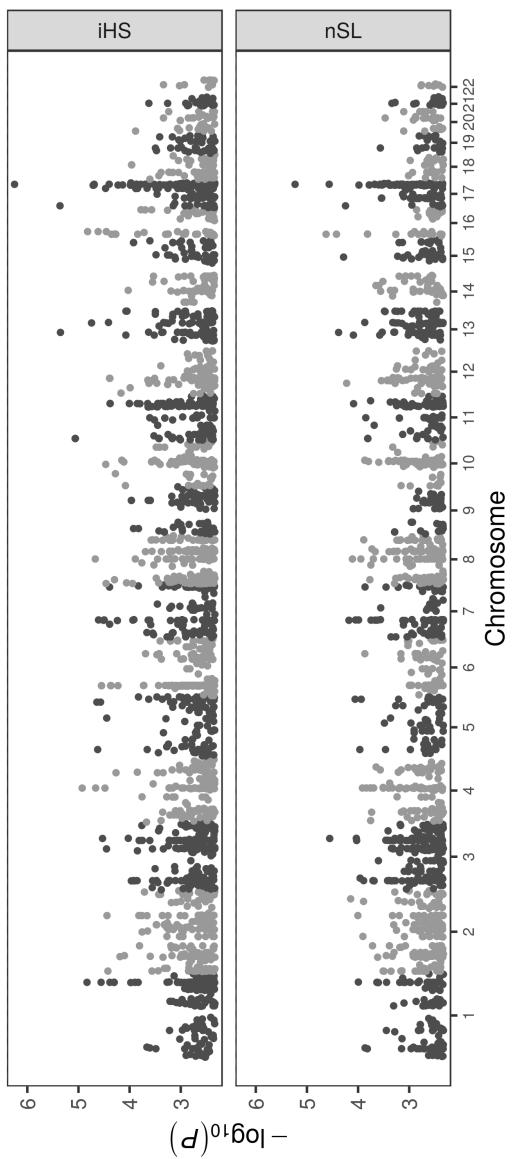


Figure 3.7: Manhattan plot for significant markers in SAM for both iHS and nSL.

There were no contiguous depressed score regions that intersected with genes associated with urate, obesity, type 2 diabetes, kidney disease, or metabolic syndrome. Regions of contiguously depressed score that also had significant markers for iHS or nSL in SAM covered three regions of the genome across two statistics. Tajima's *D* had two regions, the first was 400 kb at chr4:170025002-170425001 and had a single significant marker for iHS (rs4235024) nearby in *NEK1*. The second region was 300 kb in length and was found at chr10:76685002-76985001. It had significant markers in *KAT6B* (iHS and nSL:rs3213967), and in *SAMD8/DUSP13* (iHS and nSL: rs7912300 and rs10824274). Fu and Li's *F* had a single 200 kb region at chr1:26465002-26665001 that had a single significant marker for nSL in *CNKS1* (rs2783633), and two markers for iHS and three for nSL in *AIM1L* (iHS and nSL:rs11247916 and rs10751735; nSL: rs4659431).

3.3.3.5 Tongan genome-wide selection analysis

3.3.3.5.1 Pathway enrichment analysis of genome-wide selected loci in TON

Pathway enrichment analysis from the Enrichr KEGG 2016 pathways produced 11 different pathways that were significant after multiple testing correction (Table 3.3). Of the SFS based methods, only Fu and Li's *F* had a significant pathway and this was "Olfactory transduction". There were two pathway terms that were significant from the iHS results, these were "ABC transporters" and "ECM-receptor interaction". These were also significant in the nSL results, along with another eight terms (Table 3.3). The genes that were included in the "ABC transporters" pathway from the significant markers of both iHS and nSL were: *ABCC8*, *TAP2*, *ABCA9*, *ABCA8*, and *ABCG2*. From only iHS markers *ABCA5*, *ABCA6*, *ABCC5*, *ABCB* were included, and from only nSL markers, *ABCB8* was included. Genes with significant markers for both iHS and nSL in the "ECM-receptor interaction" pathway included *COL4A2*, *ITGB5*, *ITGA1*, *ITGA2*, *SPP1* and *ITGB6*. From only iHS markers *LAMA5*, *TNXB*, *TNC*, *COL6A3*, and *HMMR* were included, and from only nSL makers, *COL4A1*, and *ITGA9* were included. Five of the significant pathway terms were only significant in TON out of the Polynesian populations.

3.3.3.5.2 Selection from haplotypic statistics in TON - genome-wide

The significant markers for iHS and nSL were plotted genome-wide after conversion to P-values in a Manhattan plot (Figure 3.8). The most extreme iHS markers (by position, $-\log_{10}(P)$) were: chr1 - rs12042853, rs10495181, rs4240931, and rs2800853; chr7 - rs10485976, rs6962297, rs296307, rs6969276, rs7791859, and rs1405425; chr8 - rs4349972; chr10 - rs7097067 and rs7923688; chr11 - rs6578634; chr13 - rs4941616; chr16 rs154148 and rs350277; chr17 - rs6504539, rs1860316 and rs16976276. For nSL they were: chr7 - rs7791859 and rs1405425; chr8 - rs2605867; chr13 - rs4941616; chr16 - rs350277. In total there were 105 genes that were only significant in TON for iHS, and 108 for nSL. In the most extreme 100 nSL marker scores, there were 32 genes represented with *SNX29* having three markers (rs350277, rs7201595, and rs12931604), and *CDH23* having two (rs2394801 and rs10762462) from the 100. The most extreme 100 marker scores for iHS had 30 genes represented with *SNX29* having five markers (rs350277, rs7201595, rs7198595, rs7189759, and rs12931604), *PCDH15* having three (rs4272709, rs11004106, and rs4935502), and *DNAH11* having two (rs10485976 and rs1989904) of the

100. There was an overlap between the two gene lists for the most extreme 100 of six genes, with *SNX29*, and *DNAH11* having the largest number of extreme markers for both iHS, and nSL. The genes with the largest number of significant markers overall for iHS and nSL were *CPNE4* (16 markers for both iHS and nSL), *SNX29* (15 markers for iHS and 16 for nSL), *CDH23* (11 markers for iHS and 12 for nSL), *PCDH15* (12 markers for nSL), and *BANK1* (9 markers for iHS and 10 for nSL). There were 37 genes that had at least 20 populations with significant markers for XP-EHH in only TON from the Polynesian populations, including *IGF1R*, *JAZF1*, and *PEPD* which had been associated with urate or co-morbidities.

There were seven genes that had at least one significant marker for either iHS or nSL, and windows in the 1st percentile for at least three frequency-based intra-population selection and neutrality statistics. They were *C4orf45*, *SUPT3H*, *CTNNA3*, *SAMD8*, *EXT2*, *LPO*, and *NRXN3*.

3.3.3.5.3 Contiguous regions of score depression in TON

There were no contiguous depressed score regions that intersected genes associated with urate, type 2 diabetes, kidney disease, or metabolic syndrome. There was one region that intersected the obesity associated gene *CCR3* for Tajima's *D* and was 260 kb in length at chr3:46035002-46295001. There were three regions with contiguous depressed score that also had significant markers for iHS or nSL in TON. Tajima's *D* had two regions. The first similar to SAM, was at chr10:76685002-76985001 and spanned 300 kb and had significant markers in *KAT6B* (iHS and nSL: rs1551067 and rs3213967), *SAMD8* (iHS and nSL: rs10509355, rs10824274, and rs7912300), and in *DUSP13* (iHS and nSL: rs10824274 and rs7912300). The second region at chr17:56125002-56415001 was 290 kb and had a single significant marker (rs9892223) for iHS in both *EPX* and *LPO*. Fu and Li's *F* had a slightly shifted region that also intersected *EPX* and *LPO* and had the same significant markers as the Tajima's *D* region, but also included *RNF43/BZRAP1-AS1* - which also had a single significant marker for iHS (rs2257205). The region was at chr17:56205002-56455001 and was 250 kb in length.

3.3.4 Selection in disease-associated genes

3.3.4.1 Selection in Polynesian populations for genes associated with urate and metabolic disease

In order to establish if there was evidence of selection in genes associated with urate and diseases with hyperuricaemia as a co-morbidity, loci that associated with these conditions were extracted from the GWAS catalog (as per section 2.1.3, see Table S7 for references for each trait) and neutrality and selection statistics that intersected these loci were collated. Table 3.6 shows the breakdown of the number of different statistics by population that met the significance thresholds for the SFS based statistics, the haplotype based statistics, and the overlap between the frequency spectrum and haplotypic methods. The CIM population had the least number of genes of the Polynesian populations that had a significant result out of the genes that were associated with urate. The

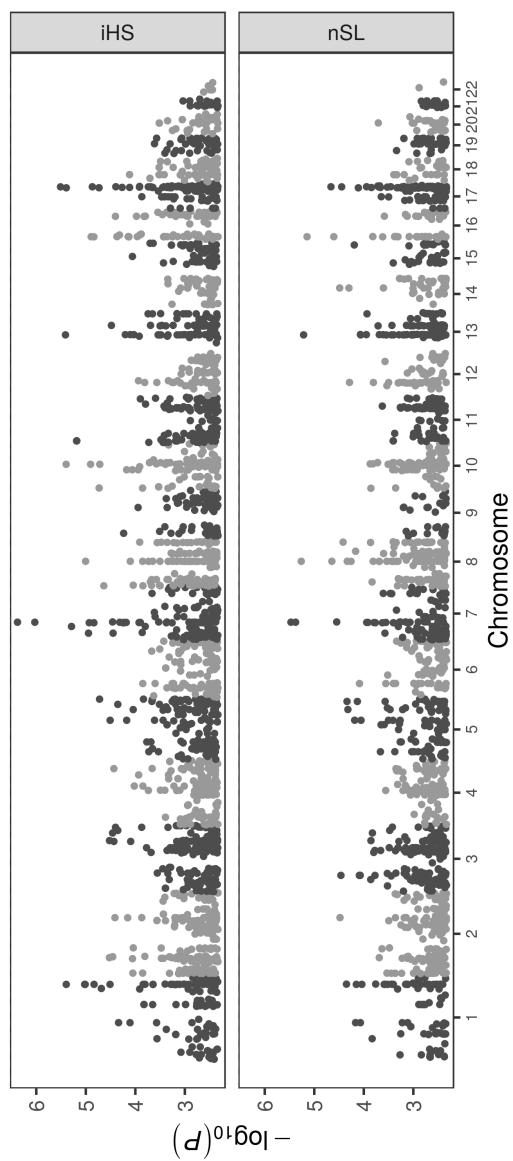


Figure 3.8: Manhattan plot for significant markers in TON for both iHS and nSL.

Polynesian populations had the fewest number of genes with significant iHS or nSL for obesity and the Western Polynesian populations had the fewest number of genes for type 2 diabetes.

Figure 3.9 shows the genes associated with urate, gout and related diseases that had a significant marker with either iHS or nSL and also had a significant window from a SFS-based statistic in at least one Polynesian population. There were only 10 genes that met this criteria, of which, the type 2 diabetes associated gene, *PTPRD*, was the only gene that met the criteria, and also had significant markers from all Polynesian populations in iHS. The obesity-associated genes *GRID1*, *FHIT*, and *ERBB4*, all had windows from Fay and Wu's *H* that were significant in all Polynesian populations. *ERBB4* also had significant windows for all Polynesian populations for Tajima's *D*.

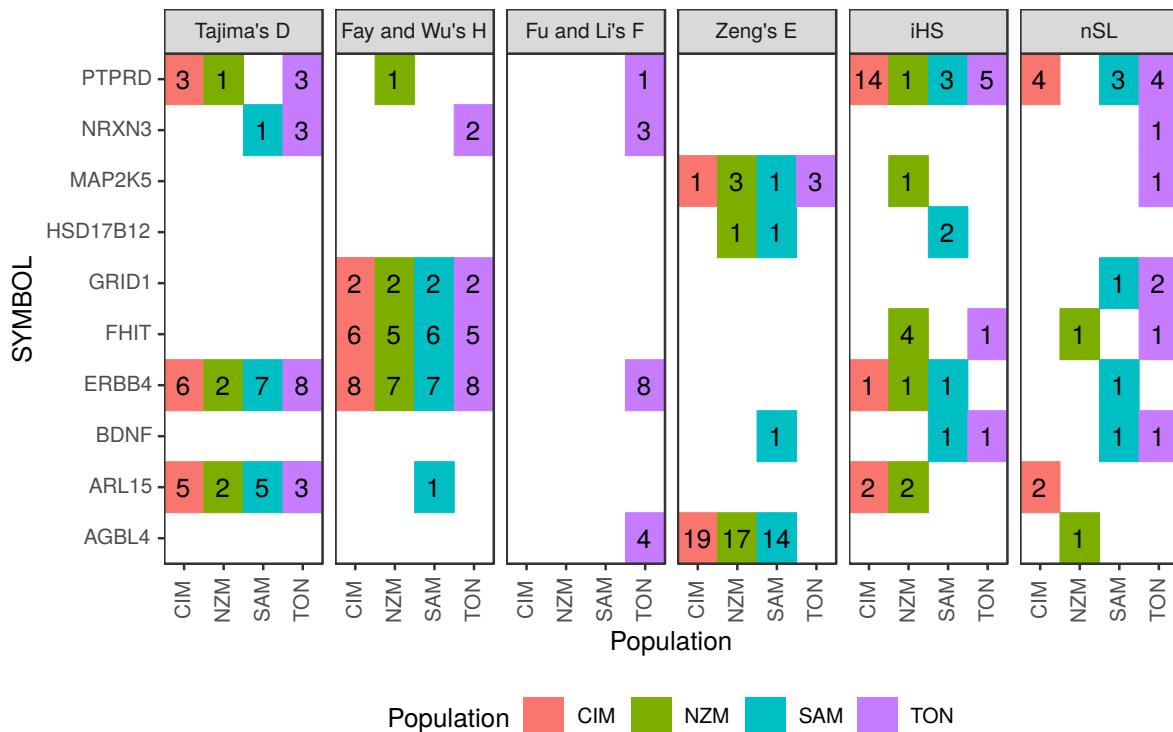


Figure 3.9: Genes associated with urate, gout, obesity, type 2 diabetes, kidney disease, and metabolic syndrome having both significance in intra-population SFS and haplotypic based methods in at least one Polynesian population. Numbers inside boxes indicate the number of significant windows (SFS methods) or SNPs (haplotypic methods) for the population.

Table 3.6: Number of genes with evidence of possible positive selection in super populations from intra-population tests from urate and co-morbidities GWAS associated loci.

Population	Super Pop.	Urate / Gout			Obesity			T2D			Kidney Disease			Metabolic Syndrome		
		SFS	Hap.	Com.	SFS	Hap.	Com.	SFS	Hap.	Com.	SFS	Hap.	Com.	SFS	Hap.	Com.
ACB	AFR	9	9	0	34	46	14	9	18	2	5	5	0	2	0	0
ASW	AFR	8	5	1	31	44	11	8	21	3	10	7	2	4	3	1
ESN	AFR	8	4	0	29	44	7	9	18	3	7	6	1	3	2	0
GWD	AFR	5	4	0	41	43	8	9	15	3	6	6	1	4	3	0
LWK	AFR	5	7	0	30	46	5	8	16	3	5	7	1	3	3	0
MSL	AFR	7	10	1	35	38	10	10	11	3	6	10	2	2	3	1
YRI	AFR	7	5	0	35	40	9	10	13	4	6	7	1	1	0	0
CLM	AMR	10	4	1	30	36	6	8	15	1	8	3	0	2	5	1
MXL	AMR	9	2	0	30	35	5	8	14	0	10	1	0	4	4	1
PEL	AMR	10	3	0	29	29	6	11	18	3	6	2	0	2	2	0
PUR	AMR	7	7	1	37	43	11	9	11	2	6	6	1	3	4	1
CDX	EAS	4	2	0	21	37	6	11	11	3	5	3	1	1	2	0
CHB	EAS	4	2	0	25	28	5	14	11	2	4	3	0	1	0	0
CHS	EAS	7	2	0	22	33	6	10	10	2	5	2	0	1	1	0
JPT	EAS	7	2	0	25	36	5	7	12	1	5	1	0	1	1	0
KHV	EAS	2	3	0	25	34	9	6	11	2	6	1	0	1	1	0
CEU	EUR	13	3	1	31	31	8	8	13	0	7	4	1	2	4	0
FIN	EUR	5	3	1	27	35	5	6	14	0	5	4	1	1	3	0
GBR	EUR	10	4	1	32	35	9	10	16	3	7	3	1	1	3	0
IBS	EUR	9	3	0	34	38	9	8	12	1	7	2	1	1	4	0
NZC	EUR	13	6	2	27	34	7	9	17	2	10	3	1	2	4	0
TSI	EUR	13	4	1	27	35	11	6	14	2	6	1	0	2	7	1
CIM	POL	5	1	0	29	17	2	11	9	2	5	0	0	1	0	0
NZM	POL	8	1	0	31	21	5	13	10	2	4	1	0	2	3	0
SAM	POL	5	4	0	27	20	4	14	4	0	7	3	0	1	1	0
TON	POL	8	5	0	22	25	4	14	7	1	7	3	0	3	2	0
BEB	SAS	7	3	0	25	25	5	7	13	2	10	4	1	0	5	0
GIH	SAS	4	1	0	24	23	6	8	14	3	4	1	0	0	3	0
ITU	SAS	6	3	0	20	24	5	6	14	2	8	2	1	1	4	0
PJL	SAS	7	2	1	21	30	3	7	16	3	7	1	0	0	3	0
STU	SAS	10	0	0	26	28	8	6	12	2	6	3	1	2	3	0

Hap = haplotypic. Com = combined. Combined is the number of genes that had both SFS and haplotypic evidence.

3.3.4.1.1 Selection in Polynesian populations for genes associated with urate and gout

The 62 loci identified in the GWAS catalog (see Table S7 for references) as being associated with urate and gout only had a small proportion with any evidence of possible selection, 18 in Polynesian populations. In the Polynesian populations, from the intra-population neutrality and selection statistics, there were 18 genes that had some evidence. However, there were only nine genes with haplotypic evidence, and only one gene (*RREB1*) that had both haplotypic and SFS-based evidence. Overall, the Polynesian populations did not have more urate genes with evidence for selection than the other populations (Table 3.6).

The two main effect loci for serum urate, *SLC2A9* and *ABCG2* both had limited evidence of selection. *SLC2A9* did not have any values for iHS or nSL because of an extended region of no recombination leading to a large gap in the reference genome which caused the extended haplotype homozygosity (EHH) calculation to be terminated. There was evidence with Fu and Li's *F* which had windows from *SLC2A9* in the 1st percentile from TON. This was also observed with the EAS populations of CHB and JPT. For the Polynesian populations, *ABCG2* only had significant results from iHS and nSL (at the same marker, rs2622626) for TON. With iHS, the SAS populations of GIH and ITU also had significant values for rs2622626. For nSL, along with GIH and ITU, the other non-Polynesian populations that had a significant value was the EAS population JPT, and the AFR population of YRI. Rs2622626 is associated with both urate and gout (Köttgen *et al.*, 2013). There were no Polynesian populations with values in the 1st percentile for the frequency-based statistics, and only Esan in Nigeria (ESN) had windows that met the threshold for Fay and Wu's *H*.

Table 3.7 shows all urate and gout associated genes that had any evidence of possible selection in the Polynesian populations. Some genes of interest included *IGF1R* and *RREB1* which both had significant markers with XP-EHH between nearly all Polynesian populations and many of the other populations excepting those in the SAS super population. *RREB1* is also associated with type 2 diabetes (Mahajan *et al.*, 2014). *BCAS3* had multiple windows that met the lower threshold for Fay and Wu's *H*, Fu and Li's *F*, and Tajima's *D* in the Polynesian populations but did not have any significant markers for the haplotypic tests of selection.

Table 3.7: Urate and gout associated loci that showed signs of possible selection in Polynesian populations.

Gene	Population	XP-EHH											
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's <i>H</i>	Fu & Li's <i>F</i>	Tajima's <i>D</i>	Zeng's <i>E</i>
<i>ABCG2</i>	TON							1	1				
<i>ACVR2A</i>	CIM								1				
<i>ALDH16A1</i>	NZM							1					
<i>ALDH16A1</i>	SAM							1					
<i>ALDH16A1</i>	TON							1					
<i>ATXN2</i>	CIM										1		
<i>ATXN2</i>	NZM										1		
<i>ATXN2</i>	SAM								5		1		
<i>BAZ1B</i>	NZM									3			
<i>BAZ1B</i>	SAM									5			
<i>BAZ1B</i>	TON									3			
<i>BCAS3</i>	CIM								8	4	2		
<i>BCAS3</i>	NZM								6	9	7		
<i>BCAS3</i>	SAM								10	1	4		
<i>BCAS3</i>	TON								3	1			
<i>IGF1R</i>	CIM	2	3	1									
<i>IGF1R</i>	NZM	3	2	1									
<i>IGF1R</i>	SAM	2	4	5	3				1				
<i>IGF1R</i>	TON	6	4	5	6		5		5				

Table 3.7: Urate and gout associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Gene	Population	XP-EHH							Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS		
<i>LRP2</i>	SAM				1		1			
<i>LRP2</i>	TON						2			
<i>LTBP3</i>	CIM								2	1
<i>LTBP3</i>	NZM								2	
<i>LTBP3</i>	SAM									2
<i>LTBP3</i>	TON									2
<i>MLXIPL</i>	TON								2	3
<i>NFAT5</i>	NZM								4	4
<i>NRXN2</i>	SAM				1					
<i>PKLR</i>	NZM								2	
<i>PRKAG2</i>	CIM	1								
<i>PRKAG2</i>	NZM	1								
<i>PRKAG2</i>	SAM						1			
<i>PRKAG2</i>	TON							1		
<i>R3HDM2</i>	CIM									5
<i>R3HDM2</i>	NZM									7
<i>R3HDM2</i>	TON									6
<i>RFX3</i>	TON					1	1			
<i>RREB1</i>	CIM	5	3	5	6		1			3
<i>RREB1</i>	NZM	4	3	5	6				1	1
<i>RREB1</i>	TON	3	1	1	2					2
<i>RREB1</i>	SAM								2	
<i>SLC2A9</i>	TON								7	

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.1.2 Selection in Polynesian populations for genes associated with obesity

There were 269 obesity-associated loci with three loci in common with urate and gout from the GWAS catalog (see Table S7 for references). The Polynesian populations had about 30% (84) of the obesity-associated genes showing some evidence of possible selection (Table 3.6). As a super population, the Polynesian populations had fewer obesity-associated genes that were significant for iHS or nSL compared to the other super populations. Genes that were associated with obesity and had evidence across multiple populations or many significant results are included in Table 3.8. Ten loci had evidence in all four Polynesian populations, from a combination of haplotypic and SFS statistics, these were *ABO*, *ARL15*, *CCR2*, *CCR3*, *COL6A1*, *ERBB4*, *LEKR1*, *PPP2R3A*, *RABEP1*, *SLC39A8*, and *ZBTB38*. *ABO* is discussed further in subsection 3.3.4.2.3 in the context of malaria. There were nine obesity associated loci that of the Polynesian populations, the Eastern Polynesian populations

had either the entire or the largest signal, these were *ADAMTS9*, *ADCY9*, *BTNL2*, *CCNLJL*, *DNAH10*, *GDF5*, *LY86*, *MTIF3*, and *PCSK5*. Conversely, eleven loci had their entire or largest signal in the Western Polynesian populations. These were *BDNF*, *CTSS*, *FER*, *FAM13A*, *GPRC5B*, *GRID1*, *JAZF1*, *LRP1B*, *PARK2*, *PEPD*, and *TRIP11*. Four of the loci, *ADAMTS9*, *ARL15*, *JAZF1*, and *PEPD* were also associated with type 2 diabetes. *FTO* did have a single marker (rs4396532) that was significant for iHS in NZM.

A contiguously depressed score region was in CIM at chr12:124195002-124505001 for Fu and Li's *F* that covered *DNAH10*, *ZNF664*, and *CCDC92*. Another region in CIM had a region at chr16:3545002-3795001 and intersected *NLRC3*, and included the significant markers for both iHS and nSL of rs13926 and rs1639150. For Tajima's *D* there were two regions of contiguous score depression that intersected obesity-associated loci, the first in NZM at chr16:67145002-67465001 intersected *KCTD19*. The second was in TON and covered *CCR3* at chr3:46035002-46295001.

Table 3.8: Obesity-associated loci that showed signs of possible selection in Polynesian populations.

Population	Gene	XP-EHH							Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS			
TON	<i>ABCA1</i>		1								
CIM	<i>ABO</i>							2			
NZM	<i>ABO</i>							1			
SAM	<i>ABO</i>							1			
TON	<i>ABO</i>							1			
CIM	<i>ACAN</i>							1			
NZM	<i>ACAN</i>							1			
CIM	<i>ADAMTS17</i>			1							
NZM	<i>ADAMTS17</i>			1							
SAM	<i>ADAMTS17</i>			1							
TON	<i>ADAMTS17</i>			1				1			
CIM	<i>ADAMTS9</i>	1				2		1	3		
NZM	<i>ADAMTS9</i>							1	3		
CIM	<i>ADCY9</i>	2	5	2	2			7	6		
NZM	<i>ADCY9</i>		5		1			2	1		
NZM	<i>AGBL4</i>							1		17	
NZM	<i>APOA5</i>							1			
TON	<i>APOA5</i>							1			
CIM	<i>ARL15</i>	1	4	3	6	1	5	2	2		5
NZM	<i>ARL15</i>	1	3		6		4	2			2
SAM	<i>ARL15</i>	1	3		6		5		1		5
TON	<i>ARL15</i>	1	3		6		5				3
SAM	<i>BDNF</i>							1	1		1
TON	<i>BDNF</i>							1	1		
CIM	<i>BTNL2</i>								7		
NZM	<i>BTNL2</i>								5		
SAM	<i>BTNL2</i>			2		3		1			
SAM	<i>CADM2</i>								2		
CIM	<i>CALCRL</i>	6	1								
NZM	<i>CALCRL</i>	4									
TON	<i>CALCRL</i>	3									
CIM	<i>CCDC92</i>	1							4	4	
CIM	<i>CCNLJL</i>				2		3	3			
NZM	<i>CCNLJL</i>						1	1			
CIM	<i>CCR2</i>	3	2	1			3				
NZM	<i>CCR2</i>	4	2	3			5				
SAM	<i>CCR2</i>			1							
TON	<i>CCR2</i>	1	2								
CIM	<i>CCR3</i>	1	2	1	6		5			2	

Table 3.8: Obesity-associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH							Fu & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS				
NZM	<i>CCR3</i>	2	4	1	6		5		1		4	
SAM	<i>CCR3</i>	3	3	1	6		5				4	
TON	<i>CCR3</i>	3	3	1	6		5			7	5	
NZM	<i>CDKAL1</i>							1				
CIM	<i>COL6A1</i>	7	2		3		2					
NZM	<i>COL6A1</i>	6	1		5							
SAM	<i>COL6A1</i>	6										
TON	<i>COL6A1</i>	7	2		6		4					
SAM	<i>CTSS</i>	3	5	5								
TON	<i>CTSS</i>	1	5	4								
CIM	<i>DGKG</i>						2					
NZM	<i>DGKG</i>				1		3					
TON	<i>DGKG</i>						3					
CIM	<i>DNAH10</i>	3							1	18	7	
NZM	<i>DNAH10</i>	1							4	4	3	2
SAM	<i>DNM3</i>	4		3								
TON	<i>DNM3</i>	3		2								
SAM	<i>EFEMP1</i>							1				
CIM	<i>ERBB4</i>	6	4		1		5	1	8		6	
NZM	<i>ERBB4</i>	7	4		4		5	1		7		2
SAM	<i>ERBB4</i>	7	4		3		5	1	1	7		7
TON	<i>ERBB4</i>	7	4		6		5			8	8	8
CIM	<i>EYA2</i>			1		1	1	1				
SAM	<i>EYA2</i>			2				2				
SAM	<i>FAM13A</i>	7	1		5							
TON	<i>FAM13A</i>	7	1		5		1					
CIM	<i>FCER1A</i>							3				
SAM	<i>FCER1A</i>							1				
TON	<i>FCER1A</i>							1				
SAM	<i>FER</i>	1			2				10		4	
TON	<i>FER</i>	5			4				2	2		
NZM	<i>FHIT</i>							4	1	5		
SAM	<i>FHIT</i>						4			6		
TON	<i>FHIT</i>	1	4	4		2	5	1	1	5		
SAM	<i>FNDC3B</i>					2						
TON	<i>FNDC3B</i>				1							
NZM	<i>FOXO3</i>							1	1			
SAM	<i>FOXO3</i>							1				
NZM	<i>FPGT-TNNI3K</i>	2					3					
TON	<i>FPGT-TNNI3K</i>	1					2	1	1			
NZM	<i>FTO</i>							1				
SAM	<i>GBE1</i>	2							1			
CIM	<i>GDF5</i>	4				2		1	2			
NZM	<i>GDF5</i>	2								3		
NZM	<i>GP2</i>			1			1					
TON	<i>GP2</i>							1				
SAM	<i>GPRC5B</i>							1	1			
TON	<i>GPRC5B</i>							1				
SAM	<i>GRID1</i>								1	2		
TON	<i>GRID1</i>								2	2		
SAM	<i>HSD17B12</i>							2			1	
NZM	<i>IQCK</i>			1								
SAM	<i>IQCK</i>							1				
TON	<i>ITGB6</i>							1	1			
CIM	<i>JAZF1</i>			1								
SAM	<i>JAZF1</i>		4	3	2	1	4					
TON	<i>JAZF1</i>	5	4	5	6	1	5					
TON	<i>KCNMA1</i>			2			4					
NZM	<i>KCTD15</i>								1			
CIM	<i>KREMEN1</i>								1			
SAM	<i>KREMEN1</i>								1			

Table 3.8: Obesity-associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH						Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS			
CIM	<i>LEKR1</i>	7		1						
NZM	<i>LEKR1</i>	4								
SAM	<i>LEKR1</i>	6		1					1	
TON	<i>LEKR1</i>	2								
CIM	<i>LEPR</i>	1			2					
SAM	<i>LEPR</i>	1			1	2				
TON	<i>LEPR</i>						1			
NZM	<i>LIN28B</i>	1						3		
NZM	<i>LINGO2</i>			1						
TON	<i>LINGO2</i>						1			
SAM	<i>LRP1B</i>				3	7	7			
TON	<i>LRP1B</i>				1	4	4			
CIM	<i>LY86</i>				1		1			
NZM	<i>LY86</i>		4	2			2			
TON	<i>LY86</i>						1			
SAM	<i>LYPLAL1</i>	1	1							
TON	<i>LYPLAL1</i>	2	1			2				
NZM	<i>MAP2K5</i>						1		3	
TON	<i>MAP2K5</i>							1	3	
CIM	<i>MSRA</i>	1	3			2	1	2		
SAM	<i>MSRA</i>			1						
TON	<i>MSRA</i>			1				1		
CIM	<i>MTIF3</i>							1		
NZM	<i>MTIF3</i>						2			
CIM	<i>NAV1</i>						1	1		
SAM	<i>NAV1</i>						1			
TON	<i>NCAM2</i>						1			
CIM	<i>NEGR1</i>	4							5	4
NZM	<i>NEGR1</i>	6			1				2	3
TON	<i>NEGR1</i>	6		1						
NZM	<i>NFE2L3</i>						1			
CIM	<i>NLRC3</i>	1	3		1				5	
NZM	<i>NLRC3</i>			1					5	
SAM	<i>NRXN3</i>	1							1	
TON	<i>NRXN3</i>	6				1		1	2	3
NZM	<i>OR10J1</i>				1					
SAM	<i>PARK2</i>	1				2		1	1	
TON	<i>PARK2</i>							1		
NZM	<i>PAX5</i>	3	1	5	3	3	5			
SAM	<i>PCSK1</i>							1		
CIM	<i>PCSK5</i>	1	3	1			2			
NZM	<i>PCSK5</i>	4	5	2			3		1	
SAM	<i>PCSK5</i>						1			
SAM	<i>PEPD</i>	7	2	4	6					
TON	<i>PEPD</i>	7	3	5	6	1				
NZM	<i>PGPEP1</i>					1				
TON	<i>PGPEP1</i>						1			
CIM	<i>PPP2R3A</i>				1					
NZM	<i>PPP2R3A</i>	4			5					
SAM	<i>PPP2R3A</i>	2			2				1	
TON	<i>PPP2R3A</i>	7			3		1			
CIM	<i>RABEP1</i>	3								
NZM	<i>RABEP1</i>	1						1	1	
SAM	<i>RABEP1</i>	6		1	1		2			
TON	<i>RABEP1</i>	7		2	1		2			
CIM	<i>RASA2</i>	1						2		
SAM	<i>RASA2</i>	1						2		
TON	<i>RASA2</i>	1								
TON	<i>RMST</i>					1				
NZM	<i>RPTOR</i>		2	3	3		5			
CIM	<i>SLC39A8</i>	5	1							

Table 3.8: Obesity-associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH						Fay & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS				
NZM	<i>SLC39A8</i>	4									
SAM	<i>SLC39A8</i>	6	1					1	4		
TON	<i>SLC39A8</i>							2			
CIM	<i>SMAD6</i>				1						
NZM	<i>SMAD6</i>						1				
TON	<i>SMAD6</i>				1						
CIM	<i>STXBP6</i>						1				
TON	<i>STXBP6</i>						1				
TON	<i>TCF7L2</i>			2	1				2		
CIM	<i>TNKS</i>				2						
NZM	<i>TNNI3K</i>	2					3				
TON	<i>TNNI3K</i>	1					2	1	1		
SAM	<i>TRIP11</i>							2			
TON	<i>TRIP11</i>							3			
TON	<i>USP37</i>					1					
CIM	<i>ZBTB38</i>	1							4		1
NZM	<i>ZBTB38</i>	1							2		
SAM	<i>ZBTB38</i>	4							1		
TON	<i>ZBTB38</i>	3							7		
CIM	<i>ZNF664</i>	1								1	2

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.1.3 Selection in Polynesian populations for genes associated with type 2 diabetes

Type 2 diabetes is a co-morbidity of hyperuricemia, and had 99 genes associated from GWAS in the GWAS catalog (see Table S7 for references), with one gene in common with urate-associated loci, and 16 genes in common with obesity. For the Polynesian populations there were 41 genes total, with 23 loci that were intersected by windows in the 1st percentile of a frequency-based statistic. There were 19 genes that had a marker with a significant value for iHS or nSL. The genes that showed evidence of possible selection in the Polynesian populations are shown in Table 3.9. There were five genes that showed evidence across all Polynesian populations, these were *ARL15*, *LPP*, *PTPRD*, *SND1*, and *THADA*. Four genes had evidence that was mostly in the Eastern Polynesian populations, these were *ADAMTS9*, *PSMD6*, *SSR1*, and *ZFAND6*. And the genes that mostly had evidence in the Western Polynesians were *BCL11A*, *JAZF1*, *KCNJ11*, and *PEPD*. As previously mentioned *ADAMTS9*, *ARL15*, *JAZF1*, and *PEPD* were also associated with obesity.

There were no regions of contiguous score depression for any of the SFS based intra-population statistics for any of the Polynesian populations.

Table 3.9: Type 2 diabetes-associated loci that showed signs of possible selection in Polynesian populations.

Population	Gene	XP-EHH							Fay & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS				
CIM	<i>ADAMTS9</i>	1				2		1	3			
NZM	<i>ADAMTS9</i>							1	3			
TON	<i>ADCY5</i>							1				
SAM	<i>AP3S2</i>		2									
TON	<i>AP3S2</i>		2									
CIM	<i>ARL15</i>	1	4	3	6	1	5	2	2		5	
NZM	<i>ARL15</i>	1	3		6		4	2			2	
SAM	<i>ARL15</i>	1	3		6		5		1		5	
TON	<i>ARL15</i>	1	3		6		5				3	
CIM	<i>ATP8B2</i>					1						
NZM	<i>ATP8B2</i>					1						
SAM	<i>BCL11A</i>		4		6	1	4		2	5	1	
TON	<i>BCL11A</i>		3		3		1		2			
SAM	<i>CAMK1D</i>					1						
TON	<i>CCDC85A</i>							1				
NZM	<i>CDC123</i>			1								
NZM	<i>CDKAL1</i>							1				
NZM	<i>CDKN2A</i>		2			1						
NZM	<i>CDKN2B</i>		2			1						
NZM	<i>FAM60A</i>			5								
NZM	<i>FTO</i>						1					
NZM	<i>GLIS3</i>						1					
TON	<i>GLIS3</i>					1		1				
CIM	<i>GLIS3</i>						1					
SAM	<i>GRK5</i>		1				1					
TON	<i>GRK5</i>		1				1					
NZM	<i>IDE</i>	1	3									
TON	<i>INS-IGF2</i>					1						
TON	<i>ITGB6</i>							1	1			
CIM	<i>JAZF1</i>			1								
SAM	<i>JAZF1</i>		4	3	2	1	4					
TON	<i>JAZF1</i>	5	4	5	6	1	5		1			
SAM	<i>KCNJ11</i>						1	1				
TON	<i>KCNJ11</i>		3			2	3					
NZM	<i>LAMA1</i>			1								
SAM	<i>LAMA1</i>					1						
CIM	<i>LPP</i>	7	2		6		5	3	1			
NZM	<i>LPP</i>	7	4		6	1	5	6	1			
SAM	<i>LPP</i>	7	2		5		4			1		
TON	<i>LPP</i>	7	3		5		4		1			
SAM	<i>PAX4</i>						1			1		
SAM	<i>PEPD</i>	7	2	4	6							
TON	<i>PEPD</i>	7	3	5	6	1						
TON	<i>PPP2R2C</i>					1						
CIM	<i>PSMD6</i>			3		2		1	1			
NZM	<i>PSMD6</i>							1	1			
CIM	<i>PTPRD</i>	6	3	5	6		5	14	4	3		

Table 3.9: Type 2 diabetes-associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH									
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H	Fu & Li's F
NZM	<i>PTPRD</i>	5	3	4	5	1	5	1	1	1	1
TON	<i>PTPRD</i>	7	3	4	2	1	5	5	4	1	3
SAM	<i>PTPRD</i>						3	3			
CIM	<i>RASGRP1</i>					1					
CIM	<i>RREB1</i>	5	3	5	6		1				3
NZM	<i>RREB1</i>	4	3	5	6				1	1	
TON	<i>RREB1</i>	3	1	1	2					2	
CIM	<i>SND1</i>			1		1				4	
NZM	<i>SND1</i>				1		1		7	4	4
SAM	<i>SND1</i>	5	3		5		5		1	1	
TON	<i>SND1</i>	2	3		2		5				
NZM	<i>SRR</i>							1			
CIM	<i>SSR1</i>	2	2	5	3			2	2		
NZM	<i>SSR1</i>	4	2	5	3			2	2		
SAM	<i>SSR1</i>			1							
TON	<i>SSR1</i>			3							
SAM	<i>ST6GAL1</i>						1	1			
TON	<i>TCF7L2</i>				2	1			2		
CIM	<i>THADA</i>	6	2		6		5		8	2	1
NZM	<i>THADA</i>	7	2		6		5		3		1
SAM	<i>THADA</i>	7	2		6		5		3		
TON	<i>THADA</i>	7	2		6		5		7	12	8
SAM	<i>TP53INP1</i>	7			1			2	2		
TON	<i>TP53INP1</i>	5			1						
CIM	<i>TSPAN8</i>							1			
CIM	<i>WFS1</i>	2									
CIM	<i>ZFAND6</i>							1	1		
NZM	<i>ZFAND6</i>							1	1		
CIM	<i>ZMZ1</i>			5			5				
NZM	<i>ZMZ1</i>			5			5				
SAM	<i>ZMZ1</i>			5			4				
TON	<i>ZMZ1</i>			5			1				

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.1.4 Selection in Polynesian populations for genes associated with kidney disease

There were 53 genes that were associated with kidney disease from the GWAS catalog (see Table S7 for references) with 11 genes having possible evidence of selection in Polynesian populations. Four genes (*DDX1*, *GP2*, *PHTF2*, and *WDR37*) had markers that were significant for iHS in the Polynesian

populations (Table 3.10). NZM had 3 markers (rs6966446, rs17158527, and rs75330800) in *PHTF2*, SAM had one marker in *DDX1*, and TON had one marker in each of *GP2* (rs7188098) and *WDR37* (rs10508203). None of these loci also had windows from the SFS-based statistics that were significant. NZM had a single marker (rs17158527) that was significant for nSL in *PHTF2*. SAM had a single significant marker (rs1001116) for nSL in *PRKAG2*, a urate locus mentioned previously. *WDR72* had two markers, rs7182198 and rs10220852, in both TON and SAM that were significant for nSL. There were no regions of contiguous score depression that intersected with kidney disease associated genes.

Table 3.10: Kidney disease-associated loci that showed signs of possible selection in Polynesian populations.

Population	Gene	XP-EHH										
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H	Fu & Li's F	Tajima's D
SAM	<i>ANXA9</i>		2	3	4							
TON	<i>ANXA9</i>		2	1	3							
CIM	<i>BNIPL</i>	1										
SAM	<i>BNIPL</i>	1								1	1	
SAM	<i>DDX1</i>							1				
SAM	<i>FAM63A</i>			2								
TON	<i>FAM63A</i>			1								
NZM	<i>GP2</i>		1			1						
TON	<i>GP2</i>							1				
NZM	<i>PHTF2</i>						3	1				
CIM	<i>PRKAG2</i>	1										
NZM	<i>PRKAG2</i>	1										
SAM	<i>PRKAG2</i>								1			
CIM	<i>SETDB1</i>			1						1		
SAM	<i>SETDB1</i>	2		3								
TON	<i>SETDB1</i>	1		3						1		
SAM	<i>SLC13A3</i>				1							
NZM	<i>WDR37</i>			1		1						
SAM	<i>WDR37</i>			1								
TON	<i>WDR37</i>			2			1					
SAM	<i>WDR72</i>	2							2			
TON	<i>WDR72</i>	4							2			

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.1.5 Selection in Polynesian populations for genes associated with metabolic syndrome

In the GWAS catalog, there were 22 genes associated with metabolic syndrome (see Table S7 for

references). In total, 9 loci had some evidence in the Polynesian populations of selection (Table 3.11). *APOA5*, also associated with obesity, had a significant marker (rs3135507) in both NZM and TON for iHS. SAM had a marker (rs12333979) that was significant for both iHS and nSL in *DGKB*. As previously mentioned, NZM had a significant marker for iHS in *FTO*, which was also associated with obesity and type 2 diabetes.

In the Eastern Polynesian populations there was a region of contiguous score depression for Tajima's *D* that intersected *EDC4* at chr16:67835002-68115001 for CIM, and chr16:67775002-68105001 for NZM. NZM also had a similar region for Fu and Li's *F*, chr16:67775002-68095001.

Table 3.11: Metabolic syndrome-associated genes that showed signs of possible selection in Polynesian populations

Population	Gene	XP-EHH							
		AFR	AMR	EAS	iHS	nSL	Fay & Wu's H	Fu & Li's F	Tajima's D
TON	<i>ABCA1</i>	1							
NZM	<i>APOA5</i>				1				
TON	<i>APOA5</i>				1				
CIM	<i>APOB</i>		1						
NZM	<i>APOB</i>	1					6		
NZM	<i>DGKB</i>			1					
SAM	<i>DGKB</i>				1	1			
TON	<i>EDC4</i>	2							
NZM	<i>FTO</i>				1				
CIM	<i>GALNT2</i>			1					
NZM	<i>LIPC</i>				1				
TON	<i>LIPC</i>				1				
CIM	<i>LPL</i>	2							
TON	<i>LPL</i>	2							

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.2 Selection in genes associated with other diseases possibly involving urate

3.3.4.2.1 Selection in Polynesian populations for genes associated with neurological disorders

Urate has associations with neurological disease, such as the potential as a biomarker for the progression of Parkinson's disease (Wen *et al.*, 2017). Loci that were associated with the neurological diseases of Parkinson's disease and Alzheimer's disease from the GWAS catalog did not intersect with any regions

of contiguous score depression of any SFS-based statistic. *LRRK2* had significant XP-EHH for all Polynesian populations with at least one population from each of the super populations. The Western Polynesian populations also had markers with significant iHS and nSL (Table 3.12). Another locus, *CERS6*, had each Polynesian population with significant XP-EHH markers with nearly all populations from the AFR, AMR, EUR super population groups and approximately half of the populations from both EAS and SAS super populations. A third locus that had significant XP-EHH markers for all Polynesian populations was *SLC2A13*. All Polynesian populations had significant XP-EHH markers with most populations from AMR and EAS super populations. NZM, CIM, and SAM also had significant markers with nearly all populations of the SAS and EUR. Only NZM had significant markers with the AFR populations. All Polynesian populations had significant markers for nSL at *SLC2A13*, and for iHS except for NZM. *VPS13C* had both Eastern Polynesian populations with two significant iHS markers (CIM: rs3784635 and rs12595158; NZM: rs3784635 and rs112236709), and windows from the 1st percentile for Fu and Li's *F* (6 windows), and Zeng's *E* (9 windows for NZM and 11 for CIM).

Table 3.12: Neurological disease associated loci that showed signs of possible selection in Polynesian populations.

Population	Gene	XP-EHH								Fay & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL				
CIM	<i>ABCA7</i>			1		2	1	1					
NZM	<i>ABCA7</i>					2		1					
CIM	<i>BIN1</i>					1							
NZM	<i>BIN1</i>					1							
SAM	<i>BIN1</i>							1					
CIM	<i>C12orf40</i>		3	4	3			1	3				
NZM	<i>C12orf40</i>	5	3	5	6			5					
SAM	<i>C12orf40</i>		1	4					2				
TON	<i>C12orf40</i>		2	4					2				
CIM	<i>CERS6</i>	7	2	1	5			1			1		
NZM	<i>CERS6</i>	6	3	2	6			1			1		
SAM	<i>CERS6</i>	7	4	5	6			3					
TON	<i>CERS6</i>	7	4	5	6			5	1	1	1		
CIM	<i>CR1</i>							1					
TON	<i>FAM126A</i>							1					
TON	<i>FRMD4A</i>							1					
CIM	<i>FRMD4A</i>								1				
NZM	<i>FRMD4A</i>								2				
SAM	<i>GBA</i>		2										
TON	<i>GPNMB</i>							1					
TON	<i>GPRIN3</i>									1	2		
SAM	<i>HLA-DQA1</i>			1		1							
CIM	<i>LRRK2</i>	1	3	4	5						1		
NZM	<i>LRRK2</i>	5	3	5	6			5					
SAM	<i>LRRK2</i>	6	3	5	6			1	1	2		3	
TON	<i>LRRK2</i>	1	3	4	5				3	3			
CIM	<i>MS4A6A</i>							1		2			
NZM	<i>MS4A6A</i>							1		1			

Table 3.12: Neurological disease associated loci that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH							Fay & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS				
SAM	<i>MS4A6A</i>							1	1			
CIM	<i>NDUFAF2</i>	7	1				5			5		
NZM	<i>NDUFAF2</i>	6					5					
SAM	<i>NDUFAF2</i>						4					
TON	<i>NDUFAF2</i>						4			8	6	
SAM	<i>OR9Q1</i>	3	3		5	1						
TON	<i>OR9Q1</i>		2						2			
NZM	<i>PICALM</i>		1		5		4					
CIM	<i>PKP2</i>	7	2				5	4	2			
NZM	<i>PKP2</i>	7	4	4	2		5	2	2			
SAM	<i>PKP2</i>	3	1				5	1				
TON	<i>PKP2</i>	2	1				2					
TON	<i>PRICKLE1</i>					1						
NZM	<i>SIPA1L2</i>							1	1			
CIM	<i>SLC2A13</i>		3	4	5		2	7	4		3	
NZM	<i>SLC2A13</i>	6	3	5	6		5		2			
SAM	<i>SLC2A13</i>	3	4	5			3	3	3			
TON	<i>SLC2A13</i>	2	4					3	2			
CIM	<i>SNCA</i>	2	3									
NZM	<i>STK39</i>	1	5									
SAM	<i>STK39</i>	1	2	5			4	8	4			
TON	<i>STK39</i>			5			2	7	4			
SAM	<i>TMEM163</i>							2	1			
TON	<i>TMEM163</i>							1	1			
CIM	<i>VPS13C</i>						2		5		11	
NZM	<i>VPS13C</i>						2		6		9	

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

3.3.4.2.2 Selection in Polynesian populations for genes associated with inflammatory and autoimmune disorders

The list of inflammatory and autoimmune disorders was from Zhang *et al.* (2013) Table 2, and included: inflammatory bowel disease, Crohn's disease, ulcerative colitis, celiac disease, systemic lupus erythematosus, rheumatoid arthritis, psoriasis, multiple sclerosis, type 1 diabetes, primary biliary cirrhosis, and vitiligo. Autoimmunity is caused by the immune system attacking self rather than foreign antigen and involves both the innate and adaptive immune response (Waldner, 2009). The genes involved with the immune response have been suggested as possible candidates of selection (Grossman *et al.*, 2013). Both gout and these inflammatory and autoimmune disorders have a common cause from

the activation of the immune system.

There were 981 genes associated with these inflammatory and autoimmune disorders, many of which were also in common with the genes associated with urate, obesity, type 2 diabetes, kidney disease, and metabolic syndrome, which is indicative of the immune response being part of all of these conditions. Some of the loci that had multiple statistics, across multiple Polynesian populations indicating possible selection were: *ARHGEF3*, *ARL15*, *ATM*, *BLK*, *CCR3*, *CNTNAP2*, *DSTYK*, *FAM167A*, *FOXP1*, *KCNB2*, *LPP*, *LRRK2*, *RREB1*, *SLC2A13*, *ALCO6A1*, *THADA*, and *XKR6* (Table S5).

All four Polynesian populations had significant XP-EHH values with populations from the AFR and AMR super populations at the *CNTNAP2* locus. The Western Polynesian populations also had significant XP-EHH values with EUR populations. There were significant markers for iHS and nSL for NZM and SAM. NZM had rs2620441 for both iHS and nSL, rs2249958 for iHS, and rs6952506 for nSL. SAM had rs2249958 and rs17170777 for both iHS and nSL, rs2620441 for iHS, and rs10255956 for nSL. All Polynesians populations also had windows in the 1st percentile for Fay and Wu's *H*, with NZM having the most with 12. All Polynesian populations also had windows in the 1st percentile for Tajima's *D* that intersected *CNTNAP2*.

CCR2 and *FCHSD2* had a higher number of significant results in the Eastern Polynesian populations, whereas *CDH23* and *JAZF1* had a higher number of significant results in the Western Polynesian populations. *CCR2* and *CCR3* were also associated with obesity, *ARL15* and *JAZF1* were associated with obesity and type 2 diabetes, *LPP* and *THADA* were associated type 2 diabetes, and *RREB1* was associated with urate and type 2 diabetes. Some of the genes had very high numbers of windows from the 1st percentile for Tajima's *D* intersect with them, such as *RAD51B* which was a contiguously depressed score region for SAM, and *XKR6*, which was a contiguously depressed score region for CIM, SAM and TON populations.

3.3.4.2.3 Selection in Polynesian populations for genes associated with malaria

Urate has a central role in malarial infection and triggering an immune response (Gallego-Delgado *et al.*, 2014). Given that immune response is a strong candidate to be under selective pressure, malaria associated genes were looked at to see if there were signals of selection. From the GWAS catalog there were two genes that had been associated with malaria at a genome-wide significance threshold (of three, see Table S7 for references), that also had evidence of selection with the Polynesian populations. The two genes that also had significant iHS scores with Polynesian populations were *ABO* and *ATP2B4*.

ABO was the only gene that had significant markers in all four Polynesian populations out of the malaria-associated genes, and this was only for iHS. All of the markers with significant iHS in *ABO* were in favour of the derived allele haplotype. CIM had two markers that were significant, the first, rs1053878, was also significant in TON. The second marker, rs55764262, was significant in CIM, NZM, and SAM, and was not in linkage disequilibrium (LD) with rs549446 (EAS R² = 0.003). Rs549446 was significant in both CHS and JPT. There were no windows from the frequency-based statistics that met the significance threshold. *ABO* has also been associated with obesity in GWAS (Comuzzie *et al.*, 2012). The marker rs8176741 had been associated with severe malaria (Timmann *et al.*, 2012),

Table 3.13: Linkage disequilibrium in EAS populations (R^2) between markers in *DDC*.

Rs number	rs2060762	rs11238131	rs4580999	rs3807558	rs2329371	rs1451375	rs1966839
rs2060762	1.000	0.600	0.449	0.688	0.723	0.492	0.490
rs11238131	0.600	1.000	0.585	0.680	0.531	0.319	0.316
rs4580999	0.449	0.585	1.000	0.674	0.510	0.611	0.607
rs3807558	0.688	0.680	0.674	1.000	0.744	0.497	0.495
rs2329371	0.723	0.531	0.510	0.744	1.000	0.707	0.698
rs1451375	0.492	0.319	0.611	0.497	0.707	1.000	0.988
rs1966839	0.490	0.316	0.607	0.495	0.698	0.988	1.000

LD was calculated in the EAS population of the 1000 Genomes Project using LDlink (Machiela and Chanock, 2015), due to not all markers being present on the CoreExome SNP array.

however, this marker had a significant iHS in Finnish in Finland (FIN) but no other populations. The other SNPs that showed significance for iHS were not in LD (EAS $R^2 = 0.002$) with rs55764262, for which CIM, NZM and SAM all had a significant result. Both CIM and TON had a different marker (rs1053878) that was significant for iHS in *ABO*.

Both SAM and NZM had SNPs that were significant for iHS in *ATP2B4*, SAM had rs10494845, and NZM had rs142206068, both were in favour of the derived allele haplotype. The malaria-associated variant was rs10900585 (Timmann *et al.*, 2012). All three variants were not in LD with each other (EAS $R^2 = 0.00$). The LD was calculated using the full 1000 Genomes Project (1KGP) marker list for EAS because rs10900585 was not on the CoreExome SNP array.

DDC encodes dopa decarboxylase and has been nominally associated at a genome-wide threshold with malaria (Jallow *et al.*, 2009). *DDC* has also been associated with BMI (Locke *et al.*, 2015a). NZM had significant markers for XP-EHH for nearly all other populations, CIM had fewer populations, but across all super population other than SAS (Table 3.14). The NZM had multiple markers within *DDC* for nSL. The markers were rs2060762, rs3807558, rs2329371, and rs1966839. The same markers in addition with rs11238131 and rs4580999 were significant for iHS. A LD matrix with marker positions is shown in Table 3.13. The marker rs1966839 (reference/derived allele = T, alternative/ancestral allele = C) had scores from iHS and nSL in favour of the ancestral allele and was in high LD ($R^2 = 0.988$) with the malaria-associated rs1451375 (reference/ancestral allele = C, alternative/derived = A) where the reference and alternative alleles correspond between the SNPs, and both are intronic variants. Under a dominant model the A allele of rs1451375 is protective of severe malaria (OR = 0.75, CI = 0.66-0.85, P = 6 x 10-6, Jallow *et al.* (2009)) and corresponded with the ancestral allele of rs1966839.

Table 3.14: Markers with significant XP-EHH in Polynesian populations at *DDC* on chromosome 7.

3.4 Chapter Discussion

3.4.1 Identifying regions under selection in Polynesian populations

One of the primary objectives of this chapter was to investigate regions that were under selection in Polynesian populations. From the intra-population statistics used, there were between 866 and 974 genes identified in the Polynesian populations that had both intra-population haplotypic and SFS statistics that were in the 1st percentile. The regions with possible evidence of selection showed that compared to the other super populations, the Polynesian populations had the fewest genes (111) that only had evidence within the Polynesian super population, whereas the African populations had the largest number of genes (457) that only had evidence in the African super population (Figure 3.3). The Polynesian populations had similar distributions of the SFS-based statistics (explored more in Chapter 4).

Pathway analysis on the regions with possible selection in the Polynesian populations revealed that many of the genes that had evidence were involved with various forms of cell signalling, with many related to calcium channels, especially in the Eastern Polynesian populations (sections 3.3.3.2.1 and 3.3.3.3.1). The majority of the pathways that had significant enrichment were from the genes that had markers significant for nSL. There was similarity between the pathways of the Eastern Polynesian populations, with many of the pathways in common being cardiac related. These pathways, such as adrenergic signalling in cardio myocytes, calcium signalling pathway, hypertrophic cardiomyopathy, and vascular smooth muscle contraction all had calcium channels as the main set of genes that were selected (sections 3.3.3.2.1 and 3.3.3.3.1). The genes in common with these pathways were *CACNA1D* (Calcium Voltage-Gated Channel Subunit Alpha1 D) which is associated with blood pressure (Lu *et al.*, 2015), *CACNA2D2* (Calcium Voltage-Gated Channel Auxiliary Subunit Alpha2delta 2) - also associated with blood pressure (Warren *et al.*, 2017), *CACNA2D3* (Calcium Voltage-Gated Channel Auxiliary Subunit Alpha2delta 3) - has an association with gout (Lai *et al.*, 2012) but has not been replicated in other studies, and *SLC8A1* (Solute Carrier Family 8 Member A1) which is associated with cardiac electrical activity time intervals (Arking *et al.*, 2014). There was an association with gout in the Polynesian GWAS performed in chapter 5 with *CACNA2D3* (rs6793459, OR 0.79 95% CI [0.68-0.90], P = 7x10⁻⁴), however it was not genome-wide significant, and this SNP was not in LD with any of the significant iHS or nSL markers.

In the Western Polynesian populations, the only pathway that was in common between SAM and TON was ABC transporters. There was only one gene in the ABC transporters pathway that had an association with urate and gout from the GWAS catalog. The gene was *ABCG2*, which has a strong effect for gout in Western Polynesian populations (Phipps-Green *et al.*, 2010) and also had haplotypic evidence in TON with iHS. One of the other genes in this pathway was *ABCC4* which recently had Western Polynesian specific variants, such as rs972711951, identified and associated with gout (Tanner *et al.*, 2017). There was evidence of selection at *ABCC4* in CIM, NZM, and SAM (Table S3).

The gene *CNTN4*, encoding Contactin 4, had the highest number of significant markers in CIM and third most in NZM for iHS. This particular locus has not been associated with urate at a genome-wide

significance level, however there is evidence to suggest there may be an association (Chittoor *et al.*, 2016), but this was not observed with gout in either of the European or Polynesian GWAS analyses performed in chapter 5 (section 5.3.2). There has also been prior selective evidence for this locus in the Biaka Pygmy and Bantu populations, but not in more closely ancestrally related populations of the Polynesian populations (Pickrell *et al.*, 2009).

When looking at genes that have been associated with urate, gout and related co-morbidities, there is no apparent difference in the number of genes showing signs of selection in Polynesian populations, when compared to the other populations (Table 3.6). However, there were differences between the Eastern and Western Polynesian populations for the iHS and nSL results, supporting the differences in ancestral background. There was also evidence of selection in loci that were associated with malaria and with type 2 diabetes and obesity, and is thus suggestive of a role for infectious disease applying selective pressure.

3.4.2 Selection of urate associated genes

The second objective was to investigate regions associated with urate, gout and related co-morbidities for evidence of selection in the Polynesian populations. There was evidence for selection in urate associated loci. However, the evidence did not appear to be specific for urate itself, but for loci that are involved in more general metabolic pathways. The main effect loci identified through GWAS, *SLC2A9* and *ABCG2*, had limited evidence of positive selection. *ABCG2* had a significant marker (rs2622626) with iHS and nSL and only in TON from the Polynesian populations, the only other populations were GIH and ITU from the SAS super population. This marker showing significance however might be due to the gout prevalence of the TON sample population which was 54%. The same result is not seen in closest population with the most similar ancestral background, SAM, where the prevalence of gout in the sample population was much lower at 12%. Although rs2622626 has not previously been associated with gout in East Asian populations (Zhang *et al.*, 2016), nor did it have an association in the Polynesian gout GWAS analysis in chapter 5 ($P > 0.1$).

Only *RREB1* and *IGFR1* had haplotypic evidence with XP-EHH across multiple other populations. The other gene of note that did not have haplotypic evidence, but did have multiple windows in the 1st percentile across multiple statistics and all Polynesian populations was *BCAS3*. *BCAS3* had been previously identified as selected in Grossman *et al.* (2013). *RREB1* was also associated with obesity and type 2 diabetes traits. Zhang *et al.* (2013) had reported a variant nearby to *RREB1* (rs675209) as having evidence of selection but this particular marker was absent from the CoreExome SNP array.

There was a similar situation with obesity, type 2 diabetes, kidney disease, and metabolic syndrome, where evidence of possible selection was found for multiple loci. The number of genes that had possible evidence of selection was similar in the Polynesian populations to that of the other populations. An interesting point however, was that the Western Polynesian populations had the fewest genes associated with type 2 diabetes identified with iHS or nSL than all the other populations, despite Polynesian populations having a high prevalence of type 2 diabetes (Winnard *et al.*, 2013).

Other complex diseases in which urate is implicated, such as the neurological, and, inflammatory and

autoimmune diseases, also showed multiple genes that displayed evidence of selection. One of the ‘themes’ that appeared was that genes were often associated with multiple diseases or gene lists, and often involved with the immune system.

3.4.3 Replication of candidate thrifty-genes

There was no evidence of selection in either of the previously identified loci of *PPARGC1A* and *CREBRF* in Polynesian populations. The CoreExome SNP array did not have rs8192678 which was the SNP reported in Myles *et al.* (2011) (and not replicated in Cadzow *et al.* (2016)) as being a thrifty genotype, but it did have rs1873532 which was in high LD (EAS R² = 0.976). The lack of evidence was consistent with the results of Cadzow *et al.* (2016) for *PPARGC1A* where there was no indication of selection from frequency- or haplotype-based methods.

No selection signal was observed at *CREBRF* in any of the Polynesian populations that were tested. One reason for this is that the CoreExome SNP array did not have the specific variants (rs12513649 and rs373863828) reported in Minster *et al.* (2016). Downstream from *CREBRF* there was a consistent signal with what had been reported. This was only found in the Western Polynesian populations and the populations of the EAS super population, so suggests that had the dataset contained the variants then detection would have been possible. However, without the two markers reported in Minster *et al.* (2016) it was unable to be determined if *CREBRF* had evidence of selection from the CoreExome data.

The pathway analysis of genes that were in the extremes for the different selection and neutrality statistics showed that pathways involved with cell signalling (and more general in nature) were the main results for the Polynesian populations, rather than specific metabolic pathways such as the urate pathway. If selection had been acting on the pathways involved with urate, gout and related conditions it would have been expected to see enrichment in genes involved with metabolic pathways. The genes that indicated possible selection, even for obesity and type 2 diabetes, which are traditionally thought of as being the thrifty phenotype, were genes that were largely either neurological or involved in the immune system. NZM did have significant results for two diabetes related pathways, “Type II diabetes mellitus_Homo sapiens_hsa04930” and “Insulin secretion_Homo sapiens_hsa04911”, however, only up to 10% of the pathways had genes that met the significance thresholds. Of the genes that were part of those pathways, up to 50% were genes that encoded subunits of voltage-gated calcium channels. While some type 2 diabetes-associated individual loci show some signs of selection, it appears as if widespread selection is not the case. Ayub *et al.* (2014) had a similar finding with respect to type 2 diabetes, where there was no support for the thrifty-gene hypothesis, testing across 65 type 2 diabetes-associated loci.

As discussed in Gosling *et al.* (2014), the thrifty-gene hypothesis in the context of the Pacific, where starvation during long voyages selected for thrifty-genes, does not fit with the anthropological history of Polynesian people, where they undertook planned voyages which settled the Pacific in a speedy manner. From the analysis results in this chapter, the thrifty-gene hypothesis appears to have none-to-minimal support from the loci that exhibit signals of possible selection.

3.4.4 Selection in malaria associated loci in Polynesian populations

Malaria is endemic in the southwest Pacific, extending out to the east as far as Vanuatu, with evidence of genetic advantage against malaria of Austronesian speaking populations in Near Oceania (Clark and Kelly, 1993). Malaria is absent, and always has been further west into Polynesia, Polynesian ancestors would have passed through Near Oceania prior to settlement of the Pacific (Clark and Kelly, 1993). In Polynesia, even though malaria is absent, high frequencies of α -thalassaemia mutations (protective against malaria) have been identified in populations not exposed to malaria (Hill *et al.*, 1985). Other malaria protective mutations, such as glucose-6-phosphate dehydrogenase deficiency and β -thalassaemia, are present in Pacific populations in higher frequencies, though generally in island groups where malaria remains extant (Flint *et al.*, 1986; Cappellini and Fiorelli, 2008).

Response to infectious challenge is a key mechanism that has been thought of as a selective pressure. Malarial infection provided a possible challenge to which urate may have been subjected to selective pressure, and as such provided a viable candidate due to the involvement of urate with malarial resistance (Gallego-Delgado *et al.*, 2014). This involves precipitated urate being released from ruptured infected erythrocytes, triggering an immune response (Orengo *et al.*, 2009; Gallego-Delgado *et al.*, 2014). Three loci that were associated with malaria and identified from the GWAS catalog were investigated for having signals of selection. Two genes that had evidence of selection in the Polynesian populations had also been identified as being associated with obesity and type 2 diabetes. *ABO* and *ATP2B4* did not have strong evidence, and it came from only iHS.

The blood antigen *ABO* has a higher susceptibility to malaria in the A, B, or AB antigen groups, than O (Zerihun *et al.*, 2011). This has led to the formulation of a potential mechanism of cytoadherence through infected red blood cells expressing PfEMP-1 and binding to the A or B antigens on other red blood cells, enabling further infection (Cserti and Dzik, 2015). There is an increasing gradient of A-type antigen from west to east in the Pacific, with populations in Melanesia having some of the lowest frequencies, whereas, within Polynesia, Western Polynesian populations have lower frequencies than Eastern Polynesian populations (Simmons, 1962). This could suggest that areas with endemic malaria have a maintained selective pressure against the A-type antigen.

ATP2B4 encodes a plasma membrane calcium transporter, another calcium related gene with evidence of selection in NZM, also sharing similarity with the many of the genes in the pathways from the pathway enrichment, in that they too were calcium channels.

DDC on the other hand, which had a nominally genome-wide significant association with malaria, had comparatively strong haplotypic evidence in the Eastern Polynesian populations. NZM had significant XP-EHH results with every other non-Polynesian population and had six significant markers for iHS, three of which were in the top 400 most extreme scores. Rs1966839 in *DDC* was in near perfect LD with the malaria associated intronic variant rs1451375 (Jallow *et al.*, 2009) and the protective allele corresponded with the allele displaying a possible signal of selection.

From the obesity GWAS associated gene list, there was *ADCY9*, which had a very strong signal for iHS for 7 markers in CIM and 2 markers in NZM, that were the only populations with significant iHS values. The variant rs10775349 in *ADCY9* had been associated with malaria in a candidate gene

approach but not in GWAS (Maiga *et al.*, 2013).

3.4.5 Study limitations

There were some technical limitations of the study that influenced the results obtained. The use of a SNP array meant that there was an ascertainment bias built into the markers observed (Nielsen *et al.*, 2007). The data regarding the statistic distributions can be found in chapter 4 (sections 4.3.2.1.1, 4.3.2.2.1, 4.3.2.3.1, and 4.3.2.4.1). The quality control process for the CoreExome SNP array also has a bias against singletons and low minor allele frequency SNPs (Guo *et al.*, 2014). The effect of this was seen in the right shifting of the distributions of the intra-population frequency-based statistics of Tajima's D , Fu and Li's F , and Zeng's E , all of which make use of low frequency or singleton variants in their calculations (Tables 4.3, 4.4, 4.5, and 4.6). This shift was not observed in Fay and Wu's H which compared the ratio of high frequency derived and ancestral allele variants with the intermediate frequency variants. The shift was seen in Tajima's D , however with the African populations being shifted to the right, such that the first percentile for many was above zero. The intermediate and high frequency variants were over represented compared to the actual situation which will also have had an impact on the haplotypic statistics due to haplotypes that have low-frequency variants being altered. Further discussion on the distributions can be found in chapter 4, section 4.4.3. The effect of using SNP array data and not sequence data, differing markers, marker density, and differing window sizes all contributed to the up-to 65% replication of previously identified regions (section 3.3.2. For the frequency-based statistics there was also only a FDR less than 10% in the Polynesian populations for the 1st percentile of Tajima's D and Fay and Wu's H . By focussing on regions that had multiple windows or multiple statistics the chance of it being a false positive should be reduced.

Another technical limitation that affected the haplotypes was phasing. The phasing of the haplotypes for the Polynesian populations made use of the 1000 Genomes Project reference haplotype panel for probabilistic phasing and could not make use of trios. This panel does not contain Polynesian-specific haplotypes and so the phasing would be influenced by this. While the majority of common haplotypes will be represented in this panel, haplotypes that are Polynesian-specific are not, and this introduces incorrect haplotypes. The phasing uncertainty was unable to be taken into account for any of the methods, this will influence the haplotypic methods of selection.

With the adaptation of the selection and neutrality tests into genome-wide usage, specifically the sliding window approach, there was the need to balance the number of segregating sites with the window size (Pybus *et al.*, 2014). This balance is entirely influenced by SNP density with whole genome sequence data offering 78 million bi-allelic SNPs in the 1000 Genomes Project Phase 3. However, the CoreExome SNP array data had lower density, with a mean of one SNP per 9142 bases. This means that window size needs to increase for SNP array data in order to have the same number of segregating sites. However, as window size increases it also reduces the resolution of the genome and with it the ability to associate a window with particular genomic features.

There was also a higher likelihood of a longer gene being reported as significant due to more windows overlapping the genic region. Not only that, but because the window size after the statistics were

calculated was trimmed to the central 10 kbp, this meant that SNPs in the calculation window might not be near the actual window used, that is, up to -45 kb to +45 kb away from the edge of the defined window used in the worst case. This would lead to some difficulty in definitive location. The flip-side to this was that it allows for upstream and downstream effects of a SNP to be incorporated. There was a trade-off between SNP density and window size. There were 305,214 bi-allelic SNP markers used in this analysis from the CoreExome SNP array. This was less than the other projects such as the HapMap project, and considerably less than the sequence data that the 1000 Genomes Project used. The marker density may have been able to be improved though the imputation of the missing markers using a haplotype reference panel. Ideally this analysis would have had a higher marker density, however it was decided that it was best to remain unimputed due to both the haplotype uncertainty at the phasing step, and the Polynesian populations not being represented in the haplotype reference panel used for phasing and imputation. An example of why having Polynesian haplotypes represented is important, is the finding of Minster *et al.* (2016), where the variant (rs373863828) had no appreciable allele frequency outside of Polynesia. Implications from the marker density used are the number of SNPs per window was lower than in other studies and specifically when looking at the degradation of haplotypes there is the possibility for early truncation due to lack of data.

The method itself of using the empirical distribution and selecting the extremes for a statistic to determine the region under selection does have precedence (Voight *et al.*, 2006; Hider *et al.*, 2013; Jonnalagadda *et al.*, 2017), however, this does assume that selection had occurred. The underlying null distribution of the genome is also unknown. It does need to be acknowledged that using the extremes of the distribution is likely to lead to many false positives (Teshima *et al.*, 2006). This was somewhat mitigated in this study by using the permutations of the actual data to calculate the empirical FDR. Having internal replication within super populations also helps to reduce the FDR, for the Polynesian populations this consisted of the Eastern and Western Polynesia population pairings.

Allele frequency methods, and also the haplotypic methods, may get confounded by the gout prevalences of the Polynesian sample sets, especially the CIM and TON sets, where the prevalence of gout was up to five-fold higher than the general population prevalence for those populations. Loci that showed evidence in multiple populations that did not have this enrichment for the disease trait would be expected to have more credibility. This is important to consider for the co-morbidities of gout too, given the commonalities in associated genes.

One of the limitations of using the GWAS catalog to create gene lists for conditions is that there is a heavy over representation of genes that are associated to disease in Europeans due to the overall bias to GWAS research being conducted in European populations (Haga, 2010; Popejoy and Fullerton, 2016). This can mean that population-specific associations are missed (Weissglas-Volkov *et al.*, 2013; Minster *et al.*, 2016). Genes that are associated with a condition can also be missed if they do not reach genome-wide significance in a GWAS but in a candidate gene approach are significant for instance *ADCY9* being nominally significant for BMI at a genome-wide level but only significant for malaria in a candidate gene approach.

3.4.6 Conclusion

The first objective of this chapter was to investigate regions of the genome that had evidence of selection in Polynesian populations. The main result regarding this objective was that the pathways that were enriched by genes that displayed signatures of selection in Polynesian populations from the genome-wide analysis were dominated by metabolic pathways. Of these pathways, calcium transport and signalling was a central theme for the genes that showed evidence of possible positive selection in Polynesian populations.

The second objective of the chapter was to investigate regions that had a previous association with urate, gout, and associated co-morbidities. The results regarding this objective showed there was indeed evidence indicative of selection in urate-associated loci in the Polynesian populations. The evidence however was not specific to urate. This was shown by the fact that there was little signal of selection appearing at the urate transporters themselves, but at the loci that were part of more general pathways. There was also evidence of selection at other loci that were associated with the co-morbidities, but again the signals were in genes that were for general cell functions, rather than specific for the particular traits. The reasons for why there would be selection still remain unclear but response to pathogenic challenge such as malaria in the case of urate is a potential mechanism.

The analysis had limiting factors with the main data using markers that originated from a SNP array, and the ascertainment bias this introduced to the statistics, especially for those dependent on the low-frequency variants. A second limiting factor was the use of empirical thresholds which will increase the number false-positives. Both of these factors should temper the conclusion of selection or lack thereof, at the loci investigated.

Chapter 4

Clustering of Selection Statistics

The previous chapter investigated regions of the genome in Polynesian populations that displayed ‘signatures of selection’, this chapter investigates the hypothesis that populations with shared ancestry will display similar signals of selection. To test this, I will be clustering the various selection statistics and then comparing the clusters to current population migratory history. This chapter contains data that could be thought of as a precursor to chapter 3, however it incorporates the results from that chapter 3 and so has been presented after.

4.1 Introduction

4.1.1 Signatures of selection

“Signatures of selection” in a population can be identified in regions of the genome that exhibit a reduction in genetic variability (Maynard Smith and Haigh, 1974; Kaplan *et al.*, 1989; McVean, 2007). This reduction in genetic variation can arise when the phenotype of a neutral benefit allele experiences a favourable change in environmental conditions, or a new allele arises conveying a selective advantage (Hermisson and Pennings, 2005). This results in an increased frequency of both the allele, and linked sites, within a population (Maynard Smith and Haigh, 1974). Genome-wide scans for signatures of selection have shown that geographically similar populations share similar genetic signatures (Coop *et al.*, 2009; Pickrell *et al.*, 2009). Geography can inhibit migration, leading to local adaptations affecting the allele frequency of local populations, leaving the non-local population allele frequency unaffected (Coop *et al.*, 2009). New Zealand Polynesians are relatively geographically isolated, with a recent settlement history, meaning that any signatures of selection found, and not shared with other populations are likely to be a recent and localised adaptation.

4.1.2 Settlement of Polynesia

The history of Polynesian migration starts with migration out of Africa (50-100 kya, Nielsen *et al.* (2017)), tracking up to the Levant and Arabian Peninsula (45-55 kya) and splitting into South Asia, Indonesia, and Australia (50kya, Kivisild *et al.* (1999); Quintana-Murci *et al.* (1999)). Concurrently the migration to Europe was occurring (45 kya). From South Asia the migration continued with the peopling of East Asia (20kya, Groucutt *et al.* (2015)). The Pacific was settled in two events, the first was the settlement of Near Oceania (40 kya), followed by the settlement of Remote Oceania, reaching Western Polynesia (Samoa and Tonga) during the Lapita expansion (3 kya, (Matisoo-Smith, 2015; Skoglund *et al.*, 2016)). Eastern Polynesia (including the Cook Islands) was then settled (1–1.2 kya, Wilmshurst *et al.* (2011)). The final region of the Polynesian triangle to be settled was that of New Zealand (NZ) by NZ Māori, occurring 800 ya (Duggan and Stoneking, 2014; Matisoo-Smith, 2015). It is from this migration history that modern Polynesian populations come to have a shared common ancestry with modern East Asian populations.

For largely economic reasons, during the 1950's and 60's there was a rural to urban migration of people in the Pacific. As part of this, many individuals and families settled in New Zealand from Samoa, Tonga, and the Cook Islands in search of work, and now the populations in New Zealand largely outnumber the populations of their islands of origin (Matisoo-Smith, 2012). Within NZ, the Polynesian populations are mostly focused within the Auckland region (Barcham *et al.*, 2009).

4.1.3 Health disparities in Polynesian populations

New Zealand Māori and Pacific Islanders in New Zealand over the past decade have been exhibiting a decrease in overall health (New Zealand Ministry of Health, 2016). Fruit and vegetable intake, and physical activity has also been reported as decreasing. All the while there has been an increase in obesity, which has an adjusted rate ratio of 1.7 for Māori versus non-Māori, and 2.4 for Pacific versus non-Pacific (New Zealand Ministry of Health, 2016). The rate of increase in body mass index (BMI) for Pacific Islanders was five times that of the rest of the world, and fasting plasma glucose was three to four times higher over the 1980-2008 period (Hawley and McGarvey, 2015). The health disparities are captured in the New Zealand statistics where Māori have a mortality rate of 1.8 that of non-Māori, with the rates of obesity and diabetes of particular note, with the Māori standardised rate of 40.6 per 100,000, five times higher than non-Māori at 8.1 per 100,000 (New Zealand Ministry of Health, 2012).

New Zealand Polynesians have inherent elevated serum urate levels and higher prevalence of gout (Winnard *et al.*, 2012, 2013) and there are genetic variants in Polynesian populations associated with increased risk of gout and elevated serum urate (Phipps-Green *et al.*, 2010, 2016). Serum urate has been associated with metabolic disorders (Choi *et al.*, 2007; Choi and Ford, 2007), and genes involved with complex diseases such as type 2 diabetes, gout, and other metabolic related disorders, have shown evidence for selection Hancock *et al.* (2008); Pickrell *et al.* (2009); Zhang *et al.* (2013)].

It is hypothesised that elevated serum urate may have undergone positive selection in Polynesian populations due to some of the beneficial properties (Gosling *et al.*, 2014), such as its role as a powerful

anti-oxidant (Ames *et al.*, 1981), or as an adjuvant for the innate immune system (Opitz *et al.*, 2009). Serum urate has also been suggested to have a protective neurological effect against dementia, and provides improved cognitive function with age (Euser *et al.*, 2009). Characterisation of selection within Polynesian populations has been largely limited to population differentiation due to population sample sizes (Kimura *et al.*, 2008; Mallick *et al.*, 2016).

Here I investigate the shared ancestry for the genetic selection of serum urate and related co-morbidities in Polynesian populations and attempt to determine whether there is a selection signature shared among populations.

4.1.4 Objectives

The objectives of this chapter are to:

- Determine if neutrality and selection tests can be used to group populations.
- Investigate the shared selective histories of genetic loci associated with gout, urate, type 2 diabetes, obesity, kidney disease, and metabolic syndrome.

4.2 Methods

4.2.1 Data

The dataset used in this chapter is the selection dataset consisting of 31 population groups, split into six super populations, created from marker matching the Polynesian populations genotyped on the CoreExome single nucleotide polymorphism (SNP) array, with the 1000 Genomes Project (1KGP) Phase 3 dataset (section 2.2.3).

4.2.2 Principal component analysis

Principal component analysis was performed on the independent genetic markers from the selection dataset (for more detail refer section 2.1.4). The first 10 components were calculated and plotted.

4.2.3 Admixture analysis

Admixture analysis was performed on the independent genetic markers for the selection dataset. Admixture proportions were calculated using all of the populations in the selection dataset and cross-validation was used to identify the K value (number of theoretical ancestral populations) with the lowest error for values of K from 1 to 15. Proportions were also calculated by first removing the American Super Population (AMR) and Polynesian Super Population (POL) populations, calculating the K value that provided the lowest error, through cross-validation, and then projecting these ancestral

populations onto the AMR and POL populations, again for values of K from 1 to 15. For further information refer to section 2.1.5.

4.2.4 Frequency spectrum

Tajima's D , Fay and Wu's H , Fu and Li's F , Zeng's E , and F_{ST} were all calculated using PopGenome v2.2.3 (Pfeifer *et al.*, 2014) for each chromosome. Aside from F_{ST} , these statistics were also calculated as per section 2.1.2.1 to generate non-overlapping windows of 10 kb. For each selection test statistic, windows were ordered by test statistic value, and the 1st and 99th percentiles of the distribution were calculated. For the windowed approach, hierarchical clustering was then performed on a binary matrix derived from the presence or absence of the windows in the 1st or 99th percentiles, across all populations. Hierarchical clustering using the value calculated for each whole chromosome (i.e., treating the start and end of the chromosome as the window boundary) used the statistic value itself to create the distance matrix. The distance measure used was Euclidean distance and the linkage criteria was "complete linkage" for both sets of clustering. Accuracy of the clustering was assessed by descending the dendrogram until six groups, one for each super population, could be made, and then calculating the maximum proportion of the individual populations of the same super population assigned to the same cluster. Exclusivity of each cluster was calculated using the proportion of populations from a given super population compared to the total number of individual populations assigned to that cluster. For example, if a cluster consisted of eight populations, two from super population A and six from super population B, then the exclusivity of the cluster would be 0.25 for super population A, and 0.75 for super population B. F_{ST} was calculated pair-wise for all populations for each chromosome.

4.2.5 Extended haplotype homozygosity

For all SNPs, integrated haplotype homozygosity score (iHS) and number of segregating sites by length (nSL) were calculated and normalised using selscan v1.1.0b (Szpiech and Hernandez, 2014) as part of the selectionTools 1.1 (Cadzow *et al.*, 2014) pipeline, as described in section 2.1.2.2. Markers with an absolute iHS or nSL value greater than 2.6 met the significance threshold, and were clustered into genomic regions using the DBSCAN package v1.1.1 (Hahsler and Piekenbrock, 2017) in R, where nearby SNPs were assigned the same group identifier. The search radius used was 200 kb and the minimum number of points for a cluster was one. Cluster regions were created by taking both the minimum and maximum position for each group identifier, population and chromosome. The distance matrix was the result of calculating $(A \cap B) / A$ and $(A \cap B) / B$, where $A \cap B$ was the sum of the intersection of clustered regions between population A and population B, and A and B was the sum of the clustered regions for population A and population B respectively. Hierarchical clustering was then performed, using complete linkage for the linkage criteria to create two sets of clusters (for an example see Figure 4.15).

4.2.6 Pathway enrichment analysis

Pathway enrichment analysis was performed by taking genes that intersected the clustered iHS and nSL regions, inputting them into Enrichr¹ (Chen *et al.*, 2013; Kuleshov *et al.*, 2016) and exporting the KEGG 2016 pathway table. This differs from the pathway analysis in chapter 3 in that it was performed on genes from clustered regions of significant iHS or nSL SNPs, rather than just the locus a SNP was from.

4.2.7 Disease-associated genes

Disease associated markers and their annotated genes were downloaded from the genome-wide association study (GWAS) catalog² (MacArthur *et al.*, 2017) and filtered for association (chi-squared P < 5x10⁻⁸). Gene lists were then created for each of the following diseases or traits; urate and gout, type 2 diabetes, obesity, kidney disease, and metabolic syndrome by filtering the study disease trait on keywords (see section 2.1.3 for details). Table S7 shows source disease trait and reference used to create categories of gene lists for gout/urate, obesity, type 2 diabetes, kidney disease, and metabolic syndrome. Selection test statistics values for each population were centred using the population median. The gene lists from these categories were used in section 4.3.4 for hierarchical clustering using the selection test statistic values for windows that intersected these gene regions. Complete linkage and Euclidean distance were used for the clustering.

4.3 Results

4.3.1 Frequency spectrum

4.3.1.1 PCA analysis

In order to confirm that individuals from a population were genetically similar principal component analysis (PCA) was performed. It was expected that there would be higher genetic variation between populations, than between individuals of a population. It was also used to confirm if the differences between populations was likely influenced by difference in genotyping platform. The genetic variation captured by PCA was plotted to visualise the relationship of individuals to one another (Figure 4.1). Starting from principal component (PC) 1, each subsequent PC captured less of the variation than the previous. Principal components for the merged 1KGP data and the New Zealand samples showed in the first two components (which capture the most variation) that the African Super Population (AFR), European Super Population (EUR), and South Asian Super Population (SAS) super populations all formed separate groupings. PC 1 was responsible for separating the AFR populations from the others, and PC 2 was largely responsible for separating the EUR and SAS from each other, and from the East Asian Super Population (EAS) and POL populations. The EAS and POL populations were extremely

¹<http://amp.pharm.mssm.edu/Enrichr/> accessed 26 February 2018

²<https://www.ebi.ac.uk/gwas/> accessed 19 June 2017

similar in PC 1 but the POL populations were spread with PC 2, with the majority of the spread being accounted for by the Eastern Polynesian populations whereas the Western Polynesian populations had a tight grouping at -0.025 on PC 2, slightly below the EAS group. This can be observed when PC 6 is used to separate the Polynesian populations (Figure 4.3). PC 3 separated the POL populations from the EAS populations (Figure 4.1 B) and PC 4 separated the AMR and SAS populations (Figure 4.1 C). The Polynesian populations were split into Eastern and Western Polynesian sub-groups with PC 6 (Figure 4.3). This confirmed that the genetic variation was higher between populations, than between individuals of a population, and indicated how genetically similar or dissimilar populations were to each other. Populations that were thought to have a large degree of admixture also indicated as such by the spread, and overlap with other better defined clusters.

The admixture of the AMR populations (1000 Genomes Project Consortium, 2015) was evident in PC 1 with a large spread centred with the EUR populations. The first four PC indicated that the six super populations could form super population clusters based on the genetic variation. There was no distinction between the Utah Residents (CEPH) with Northern and Western Ancestry (CEU), British in England and Scotland (GBR), and Europeans in New Zealand (NZC) populations (Figure 4.2), this indicated that the use of SNP array versus sequence data was not a major source of variability in the dataset.

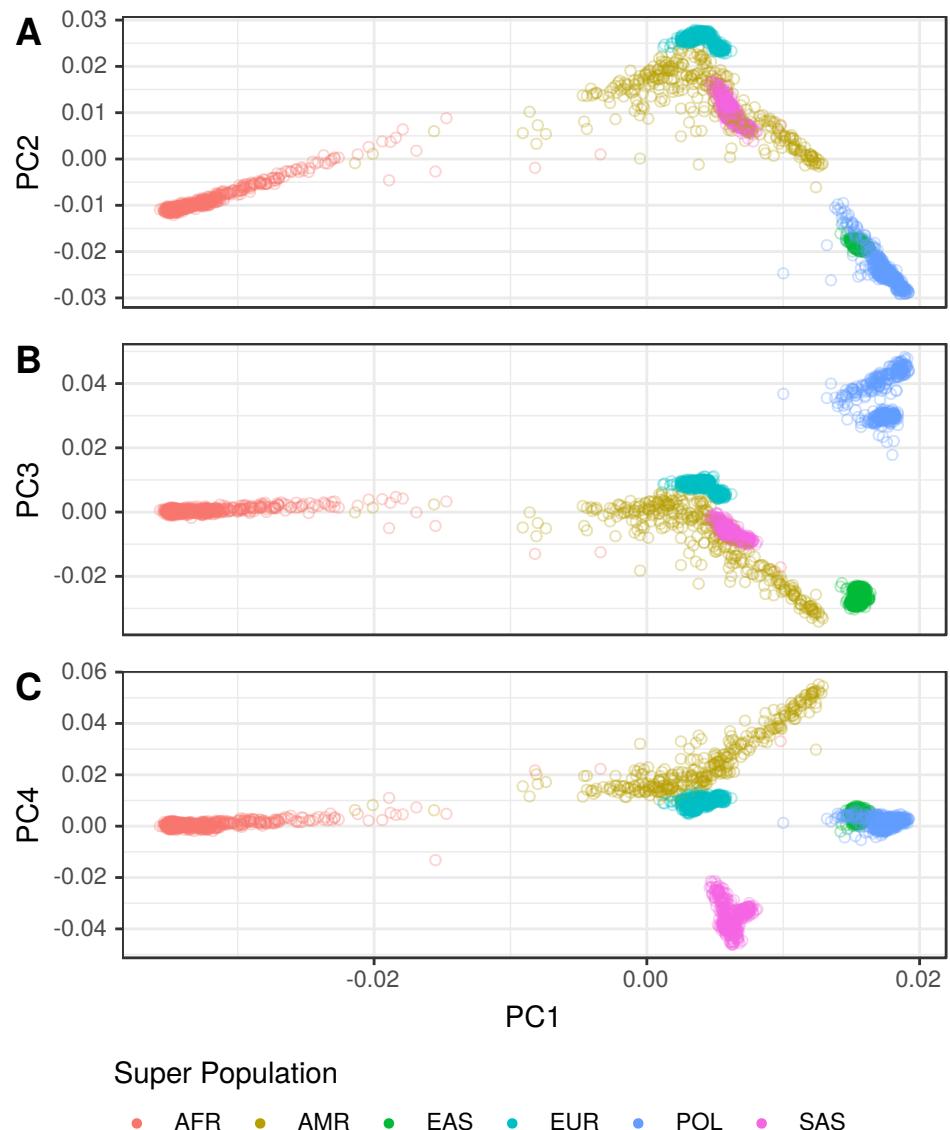


Figure 4.1: A) Principal components 1 and 2 for all populations. B) Principal components 1 and 3 for all populations. C) Principal components 1 and 4 for all populations. Coloured by super population grouping.

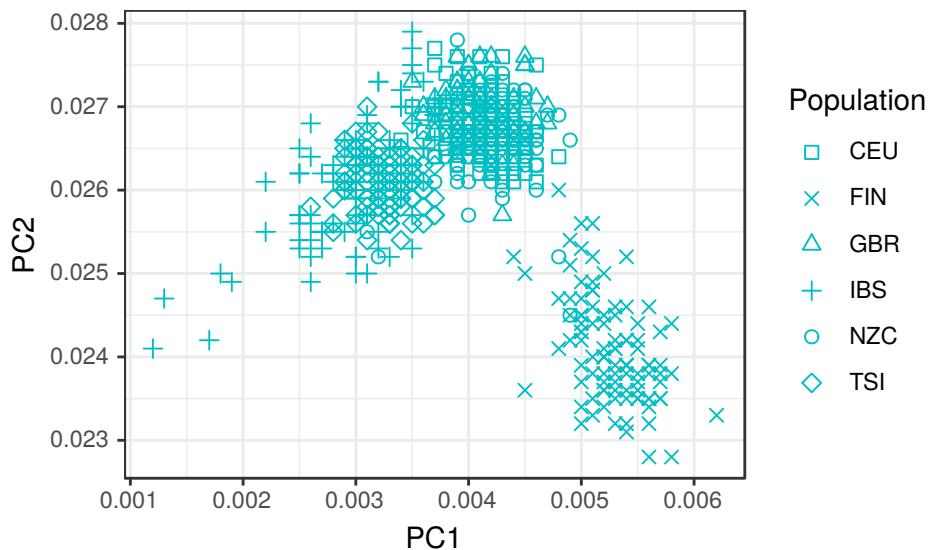


Figure 4.2: Principal components 1 and 2 for populations of the European super population.

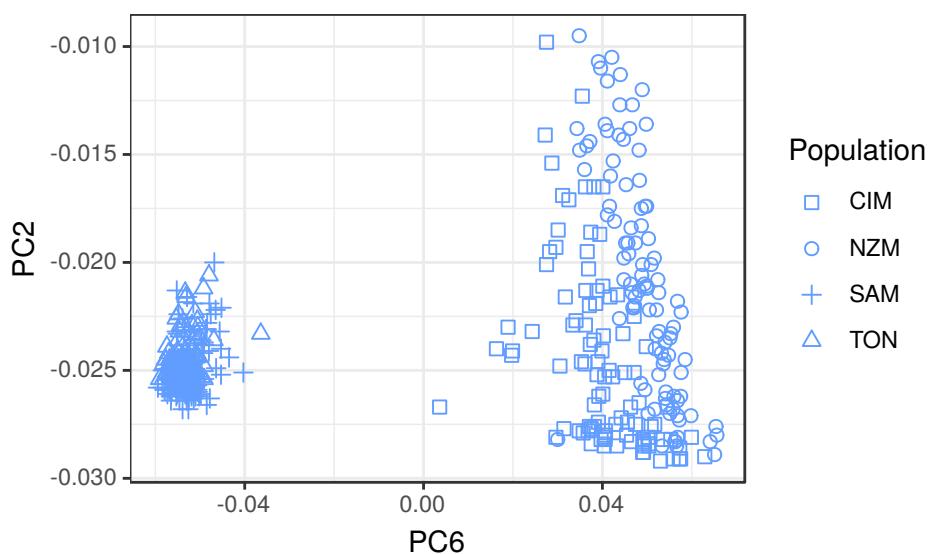


Figure 4.3: Principal components 2 and 6 for the Polynesian populations.

4.3.1.2 Admixture analysis

Admixture analysis is a method that can be used to infer the number of ancestral populations that contribute to current populations by modelling the probability of the observed genotypes using proportions of ancestral populations and the population allele frequencies (Alexander and Novembre, 2009). Proportions of theoretical ancestral populations for each population were calculated, using all of the populations to establish the number of ancestral populations (K), that resulted in the smallest cross-validation error. Five-fold cross-validation was used to calculate this error for K from 1 to 15. The K value with the lowest cross-validation error was $K = 11$, with an error of 0.3855 (Table S6). Admixture analysis was performed to identify the likely presence of admixture in populations, not for quantification purposes, for cultural sensitivity reasons.

To understand the similarities in ancestral background and degree of admixture present at a population level, the ancestral proportions for $K = 11$ were plotted for each individual for each population revealing there were some clear similarities amongst the super populations (Figure 4.4). Within the AFR populations there were three main ancestral populations, which correlated with geography, with the East African population of Luhya in Webuye Kenya (LWK) being most different to the West African populations (Esan in Nigeria (ESN), Yoruba in Ibadan Nigeria (YRI), Gambian in Western Divisions in the Gambia (GWD), and Mende in Sierra Leone (MSL)). The populations of African Caribbean in Barbados (ACB) and Americans of African Ancestry in SW USA (ASW) showed signs of admixture with similarities to the ancestral populations in the West African populations, with smaller proportions from the ancestral populations of LWK, and the European populations. The EAS populations were represented by two ancestral populations. Chinese Dai in Xishuangbanna China (CDX) was mostly one, and Japanese in Tokyo Japan (JPT) mostly the other, the remaining EAS populations were mixtures of the two. A single ancestral population represented SAS, but there were indications of admixture with the ancestral populations of the EUR and EAS populations. In the EUR populations there was again two main ancestral populations, they correlated with North and South Europe. One ancestral population was mostly in Finnish in Finland (FIN), and the other mostly in Toscani in Italia (TSI) and Iberian Population in Spain (IBS). GBR, CEU, and NZC were all extremely similar in proportions of the two and ancestral populations. The AMR populations were mostly differing proportions of the ancestral population that TSI had, and an ancestral population that was only in the AMR populations. There were also small proportions of the ancestral populations that made up the West African populations. The POL populations had two ancestral populations, one found in the East Polynesian populations, and the other in the West Polynesian populations. All Polynesian populations had indications of small amounts of admixture with the ancestral populations of the EUR populations, with the East Polynesian populations having a higher amount than the West Polynesian populations. This analysis was consistent with the results of the PCA, and other admixture analysis (Hellenthal *et al.*, 2014; 1000 Genomes Project Consortium, 2015), but also showed that there were sub-groups within the super-populations. The low level of admixture in the Polynesian populations also indicated its confounding role in subsequent analyses would be minimal.

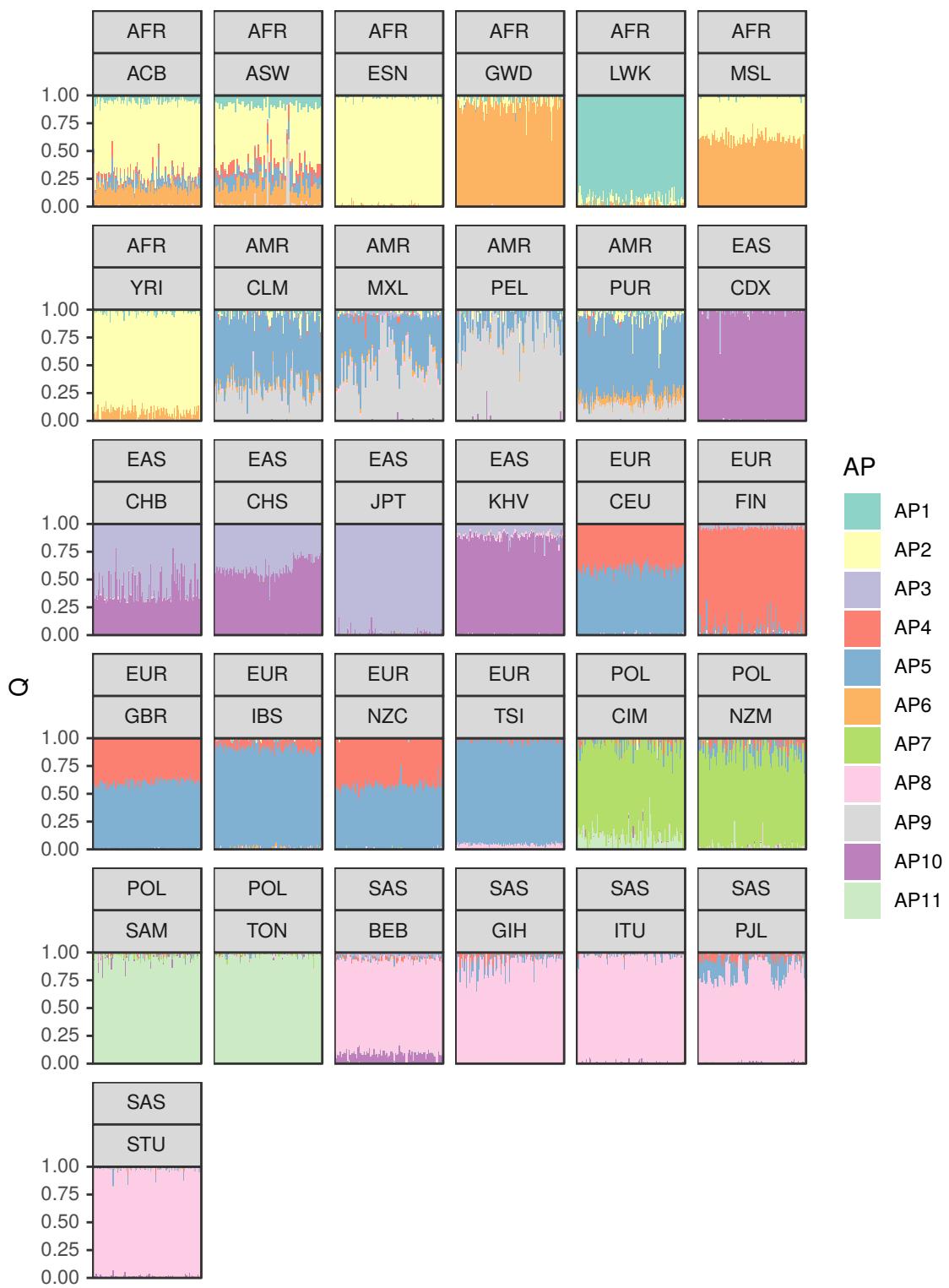


Figure 4.4: Proportions of ancestral populations as inferred from ADMIXTURE, using $K = 11$. Ancestral population (AP) represents a particular theoretical ancestral population. Q is the ancestral proportion. Individuals in the sample populations are represented on the x-axis.

4.3.1.3 Whole chromosome clustering

In order to establish at a summary genome level how similar populations were by selection and neutrality statistic were, clustering on each statistic summarising each chromosome was done. Clustering on the full chromosome results for all chromosomes from PopGenome (Figure 4.5) did not group the populations of each super population together, with the exception of Fay and Wu's H , which grouped all super populations, although it split the AMR populations into two subgroups. Tajima's D grouped the EUR and most of the SAS super populations, and had smaller sub-super population groupings of two to three populations. Cook Island Māori in New Zealand (CIM), Māori in New Zealand (NZM), and Samoans in New Zealand (SAM) were grouped with two AMR populations, (Peruvians from Lima Peru (PEL) and Mexican Ancestry from Los Angeles USA (MXL)), and this group was the least similar to the others. There was no clear pattern to the clustering order of Fu and Li's F , although it did have a few small groups of the same super population of no more than three individual populations. The CIM and SAM populations were grouped together for both statistics. Zeng's E clustered the AFR populations together and this cluster was the least similar to the other populations. There was also a grouping of the EAS populations which had the Western Polynesian populations as part of it. This indicated that at a summary level many of the populations within a super population were similar to each other.

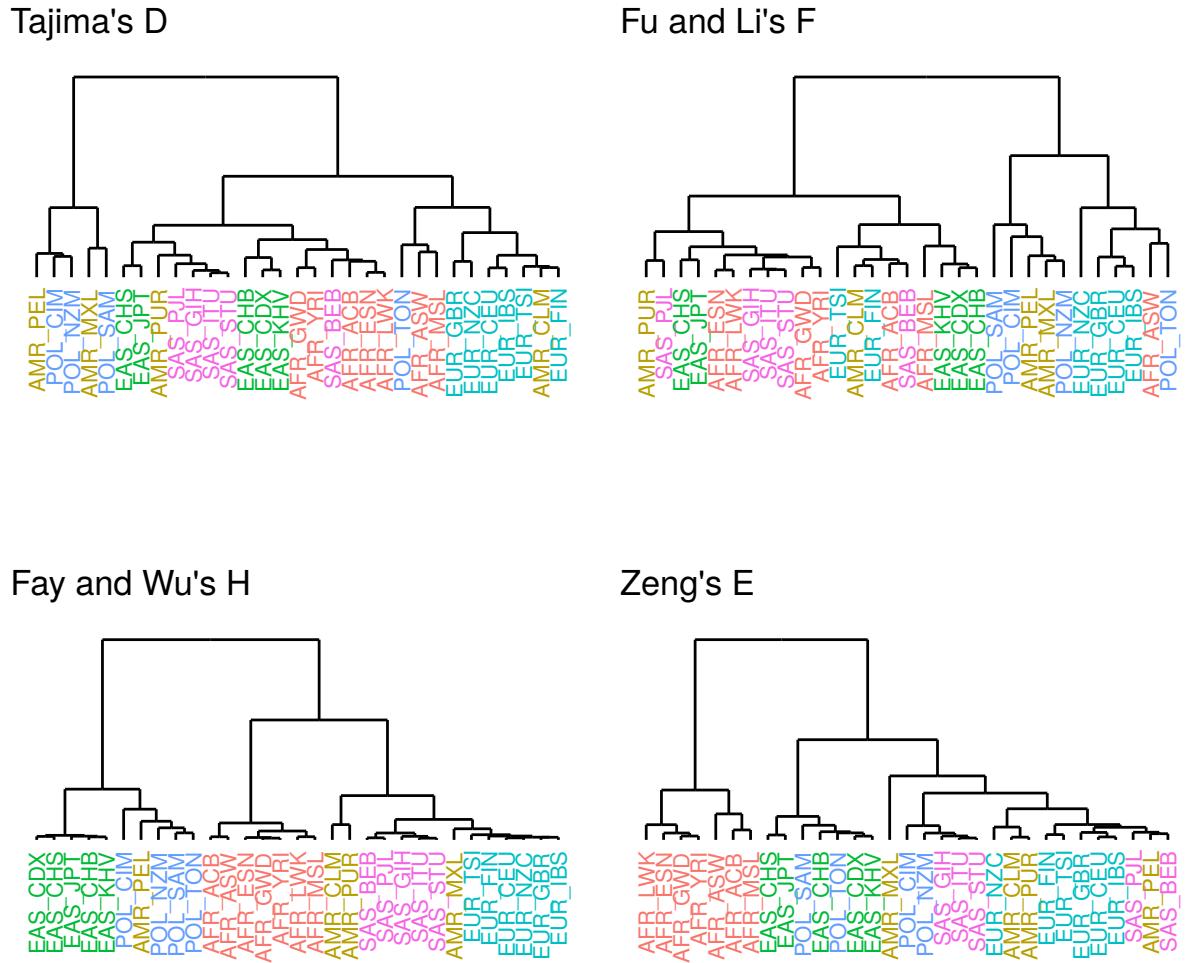


Figure 4.5: Hierarchical clustering of the populations using selection test statistics for each entire chromosome. Clustering was performed using Euclidean distance for the distance measure and complete linkage for the linkage criteria. Selection test statistic is in the label for each dendrogram. Coloured by super population.

4.3.1.4 F_{ST}

Hierarchical clustering performed on the pair-wise F_{ST} for the whole chromosomes between each individual population grouped all of the super populations, except AMR into their correct groups (Figure 4.6). The AFR populations were most similar to each other but least similar to the rest of the populations. The Polynesian populations were grouped together and were sub grouped into East and West Polynesian. The largest differentiation was between CIM and ESN (mean $F_{ST} = 0.203$), whereas the smallest differentiation was between CEU and NZC (mean $F_{ST} = 3.42 \times 10^{-4}$). The mean F_{ST} for POL and AFR was 0.184; POL and AMR was 0.110; POL and EAS was 0.065; POL and EUR was 0.128; POL and SAS was 0.095. This shows that the POL populations were least differentiated from EAS populations and most from the AFR populations, which is consistent with the migration history.

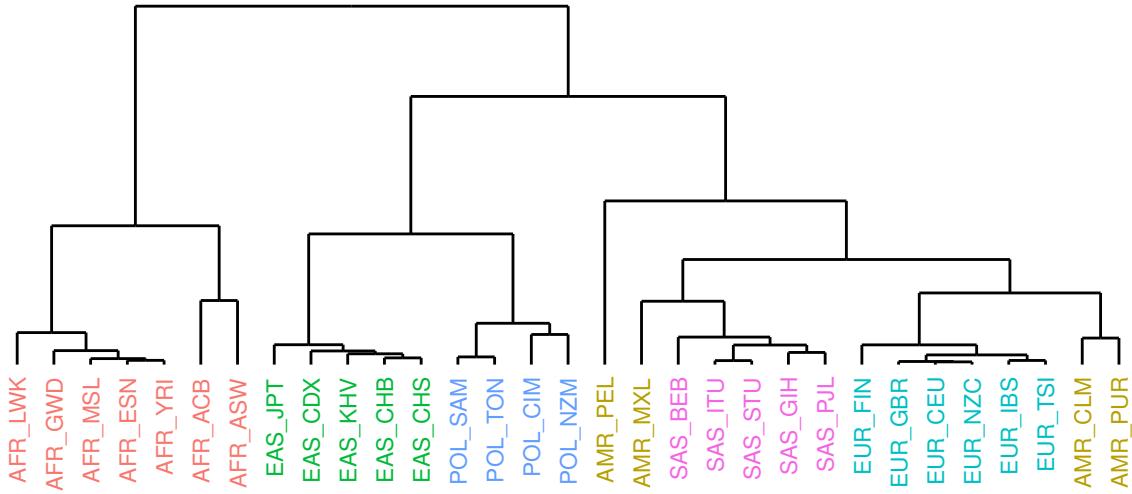


Figure 4.6: Hierarchical clustering on chromosomal F_{ST} using Euclidean distance and complete linkage. Coloured by super population.

4.3.2 Hierarchical clustering of distribution extremes

Hierarchical clustering of the distribution extremes for each population was done to investigate if there was a similarity between populations and the genetic regions that had the most extreme scores. For the frequency spectrum methods, the 1st and 99th percentiles of windows (i.e., distribution tails) by statistic value were calculated, and a binary matrix of sharing of significant windows between populations was used for the clustering. This was done with the premise being that the extremes would represent regions that had a similar selective pressure based on geographic similarity, or selective pressure that operated on a shared ancestral population. For the haplotypic methods, the proportion of regions that overlapped between populations was used to create the distance matrix for the hierarchical clustering. The groups created by the clustering were then assessed for the number of populations from a super population that were assigned the same cluster. Having all populations correctly grouped into their super population was only completely achieved by clustering using the lower tail of Fay and Wu's H . If an allowance is made for the two AMR populations to be grouped with the EUR populations, as

in the chromosomal F_{ST} clustering, then both haplotypic methods and six of the clustered tails of the frequency spectrum-based methods clustered populations into their respective super populations.

There were differing numbers of windows used for each population, ranging from a minimum of 2388 to a maximum of 2488. The mean number of windows per population was 2444.7 (SD 22.9) and was due to multiple windows having the same value. Table 4.1 and Table 4.2 show the proportion and exclusivity of the clusters respectively, for the clustering of populations by cutting the dendograms to create six groups, based on the number of super populations. Proportions were calculated independently for each super population based on the cluster group with the maximum number of populations from a particular super population, whereas, exclusivity of each cluster was calculated by using the proportions of populations from a given super population assigned compared to the total number of individual populations assigned to that cluster. Clustering based on the windows of the extreme tails of the selection statistic distributions resulted in a higher number of the individual populations being clustered into their super populations than for the chromosome-wide selection statistics, and the clusters were also more exclusive to the super populations. This shows that the aggregate frequency spectrum-based selection and neutrality statistics do not capture the population differentiation, but that the extremes for each population do replicate the grouping that population differentiation had with F_{ST} .

The following subsections will cover the distributions for the frequency-based statistics calculated genome-wide in windows to highlight that populations within a super population had similar distributions, but between super populations there were differences. For each statistic there will also be a description of the clustering that was done on the 1st and 99th percentiles, and make note of any of the metabolic disease associated loci from chapter 3 that met the selection thresholds that were found in windows from the distribution extremes. There is a subtle difference between the windows that were deemed of interest in chapter 3, and the windows in the extremes of the distributions, in that the latter does not have the greater or less than zero condition. Each statistic also represents different ‘partitions’ of the frequency spectrum with Tajima’s D representing low versus intermediate frequency variants, Fu and Li’s F representing singletons versus intermediate frequency variants, Fay and Wu’s H representing intermediate versus high frequency derived variants, and Zeng’s E representing low versus high frequency variants.

4.3.2.1 Tajima’s D

4.3.2.1.1 Tajima’s D distributions

The distributions of the genome-wide Tajima’s D values showed that there were differences between each population, and super population (Figure 4.7 and Table 4.3). The Tajima’s D means for the Polynesian populations were the smallest of the super populations at 2.003 (SD 1.049) with 1.875 (SD 1.073) for CIM, 1.929 (SD 1.042) for NZM, 2.061 (SD 1.041) for SAM, and 2.151 (SD 1.016) for Tongans in New Zealand (TON) (Table S9). The means of the Polynesian populations were lower than all of the other populations except PEL and MXL, indicating that overall the Polynesian populations had a higher proportion of low frequency to intermediate frequency variants. The SAS populations

Table 4.1: Proportion of populations clustered into their corresponding super population by selection and neutrality statistic. Proportions were calculated independently for each super population based on the cluster group with the maximum number of populations from a particular super population.

Type	AFR	AMR	EAS	EUR	POL	SAS
Fu and Li's F						
Chromosome	0.57	0.50	0.60	0.50	0.75	0.80
Lower Tail	0.71	1.00	1.00	1.00	0.75	1.00
Upper Tail	1.00	0.50	1.00	1.00	0.50	1.00
Urate/Gout	1.00	1.00	0.80	0.83	0.75	1.00
Type 2 diabetes	1.00	0.75	1.00	1.00	0.50	1.00
Obesity	1.00	0.75	1.00	1.00	0.50	1.00
Kidney Disease	1.00	1.00	1.00	0.33	1.00	1.00
Metabolic Syndrome	0.57	0.75	1.00	1.00	0.50	1.00
Fay and Wu's H						
Chromosome	1.00	0.50	1.00	1.00	0.75	1.00
Lower Tail	1.00	1.00	1.00	1.00	1.00	1.00
Upper Tail	1.00	0.50	1.00	1.00	1.00	1.00
Urate/Gout	1.00	0.75	1.00	1.00	1.00	1.00
Type 2 diabetes	1.00	0.50	1.00	1.00	0.50	1.00
Obesity	1.00	0.75	1.00	1.00	1.00	1.00
Kidney Disease	1.00	1.00	0.80	1.00	0.50	1.00
Metabolic Syndrome	1.00	0.75	1.00	1.00	1.00	1.00
Tajima's D						
Chromosome	0.71	0.25	0.60	1.00	0.50	0.80
Lower Tail	1.00	0.50	1.00	1.00	1.00	1.00
Upper Tail	1.00	0.50	1.00	1.00	1.00	1.00
Urate/Gout	1.00	0.75	1.00	1.00	1.00	1.00
Type 2 diabetes	1.00	0.50	1.00	1.00	0.50	1.00
Obesity	1.00	0.75	1.00	1.00	1.00	1.00
Kidney Disease	1.00	0.50	1.00	1.00	1.00	1.00
Metabolic Syndrome	1.00	0.75	1.00	1.00	1.00	1.00
Zeng's E						
Chromosome	0.57	0.75	1.00	1.00	0.50	0.60
Lower Tail	1.00	0.50	1.00	1.00	1.00	1.00
Upper Tail	1.00	0.50	1.00	1.00	1.00	1.00
Urate/Gout	1.00	0.75	1.00	1.00	1.00	1.00
Type 2 diabetes	1.00	0.50	1.00	1.00	1.00	1.00
Obesity	1.00	0.50	1.00	1.00	1.00	1.00
Kidney Disease	1.00	1.00	0.80	1.00	0.75	1.00
Metabolic Syndrome	1.00	0.75	1.00	1.00	1.00	1.00

Table 4.2: Mean exclusivity of clusters for a given super population by selection and neutrality statistic. Exclusivity of each cluster was calculated by using the proportions of populations from a given super population assigned compared to the total number of individual populations assigned to that cluster

Type	AFR	AMR	EAS	EUR	POL	SAS
Fu and Li's F						
Chromosome	0.25	0.23	0.47	0.67	0.40	0.28
Lower Tail	1.00	0.36	1.00	0.55	0.55	1.00
Upper Tail	1.00	0.58	1.00	0.46	1.00	0.38
Urate/Gout	1.00	0.40	0.62	0.55	0.88	0.50
Type 2 diabetes	1.00	0.61	1.00	0.43	1.00	0.36
Obesity	1.00	0.61	1.00	0.43	1.00	0.36
Kidney Disease	1.00	0.67	0.56	0.78	0.44	1.00
Metabolic Syndrome	0.64	0.21	1.00	0.86	1.00	0.45
Fay and Wu's H						
Chromosome	1.00	0.44	1.00	0.50	0.88	0.42
Lower Tail	1.00	1.00	1.00	1.00	1.00	1.00
Upper Tail	1.00	0.62	1.00	0.75	1.00	1.00
Urate/Gout	1.00	0.69	1.00	1.00	1.00	0.62
Type 2 diabetes	1.00	0.58	1.00	0.46	1.00	0.38
Obesity	1.00	0.67	1.00	0.67	1.00	1.00
Kidney Disease	1.00	1.00	0.65	0.55	0.57	0.45
Metabolic Syndrome	1.00	0.69	1.00	1.00	1.00	0.62
Tajima's D						
Chromosome	0.61	0.28	0.31	0.86	0.50	0.34
Lower Tail	1.00	0.62	1.00	0.75	1.00	1.00
Upper Tail	1.00	0.62	1.00	0.75	1.00	1.00
Urate/Gout	1.00	0.69	1.00	1.00	1.00	0.62
Type 2 diabetes	1.00	0.58	1.00	0.46	1.00	0.38
Obesity	1.00	0.69	1.00	1.00	1.00	0.62
Kidney Disease	1.00	0.62	1.00	0.75	1.00	1.00
Metabolic Syndrome	1.00	0.69	1.00	1.00	1.00	0.62
Zeng's E						
Chromosome	1.00	0.64	0.71	0.55	0.34	0.39
Lower Tail	1.00	0.62	1.00	0.75	1.00	1.00
Upper Tail	1.00	0.62	1.00	0.75	1.00	1.00
Urate/Gout	1.00	0.69	1.00	1.00	1.00	0.62
Type 2 diabetes	1.00	0.62	1.00	0.75	1.00	1.00
Obesity	1.00	0.62	1.00	0.75	1.00	1.00
Kidney Disease	1.00	0.44	0.75	1.00	0.75	0.56
Metabolic Syndrome	1.00	0.69	1.00	1.00	1.00	0.62

Table 4.3: Summary statistics for Tajima’s D by super population.

Super Population	Mean	SD	Min	1st Percentile	Median	99th Percentile	Max
AFR	2.378	0.717	-1.785	0.362	2.454	3.730	4.610
AMR	2.228	0.948	-2.252	-0.522	2.376	3.850	4.827
EAS	2.348	0.985	-2.642	-0.628	2.521	3.966	4.863
EUR	2.321	0.977	-2.200	-0.603	2.491	3.929	4.884
POL	2.003	1.049	-2.559	-0.977	2.160	3.824	4.837
SAS	2.442	0.891	-2.274	-0.257	2.590	3.948	4.772

had the highest mean Tajima’s D overall at 2.442 (SD 0.891). The AFR populations were the only populations with the 1st percentile above zero, but also had the smallest overall maximum value of 4.610. The EUR populations had the highest maximum Tajima’s D value at 4.884. The smallest Tajima’s D value of -2.642 came from the Kinh in Ho Chi Minh City Vietnam (KHV) population of the EAS super population. The AFR population had the smallest range (6.395), whereas the EAS super population had the largest (7.506). The POL population had the second largest (7.396). The distribution for Tajima’s D should normally centre on zero, however, a right shifted distribution of Tajima’s D , such as all distributions in Figure 4.7, can be attributed to ascertainment bias from using a SNP array, where there is a bias against low frequency markers.

The clustering of the individual populations for the 1% most extreme values from each of the lower and upper tails of the Tajima’s D distribution both showed the East/West split in the Polynesians (Figure 4.8). But the clustering also showed that the Polynesian populations were most similar to each other. Both tails placed the Polynesian cluster closest to the EAS cluster, consistent with the migration history. The main groupings are of the populations that were designated as the same super population, although the AMR super population in both tails has been split into two. The mean number of unique windows per population was 274.5 (SD 153.1) in the upper tail, compared with 202.4 (SD 101.2) in the lower tail, indicating a higher degree of window sharing in the 1st percentiles of populations. The groupings for both the upper and lower tails exclusively grouped each of the AFR, SAS and POL super populations. The AMR super population was split into an exclusive group of MXL and PEL, and a group of Colombians from Medellin Colombia (CLM) and Puerto Ricans from Puerto Rico (PUR) which was shared with the EUR super population. All clusters, except the AMR - EUR cluster, were each exclusive to a single super population.

4.3.2.1.2 Clustering Tajima’s D 1st percentile

In the lower tail, the number of windows that were unique to the Polynesian super population was 2188, representing 11.1% of all windows in the lower tail that were used. This compared to super population unique window numbers of 4825, 1085, 2114, 1250, 1119 for the AFR, AMR, EAS, EUR, and SAS super populations respectively. Of the unique windows for Polynesian populations, 267 were specific to CIM, 266 were specific to NZM, 300 were specific to SAM, and 404 were specific to TON. The Eastern Polynesians had 699 unique windows with 166 of those shared between CIM and NZM, whereas the Western Polynesians had 950 windows, with 246 shared between SAM and TON. There were 314 windows where all the Polynesian populations shared a window with a non-Polynesian population.

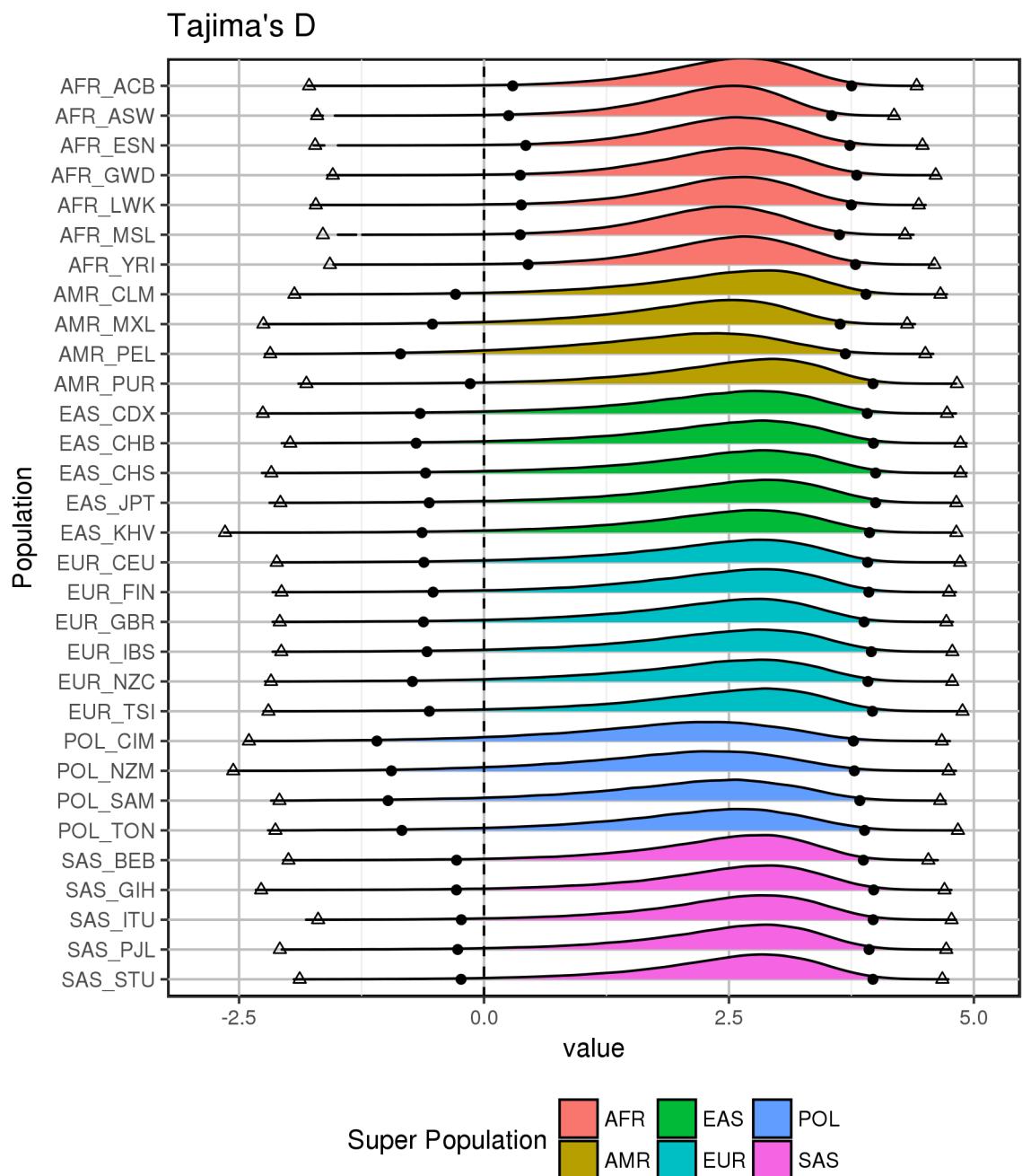


Figure 4.7: Plot of the distribution of Tajima's D by population. Triangles indicate the minimum and maximum values. Dots indicate the 1st and 99th percentiles for each population.

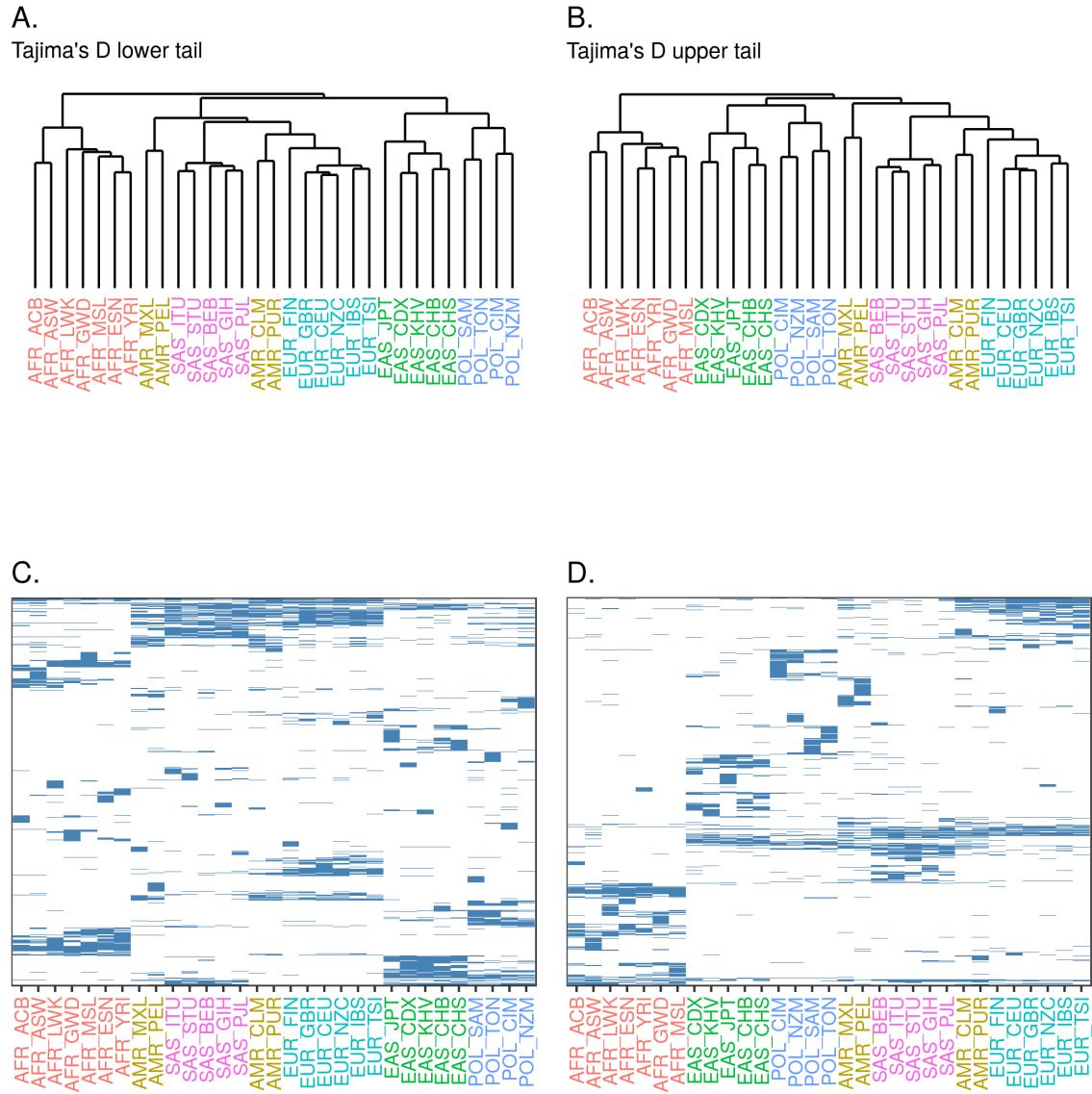


Figure 4.8: Hierarchical clustering of Tajima's D using the upper and lower 1% of the distribution. **A** and **B** Dendograms representing the clusters of Tajima's D lower and upper tails respectively, using hierarchical clustering with Euclidean distance and complete linkage, and coloured by super population. **C** and **D** Plots representing the windows present in each population from the lower and upper tails respectively. A blue line represents presence of a window, white represents the absence of a window in the 1% of the distribution for a population.

And 3130 windows, intersecting 1268 genes, where at least one Polynesian population shared a window with a non-Polynesian population (Table S13). There were 32 genes that had windows that were in the 1st percentile for all four Polynesian populations. Cell signalling as an immune response, such as the Toll-like receptor cascades, were the main pathways this list of genes covered, although this was largely due to *PPP2CB* and *NOD1*. Two other genes in the list were *CCR3* and *ARL15*, both of which are associated with obesity related traits (Comuzzie *et al.*, 2012; Shungin *et al.*, 2015), and the latter also associated with type 2 diabetes (Mahajan *et al.*, 2014).

4.3.2.1.3 Clustering Tajima's *D* 99th percentile

The upper tail had a total of 3036 windows, or 13.5% that were unique to the Polynesian populations. This compares to super population unique window numbers of 4626 (AFR), 1774 (AMR), 2476 (EAS), 1952 (EUR), 1274 (SAS). Eastern Polynesians contributed 985 unique windows, with CIM having 491 unique windows and NZM having 494 windows. The Western Polynesians had 1279 unique windows. 497 were from only SAM and 533 from only TON. There were 291 windows where all the Polynesian populations shared a window with a non-Polynesian population. And 2807 windows, intersecting 1098 genes, where at least one Polynesian population shared a window with a non-Polynesian population (Table S14).

In both tails the clustering was being driven by a small subset of the regions; this can be seen in Figures 4.8 C and D, where most of the regions that the Polynesian populations have are not shared with the other populations. It can also be seen that the regions that the Polynesian populations do have in common are also shared with the EAS populations. This is reflected in the dendograms in Figures 4.8 A and B.

4.3.2.1.4 Metabolic disease-associated genes in the extremes of Tajima's *D*

Specifically focusing on the 465 genes that were associated with urate, gout, obesity, type 2 diabetes, kidney disease, and metabolic syndrome from the GWAS catalog (Tables S7 and S8), from the clustering of the extremes, there were a total of 32 genes that had at least one of the Polynesian populations having at least one window in the 1st percentile. This dropped to 6 genes that had windows in the 1st percentile for at least one Polynesian population, and also in the 1st percentile of other populations. Some examples of this were *CCR3*, which had windows in the 1st percentile for all four Polynesian populations, and the same windows were also in the 1st percentile as the EAS populations of CDX and KHV. *ARL15* had windows that were only from the four Polynesian populations. And *DNAH10* had windows that were in the 1st percentile of the EAS populations except CDX, as well as the Eastern Polynesian populations, CIM and NZM. In the 99th percentile there were 39 genes associated with the metabolic diseases that had windows in the Polynesian populations, of those, 9 genes had windows in the 99th percentile for Polynesians and the 99th percentile of other populations. Some examples include *ADCY3* which had multiple windows with POL, EUR, EAS, and SAS. And *NEGR1* which had windows from both the EAS and POL populations.

Table 4.4: Summary statistics for Fay and Wu's H by super population.

Super Population	Mean	SD	Min	1st Percentile	Median	99th Percentile	Max
AFR	0.028	0.913	-8.113	-2.782	0.183	1.400	1.812
AMR	-0.761	1.366	-11.167	-5.098	-0.507	1.240	1.817
EAS	-1.040	1.551	-11.374	-5.977	-0.743	1.232	1.860
EUR	-0.812	1.415	-13.060	-5.306	-0.546	1.259	1.789
POL	-1.283	1.648	-11.692	-6.395	-0.982	1.162	1.732
SAS	-0.704	1.348	-11.022	-5.031	-0.453	1.271	1.827

4.3.2.2 Fay and Wu's H

4.3.2.2.1 Fay and Wu's H distributions

Comparing genome-wide Fay and Wu's H revealed that the POL super population had the lowest mean (-1.283), 1st percentile (-6.395), median (-0.982), 99th percentile (1.162), and maximum values (1.732) (Figure 4.9 and Table 4.4). Within the POL group, the minimum Fay and Wu's H value for a Polynesian population was from NZM, whereas the maximum value was from SAM. EAS was the second lowest for mean (-1.040), 1st percentile (-5.977), median (-0.743), and 99th percentile (1.232), but had the largest maximum (1.860). The EUR super population had both the smallest minimum and largest range (14.850). The AFR super population were the only group to have a positive mean (0.028), they also had the smallest range and variation (SD 0.913). This indicates that the AFR populations have a deficit of moderate-to-high frequency derived alleles, whereas the POL and EAS populations have the most windows with the highest excess of moderate-to-high frequency derived alleles.

The clustering of Fay and Wu's H (Figure 4.10) had the main groupings the same as the defined super population groups, with the exception of the upper tail where the AMR super population was split in half. The Polynesian populations were clustered together in both the lower and upper tails for Fay and Wu's H . In both tails the Polynesian cluster also had the East/West split within the group. The Polynesian population group was closest to the EAS populations with 1631 windows from the upper tail and 2212 windows from the lower tail in common between at least one Polynesian population and one EAS population.

4.3.2.2.2 Clustering Fay and Wu's H 1st percentile

The clustering of the lower tail assigned all the individuals into their super populations and each of the groups were exclusive to their super population. There were a total of 15,628 windows in the lower tail, with 1546 windows specific to Polynesian populations. This compared to 3767 windows for AFR, 679 windows for AMR, 1081 windows for EUR, and 1375 windows for EAS. The Eastern Polynesian populations had 561 windows, of those, 230 were specific to CIM and 189 were specific to NZM. The Western Polynesian populations had 593 windows, with 156 specific to SAM and 283 specific to TON. The mean number of windows specific to an individual population was 135.6 (SD 72.8) in the lower tail. There were 595 windows with all Polynesians that also shared with a non-Polynesian population. And 3208 windows, intersecting 885 genes with at least one Polynesian population that also shared

Fay and Wu's H

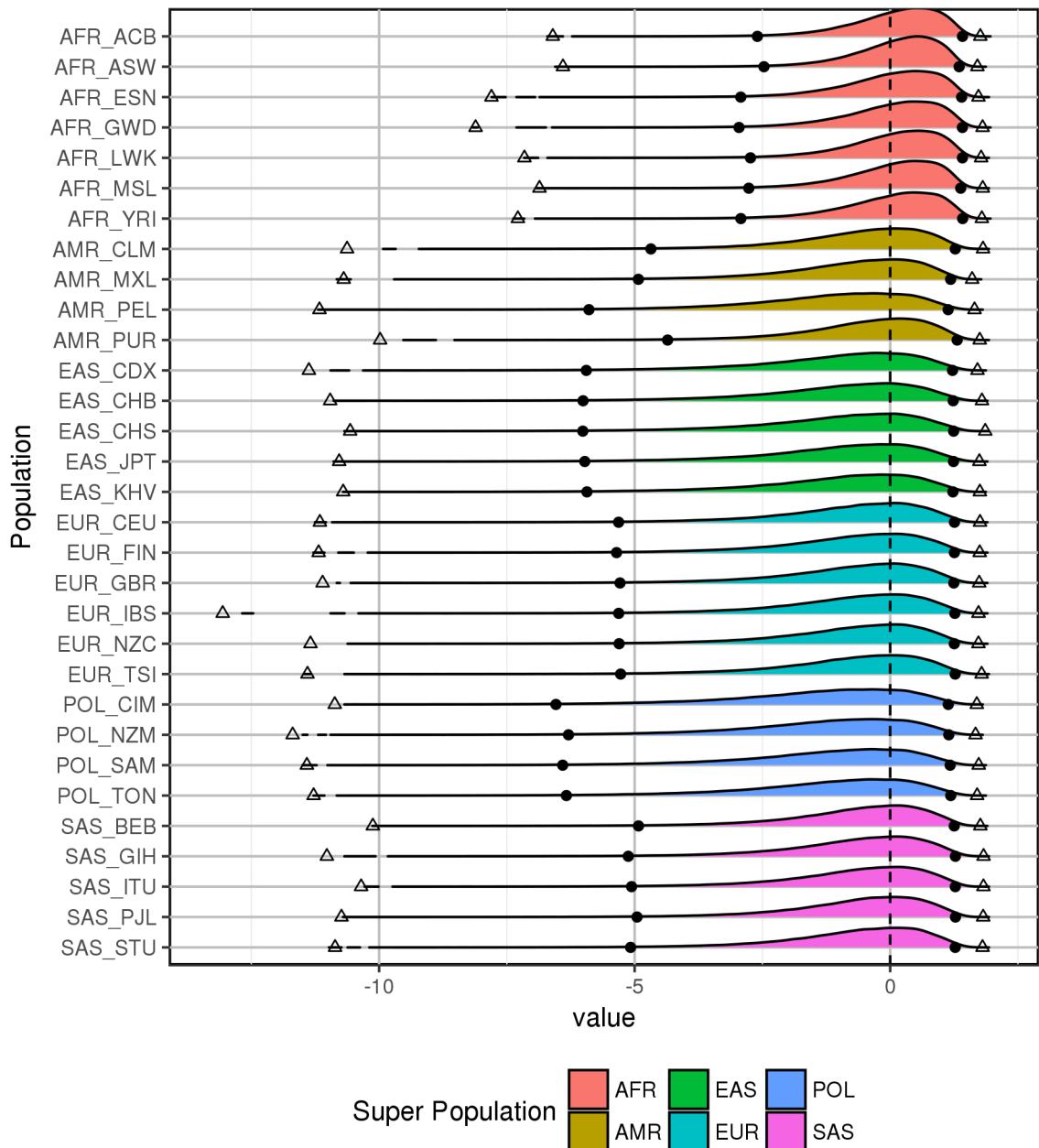


Figure 4.9: Plot of the distribution of Fay and Wu's H by population. Triangles indicate the minimum and maximum values. Dots indicate the 1st and 99th percentiles for each population.

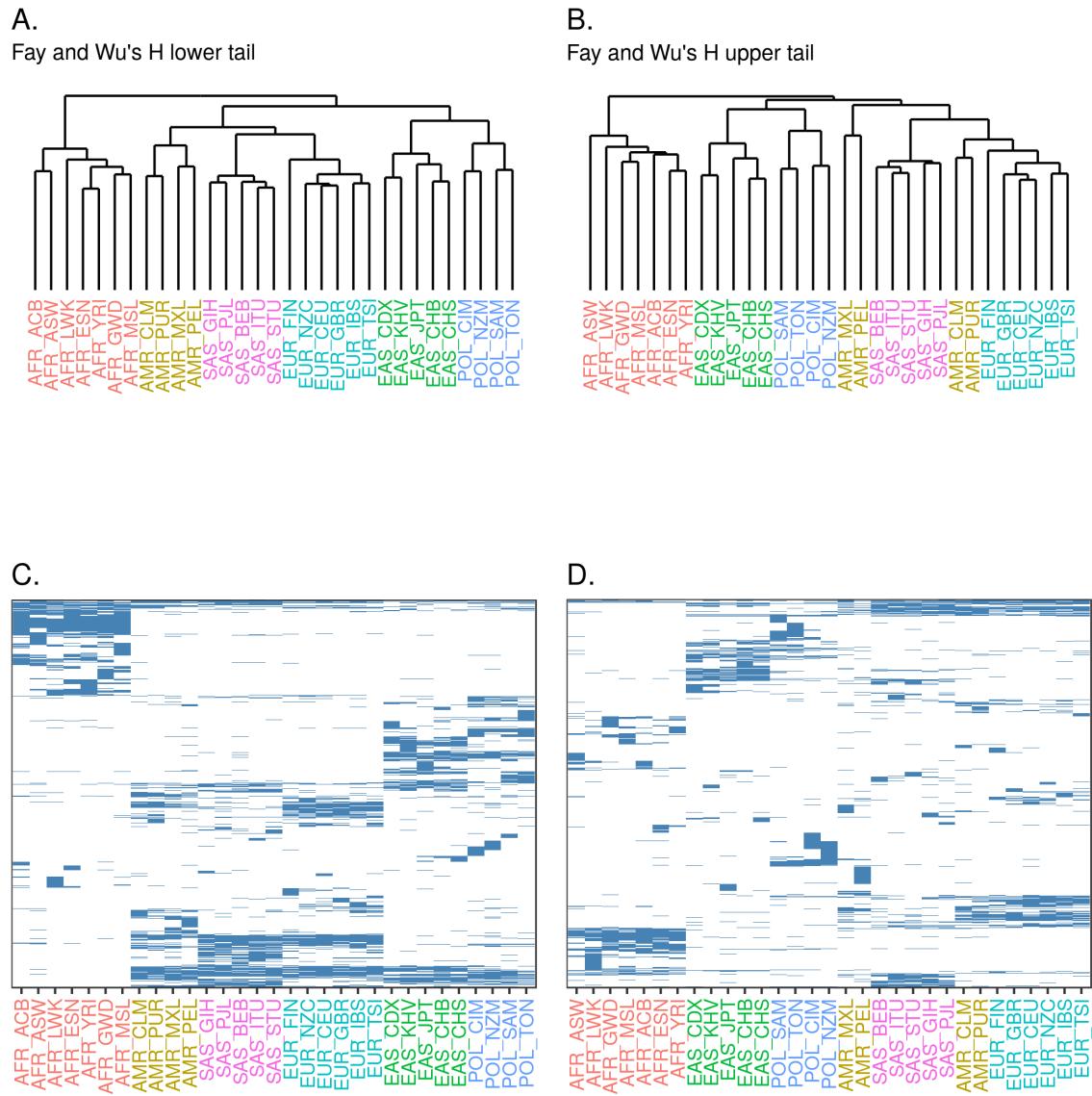


Figure 4.10: Hierarchical clustering of Fay and Wu's H using the upper and lower 1% of the distribution. **A** and **B** Dendograms representing the clusters of Fay and Wu's H lower and upper tails respectively, using hierarchical clustering with Euclidean distance and complete linkage, and coloured by super population. **C** and **D** Plots representing the windows present in each population from the lower and upper tails respectively. A blue line represents presence of a window, white represents the absence of a window in the 1% of the distribution for a population.

with a non-Polynesian population (Table S15). There were 38 genes that had windows that were only in the four Polynesian populations and no others.

4.3.2.2.3 Clustering Fay and Wu's H 99th percentile

In the upper tail of the Fay and Wu's H distribution, the clustering assigned all of the individual populations into their super populations with only AMR being split. The CLM and PUR populations were clustered as part of the EUR group. Exclusivity was also high with all clusters except the EUR - AMR being exclusive to a single super population. Of the 20,659 windows, 2545 windows were from Polynesian populations only. The Eastern Polynesian populations had 1119 specific windows, of those, 404 were unique to CIM and 371 were unique to NZM. The Western Polynesian populations had 925 unique windows, of these, 305 were specific to SAM and 332 specific to TON. The mean number of unique windows for all populations was 226.4 (SD 144.1). There were 344 windows with all Polynesians that also shared with a non-Polynesian population. And 2731 windows, intersecting 1176 genes, with at least one Polynesian population that also shared with a non-Polynesian population (Table S16).

The 1st percentile has many of the regions in the Polynesian populations in common with the EAS populations (Figure 4.10). Whereas, in the 99th percentile there were more regions that were only from the Polynesian populations that were not in common with any other populations. This shows that the regions with high frequency derived alleles are in common with the EAS populations, indicating they possibly originated prior to the Polynesian migration.

4.3.2.2.4 Metabolic disease-associated genes in the extremes of Fay and Wu's H

Of the genes that were in the windows of the 1st percentile, there were 30 genes that were in the list of genes of metabolic disease associated genes and also had windows that intersected them from the Polynesian populations. Thirteen of those genes were for windows from Polynesian populations, as well as others. Some examples of these genes were *RASA2* and *ERBB4*, and both are associated with obesity traits and involved in signalling pathways. *RASA2* had windows only from CIM and NZM. *ERBB4* had windows from all four Polynesian populations, as well as the populations of EAS - except JPT.

There were 25 genes that were in the 99th percentile that were also in the list of genes of metabolic disease associated genes. Some examples of genes that had windows in both Polynesian and other populations in the 99th percentile include *SHROOM3*, *RBMS1*, and *UBE2E2*. *SHROOM3* was an example of a gene that had windows in common from all the other super populations, except EAS. *RBMS1* had multiple windows for across all Polynesian populations, but not other super populations. And *UBE2E2* had multiple windows across the Polynesian populations, as well as the populations of EAS and SAS. *SHROOM3* was an example of a gene associated with kidney disease, with its role being in the development of the kidney (Khalili *et al.*, 2016).

Table 4.5: Summary statistics for Fu and Li's F by super population.

Super Population	Mean	SD	Min	1st Percentile	Median	99th Percentile	Max
AFR	1.992	0.543	-4.068	0.243	2.050	2.982	4.072
AMR	1.891	0.694	-4.809	-0.342	1.998	3.055	3.953
EAS	1.955	0.663	-5.262	-0.108	2.044	3.110	3.934
EUR	1.899	0.744	-5.048	-0.528	2.025	3.100	4.013
POL	1.719	0.754	-4.498	-0.685	1.844	2.970	4.040
SAS	2.012	0.643	-4.068	-0.055	2.104	3.107	3.921

4.3.2.3 Fu and Li's F

4.3.2.3.1 Fu and Li's distributions

The distributions of Fu and Li's F were all centred above zero, with the lowest mean being that of POL at 1.719 (Figure 4.11 and Table 4.5). The POL populations also had the lowest 1st percentile (-0.685), median (1.844), and 99th percentile (2.970). Conversely, the SAS populations had the highest mean (2.012), median (2.104), and were equal with the AFR populations with the highest minimum (-4.068). The AFR populations were the only group to have a positive 1st percentile (0.243). The EAS populations had the lowest minimum (-5.262), but the highest 99th percentile (3.110) and the largest range (9.195). This pattern was similar to what was seen with Tajima's D .

Clustering on the 1% Fu and Li's F from each tail of the distribution (Figure 4.12) had 13851 windows in the upper tail (Figure 4.12 B and D) and 30798 windows in the lower tail (Figure 4.12 A and C). The mean number of unique windows per population was 120.6 (SD 102.8) in the upper tail, compared with 467.2 (SD 202.1) in the lower tail. Both tails clustered the EAS populations into their own exclusive cluster.

4.3.2.3.2 Clustering Fu and Li's F 1st percentile

Clustering on the lower tail windows grouped the majority of the individual populations together into their super populations. The Polynesian populations were grouped together as a West Polynesian group and the CIM population. The NZM population was grouped with the MXL and PEL as part of a larger group with the EUR super population. Exclusivity of clusters was reasonable with four of six being exclusive. There were a total of 30,798 windows with 2722 windows specific to Polynesian populations. This compared to 6337 windows for AFR, 1322 windows for AMR, 2635 windows for EUR, and 4168 windows for EAS. The Eastern Polynesian populations had a total of 926 unique windows, 483 were unique to CIM and 407 were unique to NZM. The Western Polynesian populations had 1538 unique windows. Of these, 565 were unique to SAM and 811 to TON. There were 15 genes that overlapped windows in the 1st percentiles of all four Polynesian populations (Table S17). Of these genes, three (*OR4C11*, *OR4C15*, and *OR4C16*) were involved with the significantly enriched “Olfactory transduction_Homo sapiens_hsa04740” pathway (adjusted $P = 0.044$).

4.3.2.3.3 Clustering Fu and Li's F 99th percentile

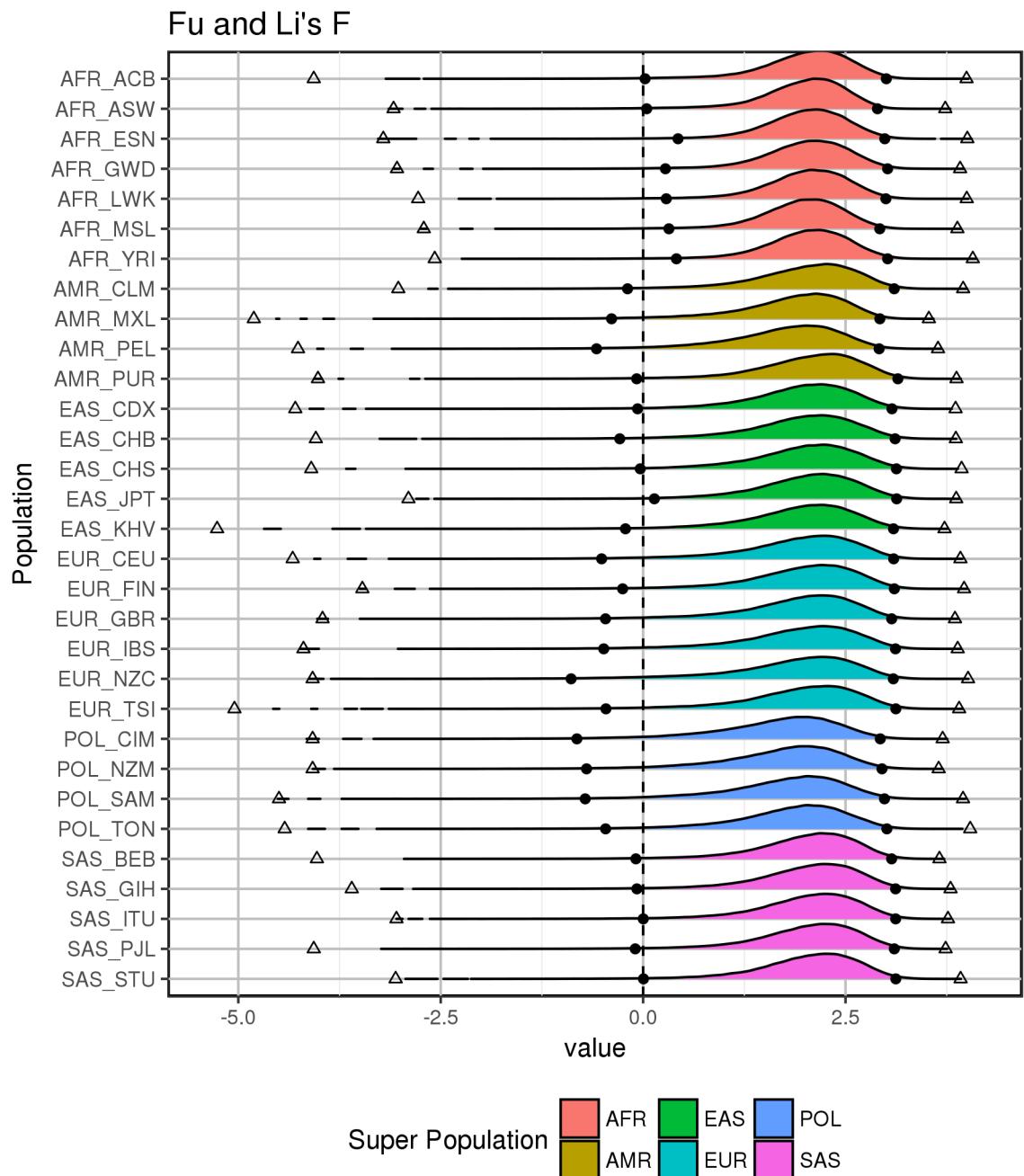


Figure 4.11: Plot of the distribution of Fu and Li's F by population. Triangles indicate the minimum and maximum values. Dots indicate the 1st and 99th percentiles for each population.

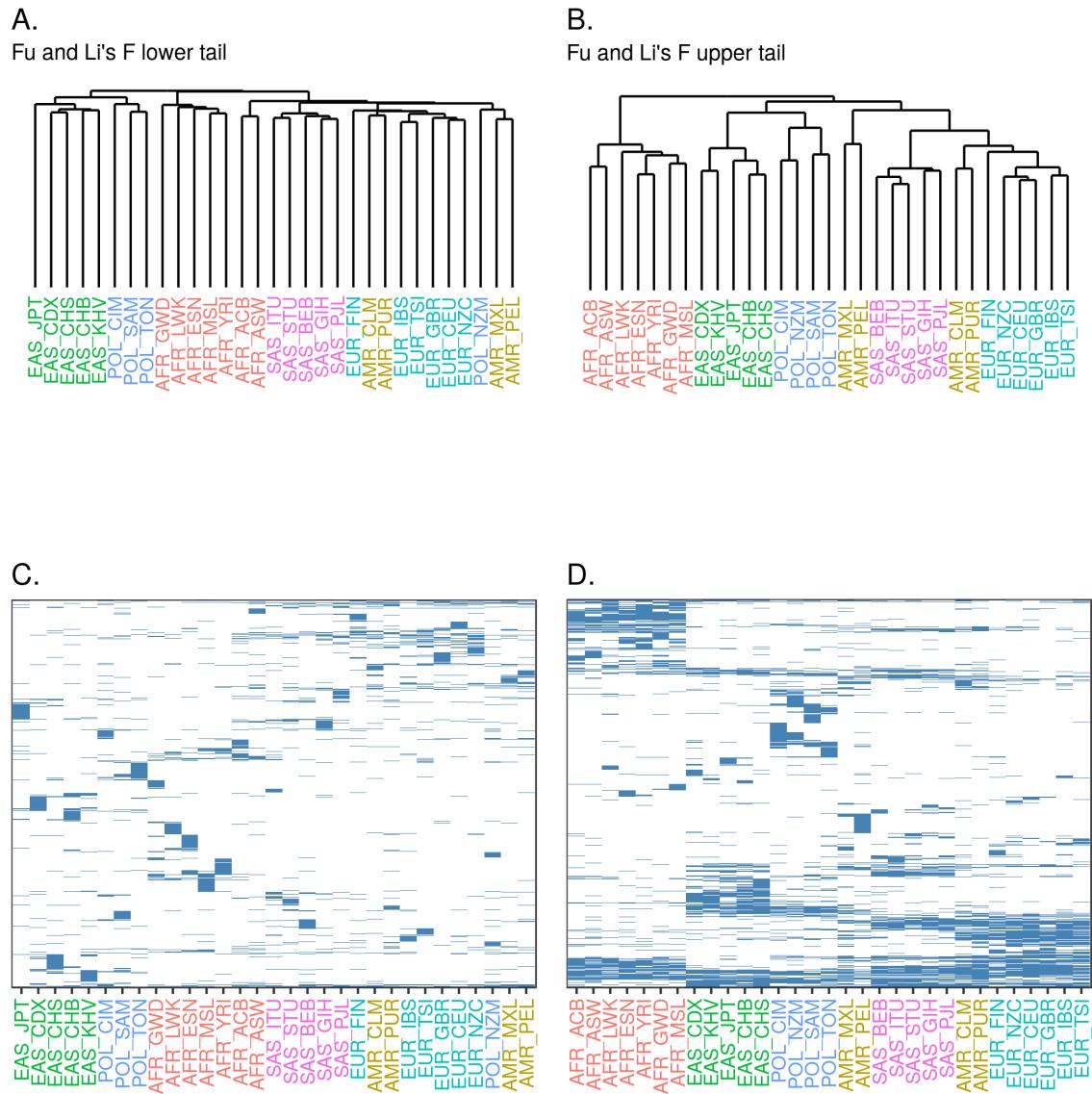


Figure 4.12: Hierarchical clustering of Fu and Li's F using the upper and lower 1% of the distribution. **A** and **B** Dendograms representing the clusters of Fu and Li's F lower and upper tails respectively, using hierarchical clustering with Euclidean distance and complete linkage, and coloured by super population. **C** and **D** Plots representing the windows present in each population from the lower and upper tails respectively. A blue line represents presence of a window, white represents the absence of a window in the 1% of the distribution for a population.

Clustering on the windows from the upper tail put the super populations into their own groups, but also split the AMR populations into two separate groups. The Polynesian populations were also clustered into an East/West split of 2 separate exclusive clusters. Exclusivity of clusters was high with five of the six clusters being exclusive to a single super population, however one cluster contained half of the AMR populations (CLM and PUR) and all of the EUR and SAS populations. There were a total of 13,851 windows with 1804 windows specific to Polynesian populations. This compared to 2020 windows for AFR, 896 windows for AMR, 777 windows for EUR, and 953 windows for EAS. The Eastern Polynesian populations had a total of 795 unique windows with 314 only from CIM and 228 from NZM. From the Western Polynesian populations there were 687 specific windows. There were 267 specific to SAM and 276 specific to TON. There were 101 genes that overlapped windows in the 99th percentiles of all four Polynesian populations (Table S18).

4.3.2.3.4 Metabolic disease associated genes in the extremes of Fu and Li's *F*

There were 50 genes that were in the list of genes of metabolic disease associated genes and also had windows in the 1st percentile, that intersected them from the Polynesian populations. Some examples of these genes were *MAFC1*, *DNAH10* and *SLC2A9*. *MAFC1* had six windows, some of which were in common with the AMR and AFR populations. *DNAH10* had 18 windows, with about half being in common with the populations of EUR, EAS, SAS, and AMR. The urate transporter *SLC2A9* had multiple significant windows in TON, JPT, and Han Chinese in Beijing China (CHB).

There were 41 of genes that were in the 99th percentile for Polynesians. Some examples were *HLA-B*, *PTPRD*, *LIPC*, *SHROOM3*, and *ABO*. *HLA-B* was in common with all the super populations, although only half of the populations of AFR. *PTPRD* was mostly only the four POL populations with a few windows that were in common with a population for either SAS or AMR. *LIPC* had windows in common for all Polynesian populations, and also shared with EAS, SAS, and AMR. *SHROOM3* was in the Eastern Polynesian populations, but also in common with EAS, and to a lesser degree EUR and AMR and AFR. *ABO* was in the Eastern Polynesian populations and also in common with EAS, SAS, and AFR.

4.3.2.4 Zeng's *E*

4.3.2.4.1 Zeng's *E* distributions

Similar to Tajima's *D*, all of the distributions for Zeng's *E* appear to have been shifted to the right, indicating a potential bias against low frequency variants (Figure 4.13 and Table 4.6). The AFR populations had the lowest overall distribution, with the lowest mean (1.850), 1st percentile (-0.096), median (1.838), 99th percentile (3.910), maximum (6.511), and range (8.093). The EUR populations that had the lowest minimum (-1.866) and the highest maximum (8.544), which also meant the largest range (10.410). The highest mean of 2.642 was from EAS. The POL populations had the second highest mean (2.559), median (2.535), 99th percentile (5.194), and maximum (7.827).

Zeng's *E* for both extremes of the distribution (Figure 4.14), clustered all populations except AMR

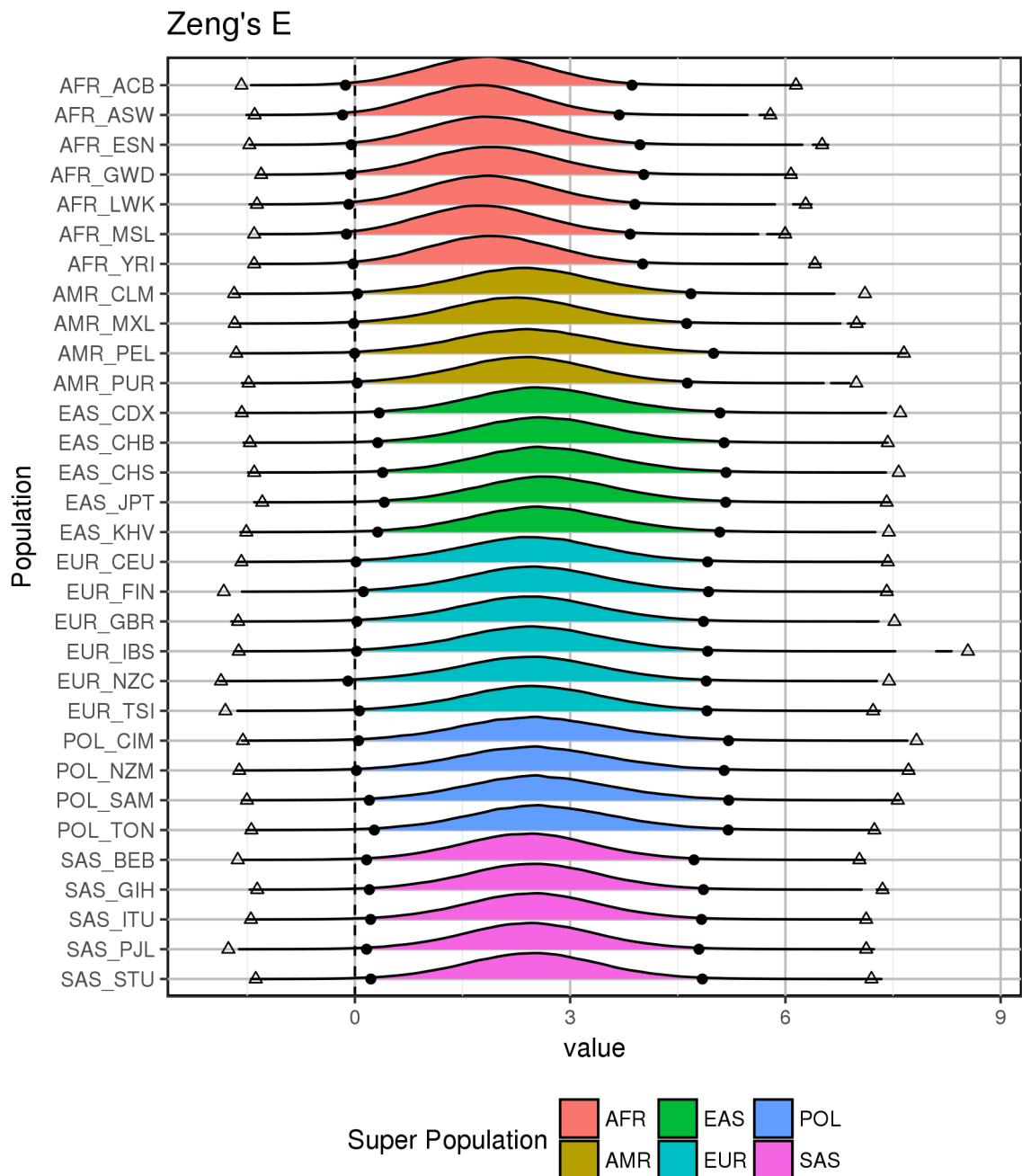


Figure 4.13: Plot of the distribution of Zeng's E by population. Triangles indicate the minimum and maximum values. Dots indicate the 1st and 99th percentiles for each population.

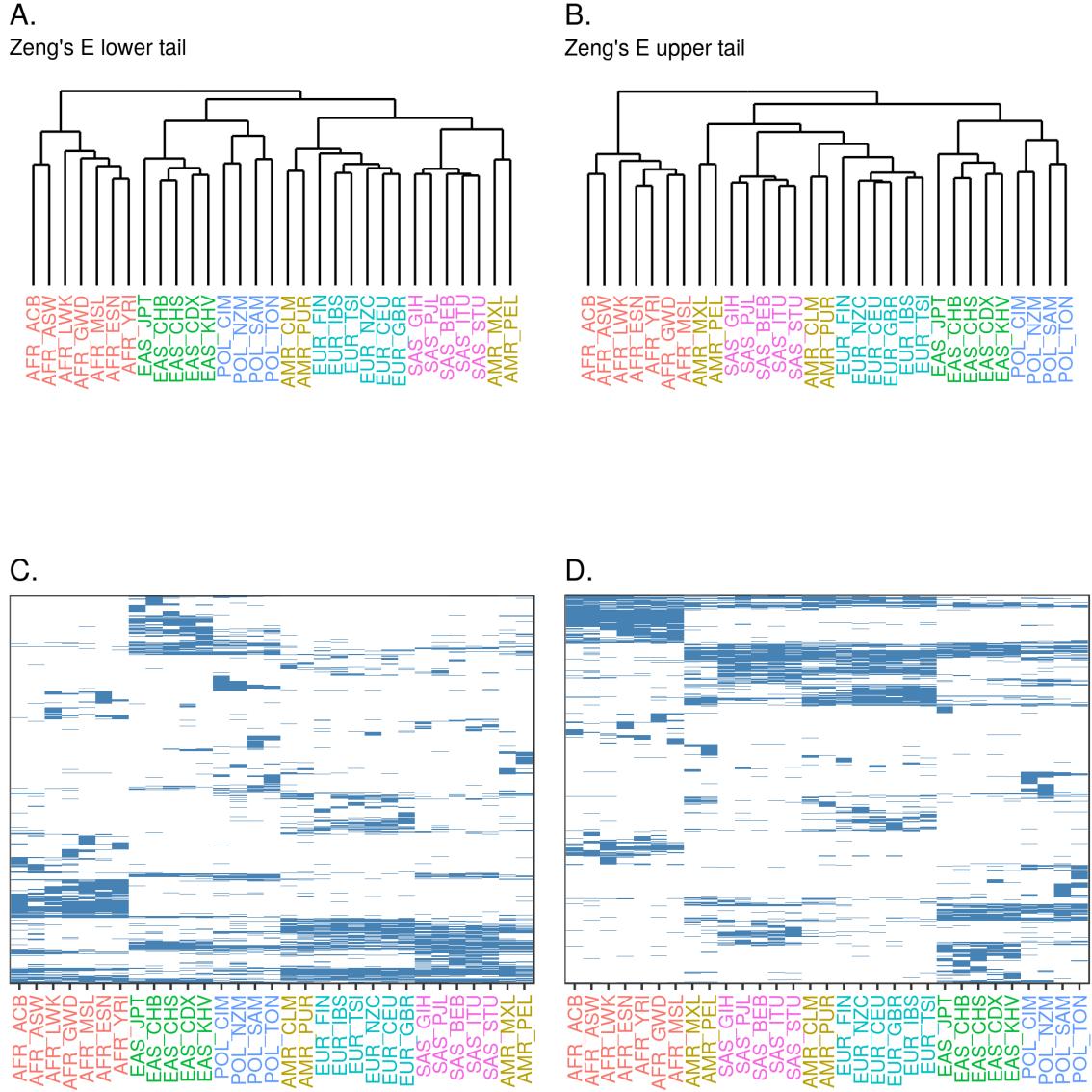


Figure 4.14: Hierarchical clustering of Zeng's *E* using the upper and lower 1% of the distribution. **A** and **B** Dendograms representing the clusters of Zeng's *E* lower and upper tails respectively, using hierarchical clustering with Euclidean distance and complete linkage, and coloured by super population. **C** and **D** Plots representing the windows present in each population from the lower and upper tails respectively. A blue line represents presence of a window, white represents the absence of a window in the 1% of the distribution for a population.

Table 4.6: Summary statistics for Zeng’s *E* by super population.

Super Population	Mean	SD	Min	1st Percentile	Median	99th Percentile	Max
AFR	1.850	0.860	-1.581	-0.096	1.838	3.910	6.511
AMR	2.346	0.994	-1.684	0.012	2.340	4.754	7.652
EAS	2.642	1.001	-1.577	0.352	2.620	5.135	7.602
EUR	2.452	1.024	-1.866	0.024	2.453	4.904	8.544
POL	2.559	1.065	-1.613	0.122	2.535	5.194	7.827
SAS	2.463	0.968	-1.763	0.192	2.458	4.812	7.354

into their respective super populations. The AMR population was split into CLM and PUR which was grouped with the EUR populations, and MXL and PEL which were grouped as a separate exclusive cluster. The groupings of all super populations was the same for both tails, however, in the dendrogram the difference comes from where the AMR populations were relative to the other super populations with the MXL and PUR populations changing slightly in how close to the SAS group they were. Exclusivity of clusters was high with only the AMR -EUR group not being exclusive to a single super population.

4.3.2.4.2 Clustering Zeng’s *E* 1st percentile

There were a total of 14,052 windows, with 1371 windows specific to Polynesian populations in the lower tail. This compared to 3002 windows for AFR, 605 windows for AMR, 976 windows for EUR, and 1149 windows for EAS. There were a total of 444 windows specific to the Eastern Polynesian populations, with 191 specific to CIM and 130 specific to NZM. The Western Polynesian populations had 618 unique windows, 204 were unique to SAM and 250 unique to TON. The mean number of unique windows was 119.5 (SD 64.1) in the lower tail. There were 36 genes that intersected windows that were only in the four Polynesian populations and no other populations (Table S19). None of these genes were associated with urate or co-morbidities.

4.3.2.4.3 Clustering Zeng’s *E* 99th percentile

In the upper tail there were a total of 13,684 windows with 1320 windows specific to Polynesian populations. This compared to 2909 windows for AFR, 584 windows for AMR, 900 windows for EUR, and 1200 windows for EAS. The Eastern Polynesian populations had 467 unique windows, the CIM had 211 unique windows, and NZM had 141 unique windows. The Western Polynesian populations had 594 unique windows. Of those, 204 were unique to SAM and 264 unique to TON. The mean number of unique windows per population was 113.5 (SD 64.3)

4.3.2.4.4 Metabolic associated diseases in the extremes of Zeng’s *E*

There were 34 genes that were in the list of genes of metabolic disease associated genes and also had windows in the 1st percentile, that intersected them from the Polynesian populations. Some examples of these genes are *MACF1*, *AGBL4*, *DACH1*, and *RBMS1*. *MACF1* had windows that were in the 1st percentile of all populations. *AGBL4* and *DACH1* had windows in nearly every population except for

most AFR populations. *RBMS1* was only in the 1st percentile for the Western Polynesian populations SAM and TON. Windows of the 99th percentile in the Polynesian populations that intersected the metabolic disease associated genes tended to also be found in the populations of EAS, and many also were in common with all other populations but not many were in common with the AFR populations. There were 37 genes that were in the 99th percentile for Polynesians. Some examples were *CHST8*, *ERBB4*, *FHIT*, *PTPRD*, and *HMGAA2*. *CHST8* had windows that were in the 99th percentile for the four Polynesian populations, as well as the populations of EAS with a couple of populations of both SAS and EUR too. *ERBB4* had windows that were in both POL and EAS populations. *FHIT* had windows for the Western Polynesian populations, SAM and TON and most of the populations from EAS and SAS. *PTPRD* had windows from the Eastern Polynesian populations of CIM and NZM, these windows were also in common with populations mostly from EUR, while some other windows were in common with the Eastern Polynesian populations and AFR. *HMGAA2* had windows in all of the super populations except AFR.

4.3.2.5 Summary of clustering on extremes regions

Clustering of the windows using that fell in the lower 1% and upper 99% of the distributions for grouped the populations of the POL super population together in all instances except the lower tail of Fu and Li's *F*. There were no windows that were present for a single population in the 1st percentile for all frequency-based statistics. The same applied for the 99th percentile. Pooling the windows for each population within a super population for the 1st percentile, across all intra-population frequency-based statistics used, had 16 genes. The AFR superpopulation had *PHF7*; EAS had *ALMS1*, *OPLAH*, *EXOSC4*, *GPAA1*, *CYC1*, and *SHARPIN*; EUR had *NTN3*; POL had *FANCM* and *ALMS1*; and SAS had *ABHD14A*, *ACY1*, *ABHD14B*, *TBC1D24*, and *NTN3*. EAS and POL had *ALMS1*, a kidney function-associated locus (Pattaro *et al.*, 2016) in common, and EUR and SAS had *NTN3* in common. There were no windows in the 99th percentile that when populations were pooled as super populations were present in all four frequency-based statistics.

4.3.3 Hierarchical clustering of selected haplotypic regions

A similar approach as with clustering of the frequency spectrum-based statistics was used to investigate similarities in regions that had evidence of possible selection in the intra-population haplotypic based selection methods. To do this, the significant iHS or nSL results were clustered into regions (Table S21), and the percentage of bases overlapping between populations was calculated in these regions. The proportion of shared regions, based on region overlap, was used to cluster populations, given the total region that was significant for a single population. This was to investigate if there was a commonality between populations in the haplotypic regions that had evidence for selection, the hypothesis being that populations in similar geographic regions would have had a similar environmental exposure, and therefore have similar regions of the genome under selective pressure. Another possibility was that the selective pressure was in a shared ancestral population.

Clustering on the proportion of shared regions (see section 4.2.5) for both iHS (Figure 4.15) and nSL

(Figure 4.16) grouped the individual populations into their super populations, with the exception of the AMR populations which were split into two separate groups. The order of populations differed due to each population having a different total for the size of region that was significant. Both iHS and nSL gave similar groupings and had similar proportions of selected regions shared.

4.3.3.1 iHS

For the iHS significant regions, the populations with the least regions shared between them were CIM and CLM, with 3.0% shared. The populations that had the largest proportion of region shared was NZC and CEU with 56.7% shared. Looking specifically at the proportion of regions shared between Polynesian populations, for the Eastern Polynesian populations there was NZM with 34.5% of regions shared with CIM, and CIM shared 35.6% of regions with NZM. In the Western Polynesian populations, SAM shared 43.8% of regions with TON, but TON shared 45.3% of regions with SAM. Out of all the significant iHS regions for Polynesian populations, only 1.7% were in common across all Polynesian populations. The Eastern Polynesians accounted for 59.9% of the regions from all of the Polynesian populations while the Western Polynesian populations made up 50.1%.

The proportion of significant iHS regions the Polynesian populations shared with AMR was 14.5%, AFR was 14.5%, EAS was 12.9%, EUR was 10.4%, and SAS was 9.6%. This was lower than the mean (20.3%) for proportion of regions shared between the other super populations. The Polynesian populations were clustered together with a high proportion of shared significant regions. The Eastern Polynesians had on average 35.0% in common, and the Western Polynesian populations had 44.6%. However, shared regions between the Eastern and Western Polynesian populations was 14.2% on average. Of the clustered regions of significant iHS that were in common between all of the POL populations, 70 genes intersected the regions. These genes were part of 57 pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) 2016 pathways in Enrichr, but none were significantly enriched. Regions that were in common between NZM and CIM had 392 genes that intersected, and were involved with 174 pathways. Again, no pathways were significantly enriched. For the regions that were in common between SAM and TON, there were 502 genes that intersected the regions. They belonged to 223 pathways, with only a single pathway significantly enriched (Benjamini-Hochberg adjusted P = 0.023) with seven of 44 genes (*ABCA5*, *ABCC8*, *ABCC5*, *TAP2*, *ABCA9*, *ABCB8*, and *ABCA8*). That pathway was “ABC transporters_Homo sapiens_hsa02010”.

4.3.3.2 nSL

The clustering results using regions of significant nSL gave the same groupings as iHS (Figure 4.16). This similarity was also seen with the differences in individual populations, where NZC and GBR had the highest percentage of shared regions at 59.9%. NZM and Bengali from Bangladesh (BEB) shared the lowest percentage of regions at 0.5%. The percentage of the significant nSL regions the Polynesian populations shared with AMR was 10.0%, AFR was 10.4%, EAS was 11.6%, EUR was 7.6%, and SAS was 7.0%. This was lower than the mean (16.5%) proportion of region sharing between super populations. The Polynesian populations were clustered together with a high proportion of shared

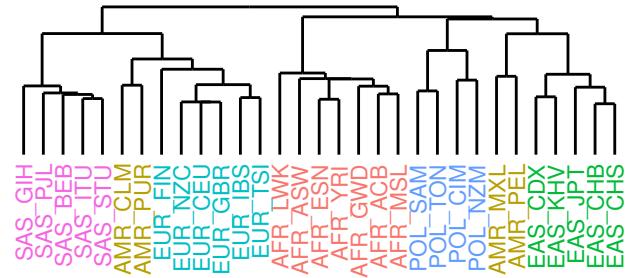
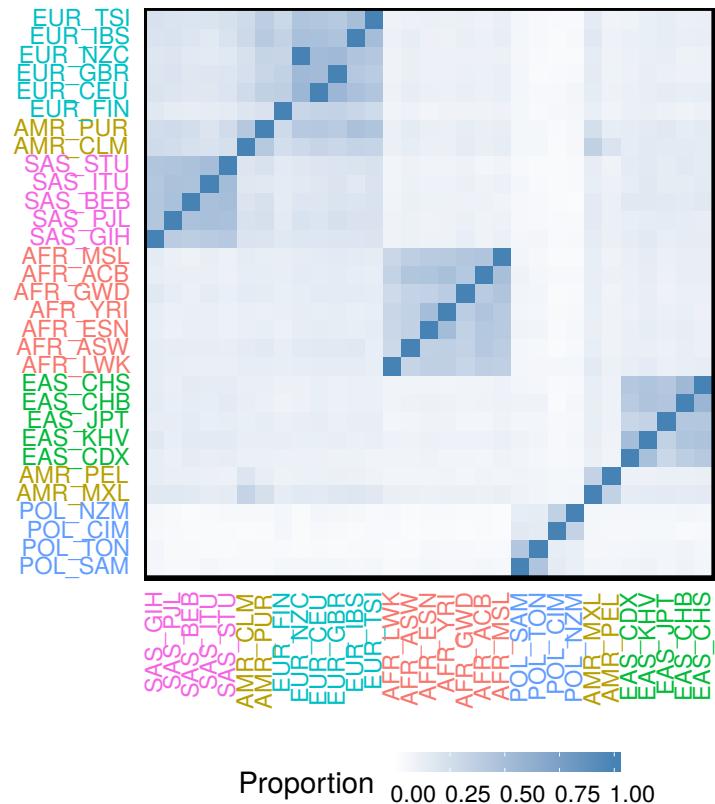
A**B**

Figure 4.15: iHS Clustering: A. Dendrogram from hierarchical clustering based on proportion of significant iHS regions shared between populations, coloured by super population. B. Heatmap showing the proportion of significant iHS regions shared. The population listed on the axis is used as A in the calculation = $(A \cap B)/|A|$. Axes have been ordered according to the hierarchical clustering.

significant regions. The Eastern Polynesians had on average 36.5% in common. NZM shared 38.3% of regions with CIM, whereas CIM shared 34.6% of regions with NZM. The Western Polynesians had 40.4% of regions shared on average. SAM shared 39.1% of regions with TON, but TON shared 41.8% of regions with SAM. However, between the Eastern and Western populations, region sharing averaged 11.2%. Of the regions from all of the Polynesian populations, the Western Polynesian populations made up 56.7%. The Eastern Polynesians made up 51.0%. The total percentage of regions common for all Polynesian populations was 1.4%. This shows that within the Eastern or Western Polynesian groups there was a larger proportion of regions that were similar for significant nSL, than between the Eastern and Western groups.

The regions that were in common for significant nSL between all of the POL populations intersected 27 genes. These genes were part of 10 pathways, but none were significantly enriched. Regions that were in common between NZM and CIM had 230 genes that intersected, and were involved with 160 pathways. A single pathway, “Glycosaminoglycan degradation_Homo sapiens_hsa00531”, was significantly enriched (adjusted P 9.223×10^{-3}) with four of 19 genes (*HPSE2*, *HYAL1*, *HYAL2*, and *HYAL3*). The regions that were in common between SAM and TON intersected 319 genes. There were 197 pathways that these genes were part of but none were significantly enriched.

4.3.3.3 Metabolic disease associated genes

In the clustered-significant regions of both iHS and nSL (Table S21) there were 405 independent regions that overlapped with genes associated with urate and metabolic diseases from all populations, for Polynesian populations the number of regions was 104. iHS had 337 regions, while nSL had 329, and in common between the two statistics were 261 regions. The regions of the Polynesian populations covered 433 SNPs for iHS and 301 for nSL, with 149 in common between them. For the genes that had significant regions in Polynesian populations, 8 regions intersected genes that had been associated with urate and gout, 67 with obesity, 30 with type 2 diabetes, 6 with kidney disease, and 5 with metabolic syndrome.

Using the significant SNPs for these iHS and nSL (as in Chapter 3) there were a total of 204 SNPs that were significant in a Polynesian population and also significant in at least one other population for iHS, and 93 for nSL (Table S3). The number of SNPs that were significant in both iHS and nSL in the Polynesian population was 48.

The urate associated gene *ABCG2* had a single SNP (rs2622626) that was significant in TON that was in common for significance for iHS in two other populations, Gujarati Indian from Houston Texas (GIH) and Indian Telugu from the UK (ITU), both were from the SAS super population. The situation was the same for nSL, but with the addition of JPT.

The SAM and TON populations had multiple SNPs in common that were significant for both iHS and nSL. An example is in *LRP1B*, where there were three SNPs: rs10173806, rs4591293, and rs2890615. Rs2890615 was also significant in ESN and Southern Han Chinese (CHS).

ERBB4 had a single significant SNP, rs6707285, in SAM, and was also significant in PUR, KHV, IBS,

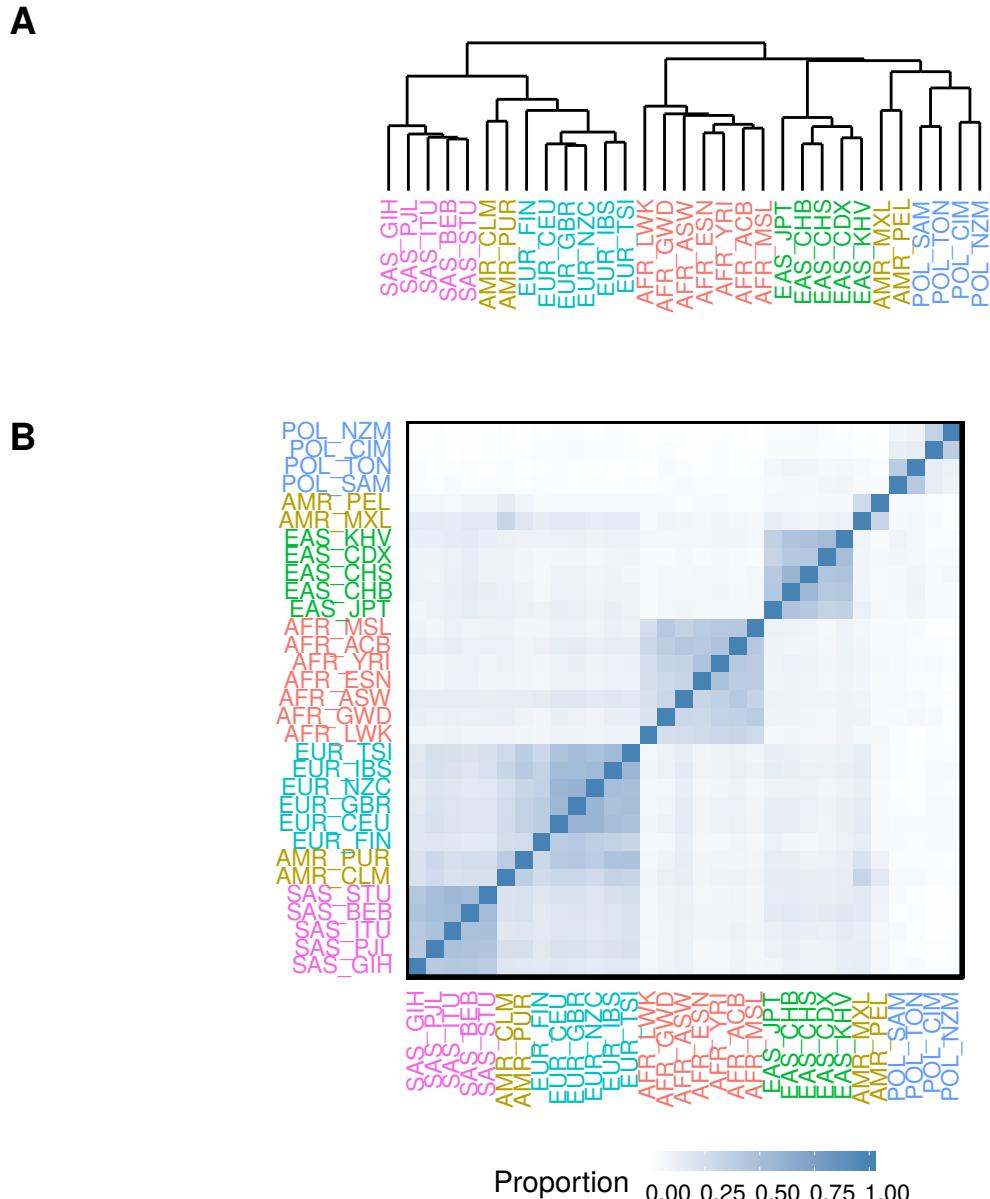


Figure 4.16: nSL Clustering: A. Dendrogram from hierarchical clustering based on proportion of significant nSL regions shared between populations, coloured by super population. B. Heatmap showing the proportion of significant nSL regions shared. The population listed on the axis is used as A in the calculation = $(A \cap B)/A$. Axes have been ordered according to the hierarchical clustering.

TSI, and GBR. *FHIT* also had a single significant SNP, but this was in the NZM, CHB, CHS, CDX, KHV, and PEL populations. Of the metabolic disease genes, *ERBB4* and *FHIT* were also found by Pickrell *et al.* (2009) in the Human Genome Diversity-CEPH Panel using iHS and cross-population extended haplotype homozygosity (XP-EHH).

One of the main regions, chr20:33470694-34556005, had 11 significant SNPs in Polynesian populations as well as up to 16 other populations, contained the obesity-associated gene *GDF5* and the neighbouring genes *CEP250* and *UQCC1*. The majority of the significant SNPs were for the ancestral allele. Both rs1570841 and rs4911502 were significant in the populations of AMR (CLM, MXL and PUR), EAS (CDX, CHB, CHS, JPT, and KHV), EUR (CEU, FIN, GBR, IBS), POL (CIM, NZM, and TON), and SAS (BEB and ITU). Both SNPs were for the ancestral allele. Four SNPs, rs4911178, rs4911494, rs6087704, and rs6087704, intersected both *GDF5* and *UQCC1*, and were significant for the derived alleles in the CIM and JPT populations.

4.3.4 Disease-associated gene clustering

To assess if there was specificity of selection for the metabolic disease gene lists or if the population groupings could be obtained from any list of genes, clustering was performed using the actual statistic values for the genes. These values were median-centred to account for the population specific shifting of distributions which would have accounted for the majority of difference in the distance calculation. Gene lists were chosen for metabolic-related disease and compared to lists of random genes (section 4.3.5.2) to assess clustering specificity to metabolic disease.

4.3.4.1 Clustering of median-centred window values for urate and gout-associated genes

There were 62 loci that had an association with urate or gout at a genome-wide significance level in the GWAS catalog. This mapped to 435 windows for each of the frequency-based selection test statistics. Tables S22, S23, S24, and S25 contain the median centred values for the windows used in the clustering. All statistics (Tajima's *D*, Fay and Wu's *H*, Fu and Li's *F*, and Zeng's *E*) grouped most of the super populations correctly and nearly all clusters had good exclusivity for each single super population (Tables 4.1 and 4.2). Tajima's *D*, Fay and Wu's *H*, and Zeng's *E* also had an Eastern/Western Polynesian population sub group within the Polynesian group (Figure 4.17). Unlike the FST clustering, the AMR populations were grouped together for Tajima's *D*, Fay and Wu's *H*, and Zeng's *E*. The EAS and POL populations were closest for Tajima's *D*, Fu and Li's *F*, Fay and Wu's *H*, and Zeng's *E*.

One of the trends across all of the statistics, was that within a super population, the range of values within a gene were very similar, but between super populations there could be completely different directions. There was also variation between loci, in terms of which populations had overall more positive or negative windows. *INHBB* was an example of a locus with Fay and Wu's *H* that showed similarity in scores of populations with similar ancestry, such as between the POL and EAS which were positive, SAS was near zero but slightly positive, EUR and AMR were near zero but slightly

negative, and AFR was more negative (Table S23). Another example was *RREB1*, where for Tajima's *D*, the POL populations were sitting around -2, with the EAS populations slightly less negative, and the other populations having windows that were around zero (Table S22). A similar pattern was seen in *RREB1* with Fu and Li's *F* (Table S24). With Zeng's *E*, *SLC22A12* and *SLC22A12* both had AFR, EAS, and POL negative, and AMR, EUR, and SAS positive (Table S25). And both *SLC17A1* and *SLC17A4* had all populations with negative Zeng's *E*, except the POL populations, which were positive. *SLC2A9* tended to be overall more negative in value for all statistics for both POL and EAS than for the other super populations.

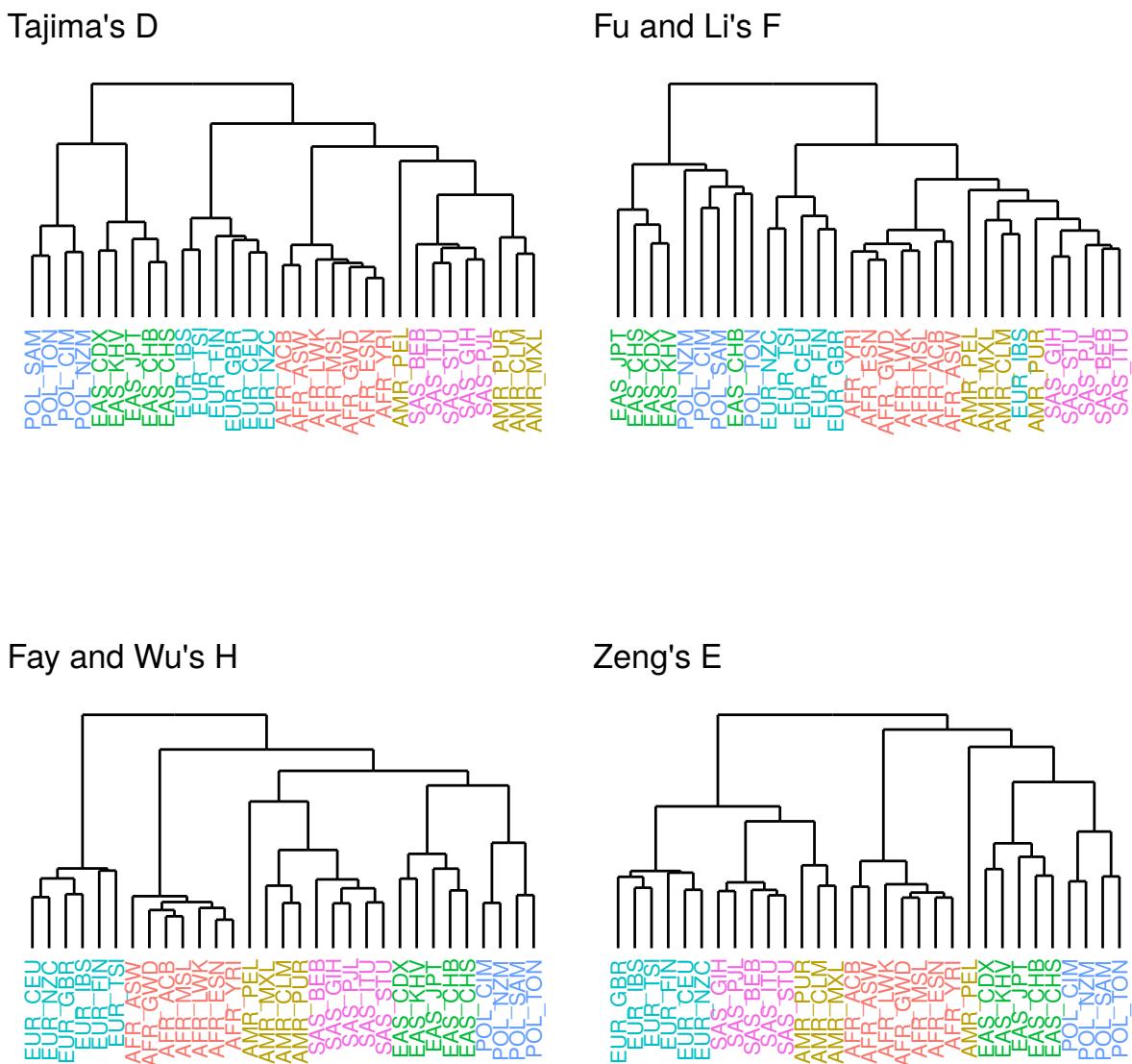


Figure 4.17: Dendrograms created from hierarchical clustering applied to windows from the 62 loci associated with urate and gout. Coloured by super population.

4.3.4.2 Clustering of median-centred window values for obesity-associated genes

The GWAS catalog had 269 loci associated with obesity, this was represented by 4099 windows and the clustering resulted in the groupings of super populations (Figure 4.18). Exclusivity of a single super population to a cluster was also high with the notable exception being the AMR populations of CLM and PUR often being part of the EUR cluster (Tables 4.1 and 4.2). Tables S26, S27, S28, and S29 contain the median centred statistic values for the windows used in the clustering. The Polynesian super population cluster was also closest to EAS in all selection statistics and also had an Eastern/Western sub-grouping within the POL group.

Much like the urate-associated genes, the same trend of within a super population, genes had similar values, but there could be completely different directions of statistic between super populations held. Some examples of this included *LEPR* and *FOXE1*, where for Fay and Wu's *H*, Tajima's *D*, and Fu and Li's *F*, for the EAS and POL populations the windows were mostly negative, whereas for the other populations they were close to zero or even positive. *ARL15* was an example of a locus that all populations had a range of values, but the POL populations was the negative most for all statistics, and in Tajima's *D*, Fay and Wu's *H*, and Fu and Li's *F* the EAS populations showed similarity with the POL populations in the values of the statistics in being overall quite negative.

4.3.4.3 Clustering of median-centred window values for type 2 diabetes-associated genes

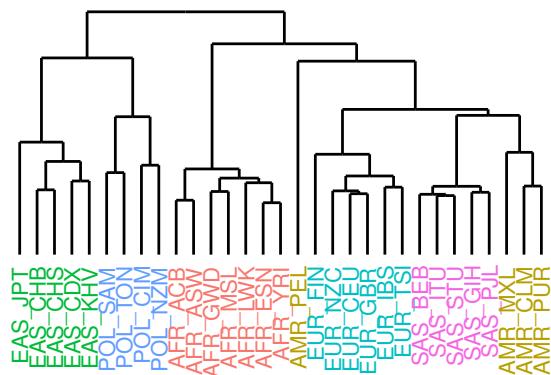
The GWAS catalog had 99 genes associated with type 2 diabetes. These were represented by 1471 windows and clustering of the type 2 diabetes associated genes grouped the individual populations into their super population groups. For all clustering the AMR populations were split, with CLM and PUR grouped with the EUR populations for Tajima's *D*, Fay and Wu's *H*, and Zeng's *E* (Figure 4.19). Tables S30, S31, S32, and S33 contain the median centred statistic values for the windows used in the clustering. The POL populations were assigned two exclusive groups through the cutree method that created an Eastern/Western Polynesian split for Tajima's *D*, Fay and Wu's *H*, Fu and Li's *F*, and Zeng's *E*. Fay and Wu's *H* and Zeng's *E* had a similar pattern to the groupings of the whole chromosome F_{ST} clustering.

Again, most of the loci that were associated with type 2 diabetes followed the same trend of values within a super population being similar, and differences between the super populations. An example where this was not the case was *SSR1*, where all populations, except the Eastern Polynesian populations were very similar and close to zero, whereas the Eastern Polynesian populations were negative in all statistics, but not extreme enough to be in the 1st percentile.

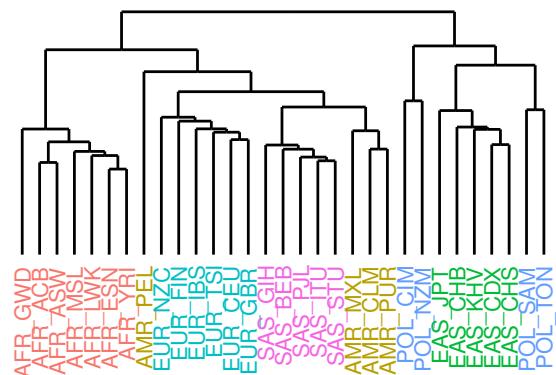
4.3.4.4 Clustering of median-centred window values for kidney disease-associated genes

The GWAS catalog had 53 genes associated with kidney disease, this was represented by 390 windows. The clustering of the kidney disease-associated genes grouped the individual populations into their super population groups for AFR and SAS using Tajima's *D*, Fay and Wu's *H*, Fu and Li's *F*, and Zeng's *E*. Populations from POL and EAS were clustered into separate clusters with Tajima's *D* and

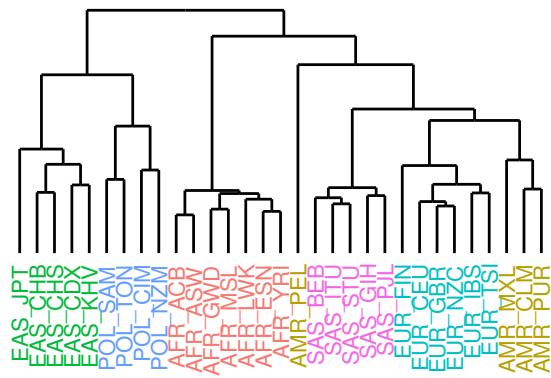
Tajima's D



Fu and Li's F



Fay and Wu's H



Zeng's E

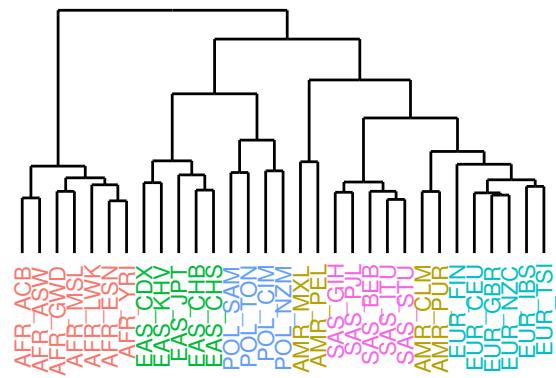
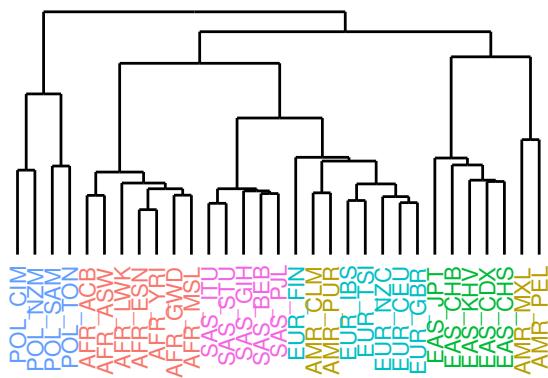
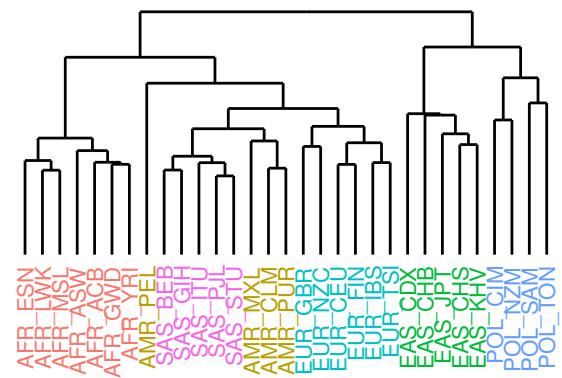


Figure 4.18: Dendograms created from hierarchical clustering applied to windows from the 269 loci associated with obesity. Coloured by super population.

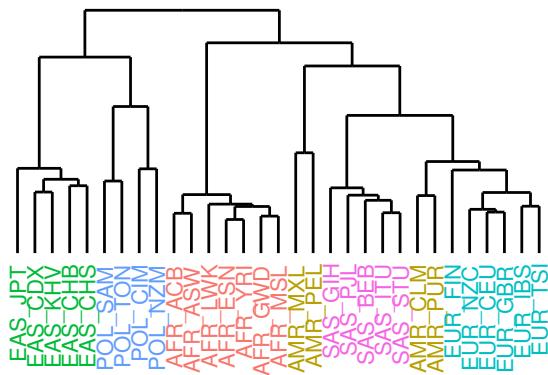
Tajima's D



Fu and Li's F



Fay and Wu's H



Zeng's E

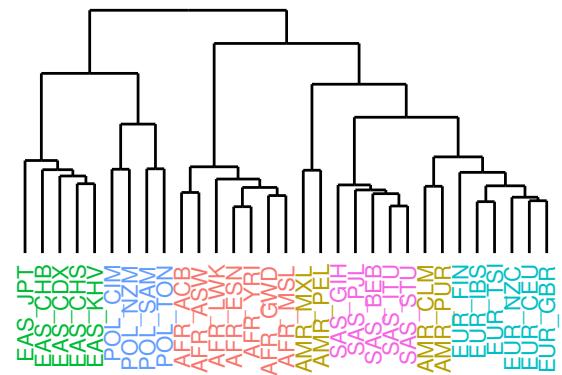
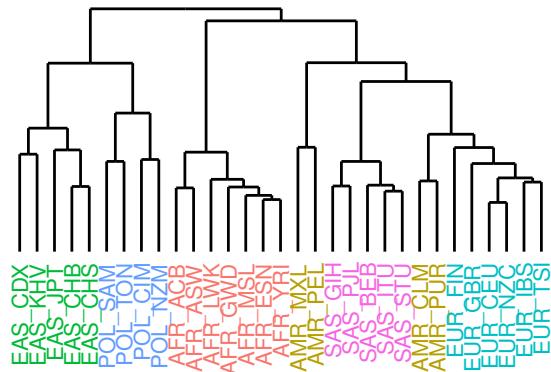
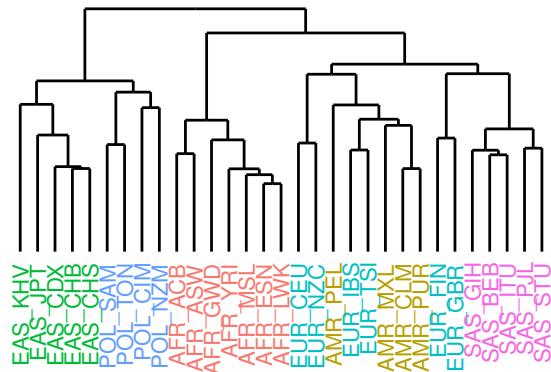


Figure 4.19: Dendograms created from hierarchical clustering applied to windows from the 99 loci associated with type 2 diabetes. Coloured by super population.

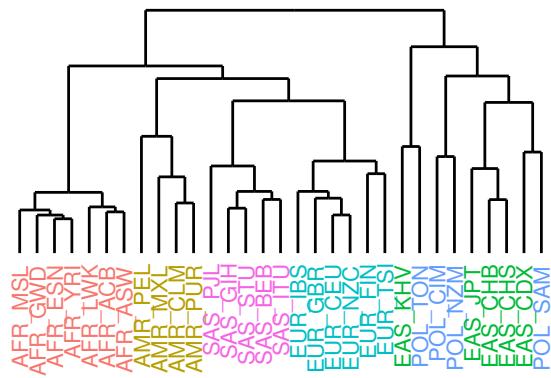
Tajima's D



Fu and Li's F



Fay and Wu's H



Zeng's E

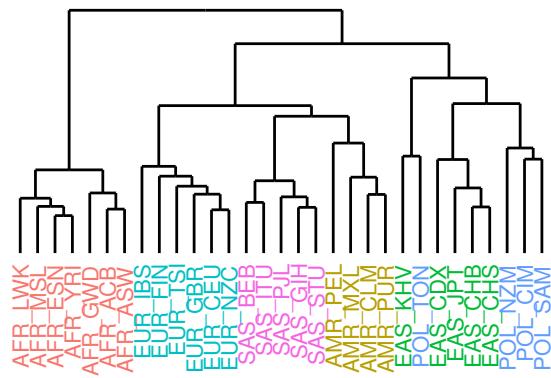


Figure 4.20: Dendograms created from hierarchical clustering applied to windows from the 53 loci associated with kidney disease. Coloured by super population.

Fu and Li's *F* but were clustered together for Fay and Wu's *H* and Zeng's *E* (Figure 4.20). Tables S34, S35, S36, and S37 contain the median centred statistic values for the windows used in the clustering. Tajima's *D* clustered populations from AFR, EAS, POL, and SAS into super population specific and exclusive clusters.

The same trend as for the other gene lists still stood with the kidney disease gene list. Although for Fay and Wu's *H* and Zeng's *E* there was a higher similarity between TON and KHV than with the other populations of their respective super populations. This was largely driven by *ALMS1*, where TON, KHV, and the populations of AFR, were negative, and the other populations were positive for Zeng's *E* and the opposite was the case for Fay and Wu's *H*.

4.3.4.5 Clustering of median-centred window values for metabolic syndrome-associated genes

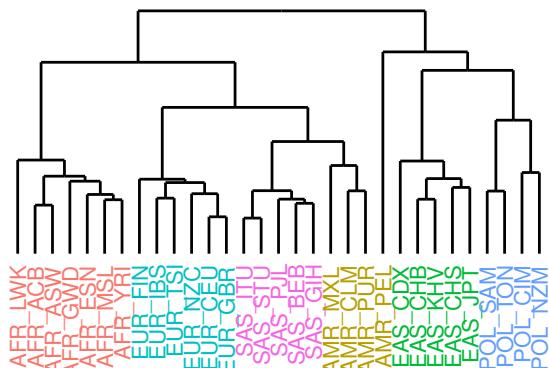
The GWAS catalog had 22 genes associated with metabolic syndrome, this was represented by 234 windows. The clustering of the metabolic syndrome associated genes grouped individual populations into their super population groups for Tajima's *D* and Fay and Wu's *H* (Figure 4.21). Tables S38, S39, S40, and S41 contain the median centred statistic values for the windows used in the clustering. For Fu and Li's *F*, the Polynesian populations were assigned two exclusive clusters that followed the Eastern/Western Polynesian split. Zeng's *E* clustered the individual populations, except for PEL, into their super populations. The Polynesian populations were grouped into entirely exclusive clusters for all statistics. The same trends of within super populations having similar statistic values and differences between super populations still held in the gene list for metabolic syndrome.

4.3.5 Random draws

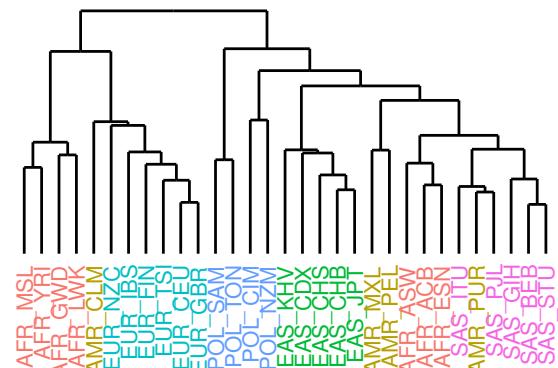
4.3.5.1 Random windows

It is possible that obtaining the population groupings by clustering the most extreme windows for each statistic, could have occurred by chance. To assess this, 10,000 iterations of random resampling of 2500 windows per population was undertaken, and these windows were then clustered. This was done so that the extreme tail clustering could be compared to the random windows to see if the extreme window clustering performed better in exclusivity and proportion of super population clustering than by drawing windows at random. The maximum proportion of populations being clustered into their designated super population was then calculated when the tree was cut to have six groups ($K = 6$), one for each super population. The mean proportion of populations grouped into their super population was less than 0.5 for all super populations, with an overall mean of 0.282 (Table 4.7). This compared to a mean proportion of 0.916 for the clustering of the extreme windows, this implies that there is similarity in the regions of extreme score within a super population, and differences in these regions between super populations. Based on the results of the admixture analysis, $K = 11$ was also tested and had higher mean exclusivity (0.424) than $K = 6$, due to more possible individual clusters. The

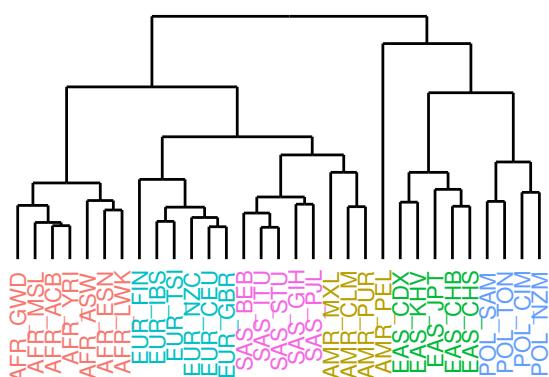
Tajima's D



Fu and Li's F



Fay and Wu's H



Zeng's E

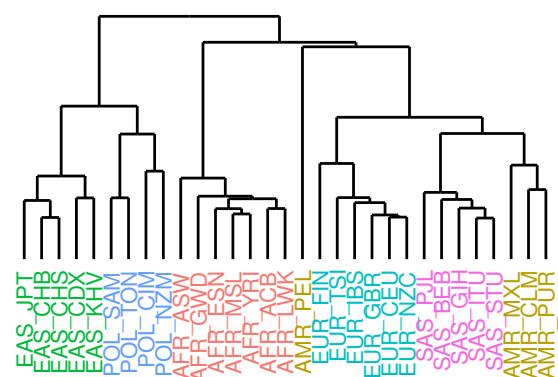


Figure 4.21: Dendrograms created from hierarchical clustering applied to windows from the 22 loci associated with metabolic syndrome. Coloured by super population.

Table 4.7: Proportion of individual populations assigned to their super population by hierarchical clustering across 10000 iterations of 2500 randomly drawn windows.

Super Population	Min	Mean	Std. Dev	Max	Prop50 (%)	Total Complete (n)
AFR	0.286	0.389	0.103	0.857	2.1	0
AMR	0.250	0.451	0.147	1.000	11.9	32
EAS	0.200	0.429	0.118	1.000	3.6	6
EUR	0.167	0.395	0.102	0.833	5.6	0
POL	0.250	0.450	0.147	1.000	11.8	30
SAS	0.200	0.423	0.116	1.000	3.3	2

Prop50 is the percentage of the iterations that had at least 50% of each super population assigned the same cluster. Total complete is the number of iterations all populations of a super population were assigned the same cluster

remainder of this section will report on $K = 6$, as this reflects the number of super populations used for the non-randomised data.

Not a single random draw grouped the all individual populations into all of their respective super populations. Only the populations of the AFR and EUR super populations did not each get fully grouped into a cluster. Grouping at least 50% of a super population was at best 12% of the draws but for the larger super populations (AFR, EAS, EUR, and SAS) it was on average 4% of the draws. There were only eight random draws that had all super populations each having half or more of their individual populations assigned to the same group. The mean exclusivity of the random draws was low, ranging from 0.26 to 0.32 and there were no random draws in which the Polynesian populations were all grouped together as a Polynesian exclusive cluster.

Compared to the clustering on the extreme tails, the random windows had poorer exclusivity of clusters with a mean exclusivity of 28%. The populations were clustered into their super populations rarely for the random windows and did not have multiple super populations being clustered as the clustering of the extremes returned (Table 4.7). This indicates that there is a degree of similarity of regions that are in the extremes of the distributions and the populations who share them. This is largely due a minimum degree of differences in similarity being needed in order for hierarchical clustering to be effective for clustering. When windows are drawn at random the differences, or distances, between populations is likely to be equivalent, so the hierarchy is flatter, meaning small differences have very large changes to the groupings.

4.3.5.2 Random genes

In order to assess whether the clustering from gene lists was specific to the disease associated genes used, lists of random genes were created by selecting genes from a list containing all annotated genes, and then clustered. Random gene lists were generated with either 25 or 100 randomly selected genes, 100 times each. Hierarchical clustering was then performed on the median centred statistic value based on the population, for windows that intersected genes from the random gene lists for Tajima's D , Fay and Wu's H , Fu and Li's F , and Zeng's E . The number of times the Polynesian super population was completely clustered ranged from 5 to 78 (Table 4.8). The Polynesian super population was completely

Table 4.8: Percentage of draws from randomly selected genes for completely clustered super populations.

Statistic	Genes (n)	POL Complete (%)	All Complete (%)	All Complete excl. AMR (%)
Fay and Wu's H	25	57	5	49
Fu and Li's F	25	13	0	0
Tajima's D	25	51	8	44
Zeng's E	25	60	7	52
Fay and Wu's H	100	78	5	75
Fu and Li's F	100	5	0	1
Tajima's D	100	74	3	73
Zeng's E	100	75	6	75

clustered more often with the lists of 100 genes for Tajima's *D*, Fay and Wu's *H*, Fu and Li's *F*, and Zeng's *E* compared to the lists of 25 genes. The number of times all populations were clustered completely was less than 10%. However, if the AMR super population was not included then it became 56 for the 100 genes, and 36.3 for the 25 genes. Fu and Li's *F* did not perform well at clustering the Polynesian populations into their super population, and only achieved this less than 13% of the time. Clustering all populations was not achieved by Fu and Li's *F*. The other frequency spectrum statistics clustered the Polynesian populations into their super population for at least 50% of the time and did better with more genes. Clustering of the other super populations was similar to the Polynesian super population when the AMR populations were not included. This indicates that clustering from disease associated gene lists is not specific to particular genes, but they contain similar discriminatory information as other lists of genes, which is likely to be capturing allele frequency differences between populations. This was evident from the fact that if the AMR populations are excluded, by chance the super populations were correctly clustered at least 50% of the time, and this increased to ~75% when more genes were used (Table 4.8). It also indicates that genes (even when randomly selected) are better at clustering the super populations than randomly drawn windows.

4.4 Chapter discussion

4.4.1 Use of selection and neutrality statistics to group populations

Other studies have previously shown the relationships between Polynesian and other Pacific populations, most often with mitochondrial DNA and Y chromosome markers (Kayser, 2010; Duggan *et al.*, 2014; Hudjashov *et al.*, 2017, 2018). Autosomes have also been used to explore Polynesian origins and relationships between Pacific populations, though generally with rather small sample sizes (Largest Polynesian population size: 24 Kimura *et al.* (2008); 9 Hudjashov *et al.* (2017); 30 Hudjashov *et al.* (2018)). This study is the first to compare four Polynesian populations, two of East, and two of West Polynesian ancestry, and of similar sample sizes to those of the 1000 Genomes Project. The analysis is important because it adds new evidence of the genetic relationship between the Polynesian populations and other populations. This is important, as (in addition to purely academic reasons) understanding the ancestry of populations helps to inform and reduce the stigma of genetic diseases (Sankar *et al.*, 2006), and to inform about the etiology of modern metabolic disease.

In Chapter 3 (section 3.4), it was noted that the populations had population specific-distributions for the frequency spectrum-based selection and neutrality statistics. The tails of the distributions contain the most extreme values and are therefore most likely to be the loci that have undergone selection - specifically the negative tail is of interest for positive selection. It was thought that perhaps the tails could be used to group the individual populations using an unsupervised method - in this case, hierarchical clustering - and reassemble the super populations. The first test used the entire chromosome and created a summary statistic that was used to create clusters. It was successful in re-assembling the super populations using F_{ST} which is a measure of population differentiation, and had previously been used within the 1KGP dataset to establish how differentiated the 1KGP samples were from each other (1000 Genomes Project Consortium, 2015). The population groupings were also the same as in the PCA that was used to select the individuals for the selection analysis (Figure 4.1). Both the F_{ST} results and PCA (for the subject selection) confirmed the genetic relatedness and differentiation of the Polynesian populations in a global context, as being most similar to the East Asian populations (Kayser *et al.*, 2008). It also confirmed that within Polynesia, the differentiation between Eastern and Western Polynesia, was consistent with the out of Africa human migration and subsequent Polynesian expansion (Matisoo-Smith, 2015; Skoglund *et al.*, 2016; Nielsen *et al.*, 2017; Hudjashov *et al.*, 2018).

4.4.1.1 Clustering of aggregate statistics

The whole chromosome summary using the intra-population frequency spectrum neutrality and selection statistics had varying levels of success in re-assembling the super populations. Fu and Li's F had mini-groups of 2-3 populations that were part of the same super population, but these mini-groups were not closely linked to the other groups of the same super population. These two statistics are based on singletons, and it is likely that these are not well represented in the markers of the CoreExome chip itself, or the quality control that the New Zealand samples underwent. During the quality control

process, a singleton was more likely to be removed, or less likely to be rescued, upon manual inspection depending on the predefined cluster boundaries for the probe intensities in Genome Studio. This would lead to an under-representation of the singletons that actually might be present in the populations. Overall, it was found that the clustering of selection and neutrality test statistics for entire chromosomes did not resemble the grouping that clustering on F_{ST} produced.

4.4.1.2 Clustering of extreme windowed statistics

Compared to the entire chromosome summaries, using windows from the extremes of the distribution was effective at clustering the individual populations into their super population groups. Fay and Wu's H was the only statistic that grouped all individual populations into their corresponding super populations, whereas Tajima's D and Zeng's E had the same split of the AMR populations as with the F_{ST} clustering. The clustering of populations into their super populations was not seen when random windows were used for clustering, indicating that there is a degree of commonality in the extreme windows within a super population group. This would mean that there was similarity in the windows, and therefore regions of the genome that were represented in the extremes of the distributions between the populations of a super population. It also suggests that because these regions differ between super populations that these regions may have been influenced from the out of Africa migration and subsequent population movements. The haplotypic based methods of iHS and nSL clustered into super populations and had the AMR population split too. There was not a high amount of region sharing outside of each super population, again showing similarity within super populations but not between.

Allele sharing between populations is moderate for common SNPs, and more so for continental groups, but is minimal for rarer frequencies, due to population divergence (Gravel *et al.*, 2011). This helps to explain both of the patterns observed with site frequency spectrum (SFS) based statistics for both the extremes and gene clustering. With the extremes, many of the super populations had windows that were not in the extremes of the other populations, these formed the 'blocks' in the heatmap, showing that there were regions of the genome that differed in extremity between super populations. This was most evident with the AFR populations, which had the majority of windows for each statistic not in common with the extremes from the other populations.

The extremes of the distributions represent different types of selection or demographic events and are different depending on the selection test statistic being used (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Zeng *et al.*, 2006). The upper end of the distribution is able to be compared as all of the values that it encompassed were positive in value across all populations for all statistics and represent the same types of selection or demography.

The clustering of the most extreme windows represents, other than a similarity in allele frequencies between populations, a pattern that is consistent with the out of Africa model of human migration (Mallick *et al.*, 2016; Pagani *et al.*, 2016). And represents the influence of selection, drift, population expansions, migrations, bottlenecks, introgression, and admixture following the population divergences over history.

4.4.1.3 Clustering on genic regions

Clustering from gene lists for the frequency spectrum-based selection and neutrality statistics consistently grouped the individual populations into their super populations, although there was inconsistency for the AMR populations. This inconsistency could be the influence of admixture in the AMR populations, between the PCA analysis and the admixture analysis (sections 4.3.1.1 and 4.3.1.2), it was clear that the four AMR populations were less homogenous than the other super populations. With the exception of the AMR populations, the clustering of populations using windows or haplotypes was consistent with previously reported results (Pickrell *et al.*, 2009; 1000 Genomes Project Consortium, 2015). The grouping of populations was also seen with random lists of genes. This would indicate that the commonalities with populations for a gene are not specific to a disease, but instead, a general locus by locus similarity.

4.4.2 Potential shared selective histories for loci associated with urate and metabolic disease

The “Out of Africa” model of human migration can be useful in explaining differences between modern populations and potential time points when these differences may have arisen. Under a neutral model the differences in allele frequency between populations are due to genetic drift (Kimura, 1979b). Demographic events such as bottleneck, migration, expansion, and admixture also contribute to the differences in allele frequencies found between populations (Wright, 1951). Conversely, selection acts in a targeted manner, therefore if the selection occurred post population split then it would be expected to look neutral in one population but not in the other. However, if the selective event happened in an ancestral population of the two then, dependent on time since divergence, it would be expected to be observed in both populations (Hermisson and Pennings, 2017). It is also possible to have ‘parallel’ selection, where both populations experience independent selection (Tennessen and Akey, 2011). Using a coalescent approach to identify the common intermediate ancestral populations, if for example, a genomic region identified as under selection in the Polynesian populations was also found in the East Asian populations, but not the Europeans, it could be inferred that the selective pressure was likely after the migration from Europe to East Asia. In looking at modern disease, a potential approach to understanding the differences in disease burden between modern populations is to look at the genetic differences between populations with shared ancestry.

The current metabolic disease burden in the Polynesian populations is higher than other populations (Winnard *et al.*, 2012, 2013). This higher disease burden could possibly have arisen due to environmental changes and selective pressure on Polynesian populations after the migration into the Pacific. As found in Chapter 3, however, there was limited evidence for selection, especially in the overlap between haplotypic and SFS-based methodologies, in the genes that were associated with urate and associated metabolic diseases (Table 3.6). Three approaches were taken to investigate the potential of a ‘shared selective’ history of the loci associated with urate and metabolic disease. These were: investigating the loci that were in either the 1st or 99th percentile, investigating the loci that were in the clustered regions for the significant SNPs of the haplotypic statistics, use of iHS and nSL, and finally, hierarchical

clustering of the median-centred values for windows that intersected the loci associated with urate and metabolic disease.

For the SFS-based statistics, many of the windows in either the 1st or 99th percentile for the POL populations that intersected the genes associated with metabolic disease were not in common with other populations. The windows that were in common, were usually with the EAS populations. This is consistent with the migratory history of the Pacific and the shared ancestry the Polynesian populations have with East Asia.

4.4.2.1 Clustering on median-centred gene values

The clustering based on the median-centred values of the statistics was able to reassemble the super population groups. Again, the similarity of the allele frequency spectrum-based on geography appeared through urate and metabolic disease genes. Both exclusivity and the proportion of populations clustered correctly were very high for nearly all populations for the four site-frequency statistics, across the metabolic disease gene lists (Tables 4.1 and 4.2). This indicated that there was sufficient similarity between populations of a super population, and differences between super populations that the statistic values were discriminatory. However, the groups were also reassembled with the random draws of genes, with an increase in the number of completely grouped super populations when more genes were used. This showed that the particular genes themselves were not specific to the clustering but the frequency distributions in genes were different and the number of genes influenced the discrimination of populations during clustering.

The simplest explanation of why similar clustering occurred with the gene lists is that with the tests of neutrality or selection that rely on the frequency spectrum, populations that are similar geographically and/or ancestrally, also have a similar allele frequency to each other (Gravel *et al.*, 2011). This would explain why, with Tajima's D , Fay and Wu's H , and Zeng's E , which are based on comparing ratios of low, intermediate and high frequency variants, all were clustered into their super population groups.

4.4.2.2 Clustering of extreme value windows

The differences in the extremes of the distributions for the site-frequency spectrum statistics showed that they could be used to cluster the populations into their super populations. The differences and commonalities in the genes that had evidence of association with a metabolic disease, in the 1st percentiles of the statistics for the Polynesian populations, were almost exclusively isolated to only the Polynesian populations. The 1st percentile represents the values with the highest likelihood of being under selection. The windows that were in common with other populations were, for the most part, only found in the populations of EAS. The windows or SNPs that were significant in the Polynesian populations within the metabolic disease gene lists were usually limited to commonality with the paired ancestrally similar population (CIM and NZM, or SAM and TON). Windows were more likely to be in common between populations for the 99th percentiles where this represented an increase in high frequency variants.

4.4.2.3 Clustering of haplotype-based statistics

For the haplotypic statistics, there were commonalities between some of the significant SNPs. The SNPs that were in common were more often in pairs of either SAM and TON, or CIM and NZM. The SNPs that were significant for a Polynesian population were more likely to have also been significant in an EAS population than the other super populations. The regions created from the clustering of significant SNPs also showed that within the super populations there was a higher commonality than between super populations. This would suggest that, similar to Coop *et al.* (2009), geographically similar populations are subjected to similar selective environments. The time to the most recent common ancestor of two populations also affects the detection of shared regions as the more time that has passed, the more recombination will have an opportunity to break down the haplotype.

4.4.3 Limitations

The technical limitations of this dataset based on the markers from the CoreExome SNP array raised in section 3.4.5 also apply to this analysis. Particular to the analysis in this chapter is how the clusters were determined. The `cutree` method implemented in R (R Core Team, 2017) that was used to slice the dendograms, creating the groups used to assess the proportions and exclusivity of the clustering was relatively crude and did not always agree with the clusters that were ‘visually apparent’ in the dendrogram, although ‘visually apparent’ could be partially attributed to confirmation bias. Hierarchical clustering was chosen as the clustering algorithm over K-means because the branching nature could be thought of as a metaphor for the population migration histories. It is also possible that different linkage and distance criteria for the clustering algorithm may have produced different clusters. However, a brief investigation of this, using each of the distance and linkage measures implemented by the `hclust` method in R indicated that any differences in clusters would likely be minor (data not shown).

The `cutree` method also required that K clusters be returned regardless of what the tree actually looked like, and $K < n$. The use of $K = 6$ was chosen due to there being six super populations, however, CLM, PUR, MXL, and PEL being classed as a single super population was not reflected in the clustering of the chromosome-wide F_{ST} . Many of the selection tests split AMR into 2 groups, CLM/PUR and MSL/PEL. This pairing is consistent with what had previously been found from similarity of admixture and F_{ST} (Gravel *et al.*, 2013; 1000 Genomes Project Consortium, 2015). These AMR populations had also been identified as being admixed and so it was interesting that for the upper tail of Fu and Li’s F they were clustered with the NZM population, which is also known to be admixed, mostly with New Zealand European (Hollis-Moffatt *et al.*, 2012; Gravel *et al.*, 2013). Because these AMR populations were still included, and the requirement for six clusters, this affected the exclusivity measurement of clusters. The proportion measurement was also affected because sometimes a super population was spread across multiple but exclusive clusters. An example of this was the Polynesian populations and the 99th percentile of Fu and Li’s F where the Eastern Polynesian populations were an exclusive cluster, and the Western Polynesian populations were an exclusive cluster. This meant more clusters were assigned to Polynesian populations than should have been (under the $K = 6$ assumption

there would be one cluster per super population), and for six clusters to remain, the SAS populations ended up combined with the EUR populations, based on the distance and linkage criteria used for the clustering.

The only statistic that consistently did not cluster the Polynesian populations was Fu and Li's F . Fu and Li's F is based on singletons, and for the clustering, the 1st percentile windows was a sparse matrix of windows that were in common between populations. The sparsity of the matrix influenced how well the clustering performed due to the lack of windows that were in common between any pair of populations. At the extremes of similarity (e.g., very similar or very dissimilar), the clusters are disproportionately influenced by only a few windows, which means a large amount of information was not being used for the differentiation of populations. As the lower tail represents an excess of singletons, this is likely to have been impacted by using SNP array data as there was a very strong bias against singletons during the quality control protocol (Guo *et al.*, 2014). In this situation, because the data has reduced singletons, the distribution for Fu and Li's F was shifted to the right, with the upper tail representing a deficit of singletons (Fu and Li, 1993). This shift now means there is an under representation of regions in the lower tail.

As reported in section 3.3.1, only the false discovery rate (FDR) for the Polynesian populations was below 10% for the 1st percentiles of Tajima's D and Fay and Wu's H for the frequency spectrum-based statistics. This means that while the extremes would be considered the most likely ends of the distribution that would contain regions of the genome that had been selected, this cannot be claimed with confidence. Therefore, the mechanism through which they became the extremes of the distribution is possibly through selection, however, other neutrality mechanisms are still valid. While it is true that the extremes may include selected loci, it is also true that because of the use of a distribution proportion threshold, there will always be windows or SNPs that would be identified as possibly being under selection (Teshima *et al.*, 2006). This means that while there were similarities in the extremes of the distributions for the frequency-based statistics, they are not necessarily under selection, and based on the FDR calculated from the permutations, they are likely to be false positives.

Another limitation was a lack of ‘internal replication’, for example, the Eastern and Western Polynesian pairs, where a region would only meet the threshold for significance in a single population and not in the paired population. This could also just be a side effect of discretising a continuous distribution, and especially for values sitting close to the threshold it is easy to imagine a situation where in one population it was significant and the other it was not. An alternative approach could have been to rank all of the windows based on statistic value and determine differences and similarities in rank order between populations.

The rightwards shift in the distributions of Tajima's D , Fu and Li's F , and Zeng's E largely is influenced by the low frequency variants and singletons being under-represented. This effect is due to the ascertainment bias on the SNP arrays, where there is an excess of common alleles, when compared with sequence data (Ramírez-Soriano and Nielsen, 2009). This shift was more noticeable in the AFR populations with the 1st percentile being above zero for Tajima's D and Fu and Li's F . This meant that the windows that were being compared were not always the going to be inferring the same type of selection or population demographic events as those that were from below zero. The rightward

shift would not have affected all genes in a uniform manner - the impact will be different for different genes, based on the frequency of the markers they contain. This means that it cannot be assumed that the genes in the 1st percentile for the AFR populations are the equivalent in terms of ranking by statistic value, to those in the other populations. Nor can it be assumed that the distribution that was observed using the markers from the CoreExome would be the same that would be observed from using sequencing data.

Grouping based on the clustering of the significant results in iHS or nSL may falsely link SNPs into a ‘region of significance’. The 200 kb limit was the maximum distance that selscan by default would allow the calculation of the extended haplotype heterozygosity to extend across a gap of information before terminating. The calculation of region sharing also does not take into account the directionality of the selection. For example, one population might have favoured the ancestral allele, whereas a second favoured the derived allele, the region sharing calculation treats these as being the same and would consider them a shared region.

4.4.4 Conclusions

It can be concluded that the extremes of the distributions for the intra-population selection and neutrality statistics can be used to group populations into their super populations. There was similarity in the regions in the extremes of the selection and neutrality statistics for the populations of a super population, but limited commonality in the extreme regions between super populations. Random regions of the genome had varying levels of commonality between populations but was lower than the regions with extreme selection and neutrality statistic values.

Genes that are associated with disease performed well at grouping populations into their super population groups with hierarchical clustering. However, this was not specific to metabolic disease associated genes, but applied to random sets of genes, with an increase in number of genes improving the clustering. This suggests the result was not specific to the loci used, and instead was due to allele frequency similarities between populations.

In all combinations of clustering of selection and neutrality statistics used, the East Asian populations were the most similar to the Polynesian populations. There was additional evidence of the differences and similarities of the Polynesian populations, between the Eastern and Western Polynesian's, which reflected the migration history . And there was evidence of regions having a shared ancestry between these populations, along with evidence of a common ancestral population with the modern East Asian populations.

Due to the limitations of the source dataset, additional analysis using whole genome sequence data would be beneficial.

Chapter 5

Selection and Association Studies

This chapter investigates the performance of different gout case definitions on a genome-wide association study (GWAS). It also looks at how selection statistics could be used for the prioritisation of variants from a GWAS.

A subset of these results were published as part of Cadzow *et al.* (2017); the gout definition testing, heritability of gout, and the replication of the Köttgen *et al.* (2013) results for gout, using the UK Biobank data. The results and discussion from that paper have been included and expanded on here.

5.1 Genetic associations with Gout

5.1.1 Performance of gout definitions

Genome-wide association studies provide a statistical framework on a genomic scale for an association analysis between a phenotype of interest, and the genotypes of a genetic marker, generally using linear regression for a continuous trait, or logistic regression for a dichotomous trait (Lee *et al.*, 2011; Zuk *et al.*, 2012). The association test, in a GWAS context, is repeated for all available markers across the genome. In epidemiological studies the use of an accurate case definition is important (Olijhoek *et al.*, 2007), however, there is variation in the gout definitions used from the multi-purpose cohort studies that are used for genetic analysis. Multi-purpose cohort studies are cohorts that are collected in a way that enables the use of the data to answer many different research questions; some well known cohorts include the Wellcome Trust Case-Control Consortium (WTCCC) cohort, the Atherosclerosis Risk in the Community (ARIC) study, and the Framingham Heart Study (FHS) (Investigators, 1989; Burton *et al.*, 2007; Splansky *et al.*, 2007).

The use of a consistent case definition is also important for genetic analyses using multiple cohorts. If the definition is consistent then it is possible to combine the genetic data from each cohort and performing the GWAS, allowing for increased sample size and therefore increased power. Another common way of combining genetic association studies is through meta-analysis. In both instances,

having the same case definitions across studies reduces the variation. The power of genetic case-control studies increases with accurate case definition. The use of gold-standard case definitions means that there are the maximum number of genuine cases, and the maximum number of disease-free controls (Colhoun *et al.*, 2003). For gout, the gold standard case definition is the presence of mono-sodium urate crystals in the synovial fluid of the joint (Wallace *et al.*, 1977), measurement of which is not possible for multi-purpose cohorts. A substitute for crystals in the synovial fluid is the American College of Rheumatology (ACR) criteria (Wallace *et al.*, 1977). These criteria require a minimum of six of twelve conditions to be met for an individual to be classed as a ‘case’. It is uncommon for multi-purpose cohorts to report either the gold-standard, or ACR criteria, but instead, for a self-report diagnosis of conditions and diseases to be reported.

The UK Biobank is a publicly available data set consisting of 500,000 individuals, mostly of European ancestry, from around the United Kingdom, aged between 40 and 69. It contains genetic, health, and lifestyle information. As part of the UK Biobank collection process, participants filled in detailed surveys on health and lifestyle questions, provided biological samples, and provided access to administrative health records. The UK Biobank includes the self-report diagnoses that were collected by survey at the time of recruitment, but it also has hospital admission information, for both primary and secondary diagnosis. This collection of both self-report information combined with hospital admission information provides a unique opportunity in a large cohort to investigate possible gout definitions and their influence on GWAS.

5.1.2 Genetic associations for gout in Polynesian populations

As has been previously mentioned (sections 1.7.2, 4.1), New Zealand Polynesians have a higher prevalence of gout and metabolic disease (Winnard *et al.*, 2013). Of the 28 genetic variants associated at a genome-wide significance threshold with serum urate levels that have been previously reported in European populations, 17 also had significant associations with gout (Köttgen *et al.*, 2013). In the Polynesian populations of New Zealand, association with gout has been replicated for nine of these variants, with some Polynesian-specific effects being noted (Phipps-Green *et al.*, 2016). The inherent higher prevalence of gout in Polynesian populations, combined with a lack of adequate sample sizes to perform GWAS with sufficient power to detect loci (other than the main effect loci of *ABCG2* and *SLC2A9*), means that alternative approaches are needed to find additional evidence to reduce the number of potentially false negative results. Traditionally meta-analysis of GWAS has been used to boost statistical power, or trans-ancestral meta-analysis can be used to incorporate differing linkage disequilibrium (LD) patterns between ancestries when performing meta-analysis (Morris, 2011).

5.1.3 Use of selection statistics to inform GWAS

GWAS in the last few years have utilised larger and larger cohorts, in a quest to discover the source of ‘missing heritability’ (Visscher *et al.*, 2012; Yang *et al.*, 2017). The requirement for larger cohorts comes from the inverse relationship between effect-size and statistical power. The source of missing

heritability is thought to be found in the polygenic effects of many small effect genetic variants and unmarked uncommon variants of stronger effect sizes (Spencer *et al.*, 2009). GWAS also suffer from the need to control the false positive rate and as such have a stringent significance threshold for association (Fadista *et al.*, 2016). One of the side-effects of this threshold is that variants that have a true effect but due to cohort size do not reach statistical significance are lost in the ‘noise’ of the GWAS signal. One of the possibilities to be able to identify such variants is to use a combination of selection statistics with GWAS to prioritise GWAS results (Ayodo *et al.*, 2007; Casto and Feldman, 2011; Field *et al.*, 2016a).

5.1.4 Objectives

The objectives of this chapter are to:

- Test performance of gout definitions on GWAS.
- Investigate the use of selection statistics in conjunction with GWAS.

5.2 Methods

5.2.1 European Gout GWAS

This GWAS was performed using the interim data release from the UK Biobank (approval number 12611) downloaded November 2015. There was genetic information for 152,249 individuals. For the GWAS itself, it was limited to be only people who had self-reported an ethnic background of British, Irish, or ‘any other white background’ to limit the effect of genetic ancestry confounding the analysis.

Inclusion criteria were: European ethnicity, aged between 40 and 69 years, and having genome wide genotypes available. The exclusion criteria used were: mismatch between self-reported sex and genetic sex, genotype quality control failure, relatedness, or either a primary or secondary hospital diagnosis of kidney disease (International Classification of Diseases, Tenth Revision (ICD-10), codes I12, I13, N00-N05, N07, N11, N14, N17–N19, Q61, N25.0, Z49, Z94.0, Z99.2), participants aged 70 years and over, and those with kidney disease, because these are risk factors for secondary gout.

Individuals from the UK Biobank cohort were genotyped on an Axiom array with 820,967 markers (Affymetrix, Santa Clara, CA, USA). The genotypes were phased using SHAPEIT3, and then imputed using IMPUTE2 with both the UK10K impute reference panel¹, and the haplotype reference consortium impute panel (McCarthy *et al.*, 2016), to bring the total number of single nucleotide polymorphisms (SNPs) to approximately 73.3 million. Genotyping, phasing, and imputation were performed by the UK Biobank.

A replication cohort was created from the full UK Biobank genetic dataset by using the same criteria as for the interim release but with the participants from the interim release excluded. Genotypes for

¹<https://www.uk10k.org>

these individuals were downloaded in March 2018. The number of individuals in this replication cohort was 261132.

5.2.1.1 Gout definitions

Four main classification criteria were used to define gout cases. *Hospital defined gout* involved either a primary or secondary hospital diagnosis for gout (ICD-10 code M10, including sub-codes). *Self-report of gout* was defined as reporting having gout at the time of the study interview with a nurse, in response to the question “In the touch screen you selected that you have been told by a doctor that you have other serious illnesses or disabilities, could you now tell me what they are?”². *Use of urate lowering therapy (ULT)* used self-report data for the use of allopurinol, febuxostat, or sulphapyrazone, and not having a hospital diagnosis of leukaemia or lymphoma (ICD-10 codes C81-C96). *Winnard defined gout* was based on Winnard *et al.* (2012) and consisted of a hospital diagnosis of gout, or self-report of gout specific medication (ULT or colchicine). The largest number of people were classified through either self-report or the use of urate lowering therapy (Figure 5.1). Additional exclusion criteria were applied to participants not classified as having gout from these definitions in order to remove possible cases from the control cohort. The criteria were cortico-steroidal use, non-steroidal anti-inflammatory drug use or probenecid use. The UK Biobank replication cohort used the gout definition of self-reported gout or self-reported ULT use to define cases.

Confidence intervals for proportions of each gout definition out of the combination of all gout definitions were calculated using the `prop.test` function from the *stats* package in R.

5.2.1.2 Gout association test

A GWAS was performed using logistic regression in Plink1.9b (Chang *et al.*, 2015; Purcell and Chang, 2015) for each gout classification, with age, sex, and body mass index (BMI) as co-variates. The association results for each GWAS were then compared with the 30 SNPs reported in Table 1 of Köttgen *et al.* (2013) that had previously been associated with urate and gout in a cohort of >140,000 Europeans.

5.2.1.3 Heritability

For each gout classification, the genetic component of heritability explained was calculated. This was done by first randomly selecting 10,000 controls and then combining this same subset with cases from each of the gout classifications. The genetic variance was calculated for each chromosome separately and then combined. The dichotomous case-control phenotype was transformed onto a continuous liability scale using the restricted maximum likelihood analysis function in the GCTA software (v1.26.0, Yang *et al.* (2011b)) and a general population prevalence of gout of 2% (section 2.1.6). The liability scale operates on a liability threshold model whereby an unseen continuous trait has a threshold above

²<http://biobank.ctsu.ox.ac.uk/crystal/docs/Interview.pdf> accessed 28 May 2018

which one is a case (Lee *et al.*, 2011; Zuk *et al.*, 2012). Heritability estimates were compared using the formula $h_1 - h_2$ and the standard error (SE) calculated using formula (5.1).

$$SE = \sqrt{SE_1^2 + SE_2^2} \quad (5.1)$$

5.2.2 Polynesian Gout GWAS

A multi-cohort gout dataset was assembled by combining individuals from the Genetics of Gout, Diabetes, and Kidney Disease in Aotearoa cohorts and the Ngāti Porou Hauora Charitable Trust cohort. Genotyping on the Infinium CoreExome v24 bead-chip was performed at the University of Queensland (Centre for Clinical Genomics) for the Genetics of Gout, Diabetes, and Kidney Disease in Aotearoa cohorts and at AgResearch (Invermay Agricultural Centre) for the Ngāti Porou Hauora Charitable Trust cohort. The genotype quality control procedures described in section 2.2.2.1 were applied per cohort, and then as necessary after combining the cohorts. The genotype quality control, and combining of cohorts were performed by Dr Tanya Major, Merriman Lab. Principal component analysis (PCA) was performed on the combined dataset with the first 10 components outputted and used as covariates to adjust for population substructure in the GWAS.

The GWAS for gout using individuals with Polynesian ancestry was conducted by Dr Tanya Major (Merriman Lab) using Plink v1.9b32. Gout affection was determined by the ACR criteria (Wallace *et al.*, 1977), or a doctor diagnosis, or if enrolled in a gout drug trial. All participants were older than 18, and provided informed written consent. Exclusion criteria included unknown gout affection status, non-Polynesian self-reported ancestry, or a mismatch between self-reported ancestry and principal component analysis (PCA) clustering. The GWAS was adjusted for age, sex, and the first ten principal components (PCs) calculated from 2,858 ancestry informative markers (as identified by Illumina and used in Guo *et al.* (2014)). Markers were removed if the minor allele frequency was less than 0.01, or had a Hardy-Weinberg Equilibrium (HWE) chi-squared exact test $P < 1 \times 10^{-6}$.

5.3 Results

5.3.1 UK Biobank

5.3.1.1 Participant and association model characteristics

There were genome-wide genotype data available for 105,421 participants, after applying the inclusion and exclusion criteria. The mean BMI for all participants was 27.36 kg/m^2 , the mean age was 56.97 years, and 50.82% were male. A breakdown by affection status for anthropomorphic traits, drug use, and gout classifications is provided in Table 5.1.

Prior to performing the GWAS analyses using the different classifications for gout, the best model (determined by Akaike information criterion (AIC) (Akaike, 1974)) for phenotype co-variate adjustment

was determined by logistic regression. Known co-variates for gout included age, sex, and BMI. Each of these co-variates were tested individually and together. Each covariate had a significant association with gout on their own: age ($P = 8.72 \times 10^{-88}$), sex ($P = 8.26 \times 10^{-247}$), and BMI ($P = 1.84 \times 10^{-271}$). In the combined model, with all three covariates there was an AIC for the model of 19558.30 compared to single variable models for age (AIC 22706.05), sex (AIC 21041.80) and BMI (AIC 22001.03). Based on the lower AIC for the combined model, age, sex, and BMI were used as the covariates in the GWAS for gout.

5.3.1.2 Performance of gout classification criteria

There were a total of 2432 individuals that had any classification of gout and 102,989 controls. Of those with any classification of gout, there were 382 with a hospital diagnosis, 1652 with urate lowering therapy, 1861 with the Winnard criteria, and 2066 with self-reported gout. The overlaps in classification between each group can be seen in Figure 5.1. There was considerable overlap between the self-report, urate lowering therapy, and Winnard classifications with 1242 individuals meeting all three classifications. There were 571 (27.6%) of 2066 participants who could only be defined through self-report. And for hospital diagnosis, 126 (33.0%) of 382, did not meet the self-report of gout, or ULT use definition criteria.

The prevalence of gout from all definitions was 2.4%. This ranged from the lowest prevalence of gout of 0.36%, from the hospital diagnosis definition, through to the highest of 2.18%, by self-report of gout or ULT usage in the study population (Table 5.2). The UK Biobank replication cohort had a total of 5065 cases and 256,067 controls.

Table 5.1: Clinical details of participants in the UK Biobank.

	Case	Control	Replication	
			Case	Control
n	2432	102989	5065	256067
Male (%)	97.88	52.94	98.22	54.75
Mean Age (SD)	60.18 (6.67)	56.9 (7.93)	59.88 (6.96)	56.94 (7.98)
Mean BMI (SD)	30.82 (4.89)	27.28 (4.7)	30.64 (4.97)	27.13 (4.61)
Gout medication				
Febuxostat	0	0	0	0
Allopurinol	1651	0	3633	0
Sulphinpyrazone	9	0	22	0
Colchicine	63	0	135	0
Gout classification				
Hospital	382	0	1381	0
Self-report	2066	0	4553	0
Winnard	1861	0	3932	0
ULT	1652	0	3610	0

Mean age is reported in years. BMI is reported in kg/m²

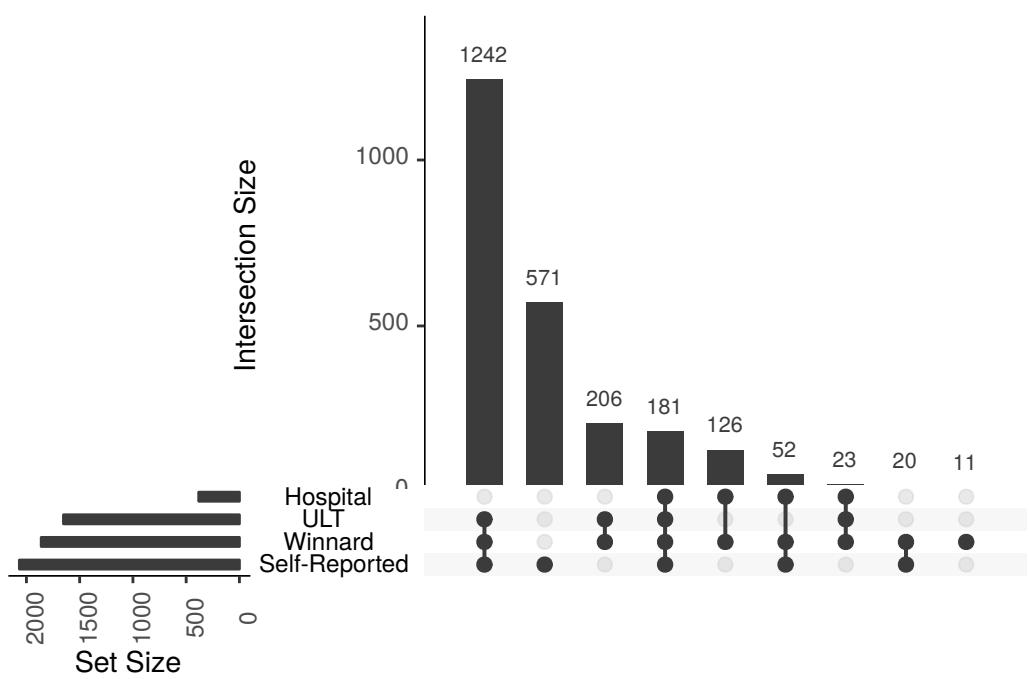


Figure 5.1: An “upset plot” showing the number of samples (left) and intersections (upper) for the different gout classification criteria.

Table 5.2: Number, prevalence (95% CI) of participants defined as gout cases

Definition	No. of subjects, prevalence (95% CI) in entire study population (n = 105,421)	No. of subjects, prevalence (95% CI) in male participants (n = 51,844)	No. of subjects, prevalence (95% CI) in female participants (n = 53,577)	Percentage (95% CI) of study population with gout using any definition (n = 2432)
Self-report of gout diagnosis	2066, 1.96% (1.88-2.05)	1921, 3.71% (3.55-3.87)	145, 0.27% (0.23-0.32)	84.95% (83.45-86.34)
Self-report of gout or ULT use	2295, 2.18% (2.09-2.27)	2122, 4.09% (3.92-4.27)	173, 0.32% (0.28-0.38)	94.37% (93.36-95.23)
ULT use	1652, 1.57% (1.49-1.64)	1529, 2.95% (2.81-3.1)	123, 0.23% (0.19-0.27)	67.93% (66.02-69.77)
Winnard definition	1861, 1.77% (1.69-1.85)	1707, 3.29% (3.14-3.45)	154, 0.29% (0.24-0.34)	76.52% (74.77-78.18)
Hospital diagnosis	382, 0.36% (0.33-0.4)	346, 0.67% (0.6-0.74)	36, 0.07% (0.05-0.09)	15.71% (14.3-17.23)

5.3.1.3 Replication of Köttgen *et al.* (2013) urate associated loci

Analysis of the 30 reported urate-associated SNPs from Köttgen *et al.* (2013) revealed similar odds ratios (ORs) for all gout definitions (Figures 5.2 and 5.3, Table 5.3). Experiment-wide significance was defined as $P < 0.0017$, and genome-wide significance as $P < 5 \times 10^{-8}$ (Fadista *et al.*, 2016). There were differing numbers of SNPs that had a significant association with gout, both at genome- and the experiment-wide thresholds between the gout definitions used. Meeting genome-wide significance, there were five SNPs (one in each of *ABCG2*, *GCKR*, *SLC17A3*, *SLC22A12*, and *SLC2A9*) for both the self-reported gout or ULT usage definition and the self-reported gout definition, four SNPs (*ABCG2*, *GCKR*, *SLC22A12*, and *SLC2A9*) for the ULT usage definition, three SNPs (*ABCG2*, *GCKR*, and *SLC2A9*) for the Winnard definition, and two SNPs (*ABCG2* and *SLC2A9*) for the hospital diagnosis definition. For all definitions, the effect size and strength of association was larger for *ABCG2* than *SLC2A9*. At the experiment-wide significance threshold, there were 13 SNPs for the self-reported gout or ULT use definition, 12 SNPs for the self-reported gout definition, 11 for ULT use definition, 10 for the Winnard definition, and three SNPs for the hospital diagnosis definition.

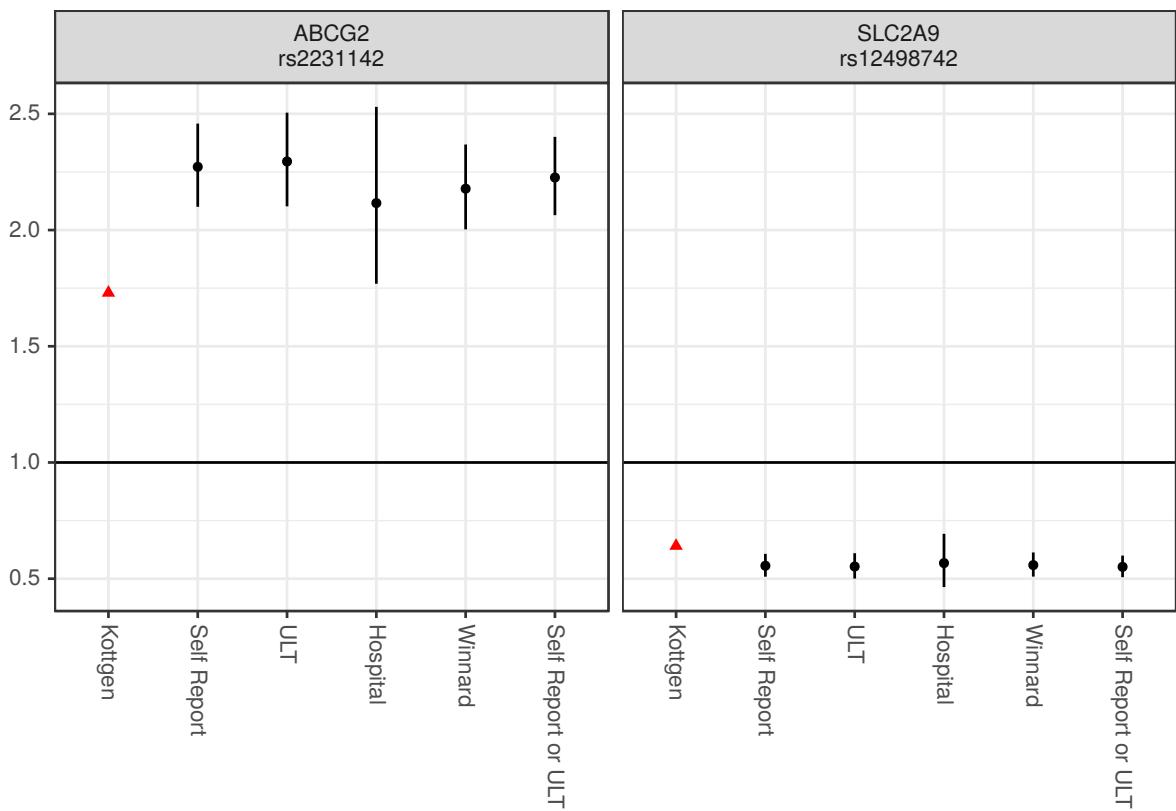


Figure 5.2: Plot showing odds ratios (95% CI) for the two largest effect SNPs reported in Köttgen *et al.* (2013) based on gout definitions, rs2231142 and rs12498742. Original Kottgen reported odd ratio are shown as red circles. The reported Kottgen odds ratio was inverted to maintain consistency of reported effect allele (A1) for rs12498742.

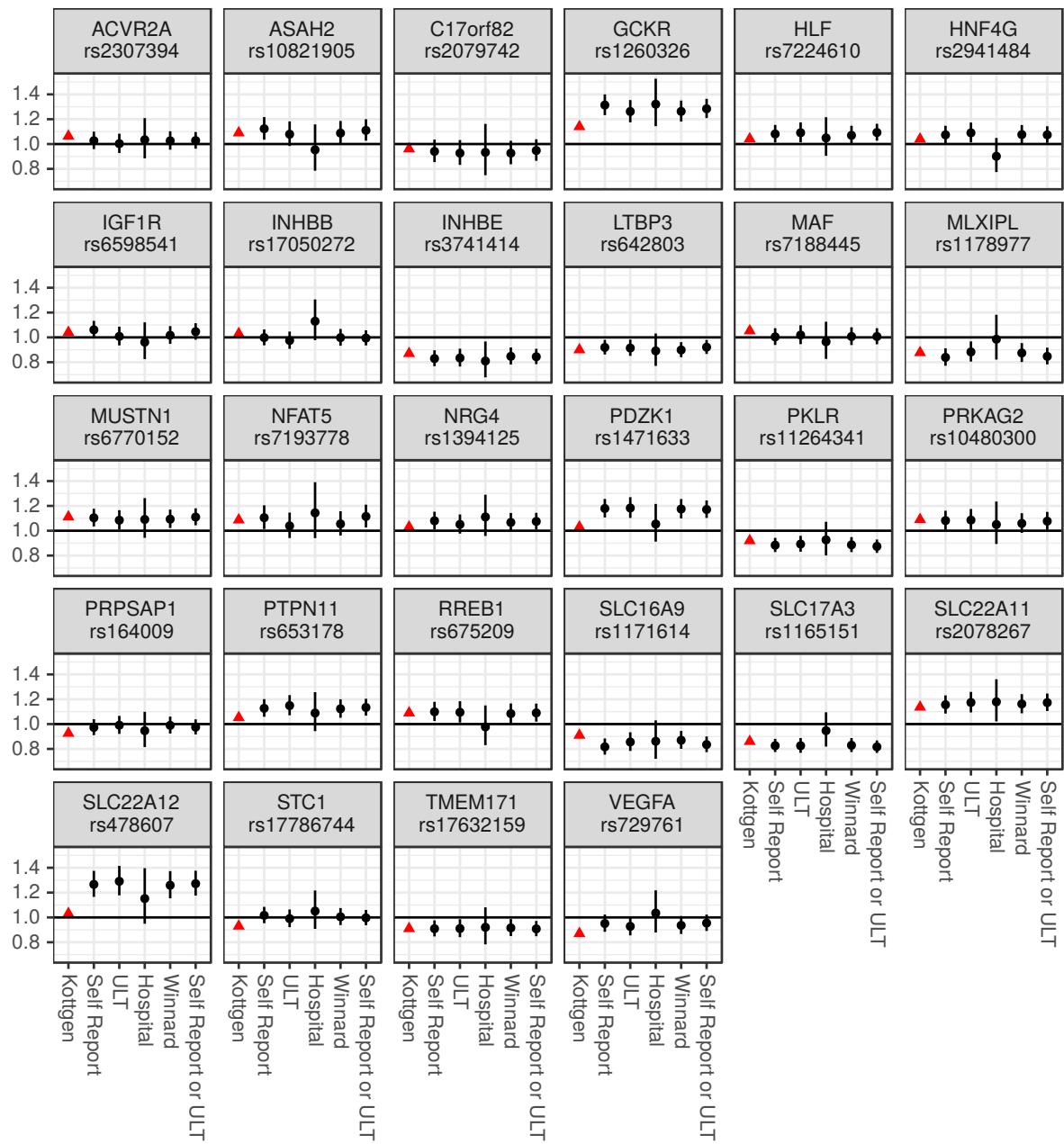


Figure 5.3: Plot showing odds ratios (95% CI) for the 28 SNPs reported in Kottgen *et al.* (2013) based on gout definitions. Original Kottgen reported odds ratios are shown as red circles. The reported Kottgen odds ratio was inverted to maintain consistency of reported effect allele (A1) for the following markers: rs2307394, rs2079742, rs7224610, rs7188445, rs1178977, rs6770152, rs7193778, rs164009, rs653178, rs2078267 and rs478607.

5.3.1.4 Heritability explained

To see if there was a difference between gout classifications in the genetic component of the variance explained by an additive model in gout, heritability estimates were calculated for each gout classification. To do this, 10,000 controls were randomly selected and were used with each different gout criteria, and heritability estimates for each chromosome were calculated using GCTA 1.26.0 (Yang *et al.*, 2011b). The heritability estimates were 0.289 (SE 0.034) for the self-report of gout or ULT use definition, 0.283 (SE 0.036) for the self-report of gout definition, 0.282 (SE 0.040) for the Winnard definition, 0.308 (SE 0.044) for the ULT use definition and 0.236 (SE 0.160) for the hospital diagnosis definition. There were no statistically significant differences between the heritability estimates for any of the different classifications. The heritability estimates for gout were similar to the estimates of genetic variance of serum urate previously reported (Köttgen *et al.*, 2013).

Table 5.3: The odds ratios for the 30 variants reported in Köttingen *et al.* (2013) for the different gout classification methods.

Grail gene	Marker	A1	A2	Hospital	Self-report	Self-report or ULT	ULT	Winnard
<i>ABCG2</i>	rs2231142	T	G	2.12 [1.77 - 2.53]	2.274 $\times 10^{-16}$	2.27 [2.10 - 2.46]	1.059 $\times 10^{-92}$	2.23 [2.06 - 2.40]
<i>ACVR2A</i>	rs2307394	C	T	1.03 [0.88 - 1.21]	0.672	1.03 [0.96 - 1.10]	0.438	1.03 [0.91 - 1.10]
<i>ASAH2</i>	rs10821905	A	G	0.95 [0.79 - 1.16]	0.629	1.12 [1.03 - 1.22]	0.005	1.11 [1.03 - 1.20]
<i>C17orf82</i>	rs2079742	C	T	0.93 [0.75 - 1.16]	0.536	0.94 [0.85 - 1.04]	0.213	0.95 [0.87 - 1.04]
<i>GCKR</i>	rs1260326	T	C	1.32 [1.14 - 1.53]	1.525 $\times 10^{-4}$	1.31 [1.23 - 1.40]	2.709 $\times 10^{-17}$	1.28 [1.21 - 1.36]
<i>HLF</i>	rs7224610	C	A	1.05 [0.90 - 1.22]	0.526	1.08 [1.01 - 1.15]	0.020	1.09 [1.03 - 1.16]
<i>HNF4G</i>	rs2941484	T	C	0.90 [0.77 - 1.05]	0.179	1.07 [1.01 - 1.15]	0.032	1.07 [1.01 - 1.14]
<i>IGF1R</i>	rs6588541	A	G	0.96 [0.82 - 1.12]	0.616	1.06 [0.99 - 1.13]	0.082	1.05 [0.98 - 1.11]
<i>INHBB</i>	rs17050272	A	G	1.13 [0.98 - 1.31]	0.099	1.00 [0.94 - 1.06]	0.957	0.99 [0.94 - 1.06]
<i>INHBE</i>	rs3741414	T	C	0.81 [0.68 - 0.97]	0.020	0.83 [0.77 - 0.90]	2.032 $\times 10^{-6}$	0.84 [0.78 - 0.91]
<i>LTBP3</i>	rs612803	T	C	0.89 [0.77 - 1.03]	0.121	0.92 [0.86 - 0.98]	0.010	0.92 [0.87 - 0.98]
<i>MAF</i>	rs7188445	A	G	0.96 [0.83 - 1.13]	0.648	1.00 [0.94 - 1.07]	0.905	1.01 [0.94 - 1.07]
<i>MLXIPL</i>	rs178977	G	A	0.99 [0.82 - 1.18]	0.872	0.84 [0.77 - 0.91]	3.768 $\times 10^{-5}$	0.85 [0.78 - 0.92]
<i>MUSTN1</i>	rs6770152	G	T	1.09 [0.94 - 1.20]	0.241	1.10 [1.03 - 1.18]	0.003	1.11 [1.04 - 1.18]
<i>NFAT5</i>	rs7193778	C	T	1.14 [0.94 - 1.39]	0.178	1.10 [1.01 - 1.21]	0.023	1.11 [1.03 - 1.21]
<i>NRG4</i>	rs1394125	A	G	1.11 [0.96 - 1.29]	0.164	1.08 [1.01 - 1.15]	0.022	1.07 [1.01 - 1.14]
<i>PDK1</i>	rs1471633	A	C	1.05 [0.91 - 1.22]	0.477	1.18 [1.11 - 1.26]	3.034 $\times 10^{-7}$	1.17 [1.10 - 1.24]
<i>PKLR</i>	rs11264341	T	C	0.93 [0.80 - 1.07]	0.306	0.88 [0.83 - 0.94]	1.644 $\times 10^{-4}$	0.87 [0.82 - 0.93]
<i>PRKAG2</i>	rs10480300	T	C	1.05 [0.89 - 1.24]	0.553	1.08 [1.01 - 1.16]	0.029	1.08 [1.01 - 1.15]
<i>PRPSAP1</i>	rs164009	G	A	0.95 [0.82 - 1.10]	0.472	0.97 [0.91 - 1.04]	0.409	0.98 [0.92 - 1.04]
<i>PTPN11</i>	rs653178	C	T	1.09 [0.94 - 1.26]	0.245	1.13 [1.06 - 1.20]	1.861 $\times 10^{-4}$	1.13 [1.07 - 1.20]
<i>RREB1</i>	rs675209	T	C	0.98 [0.83 - 1.15]	0.783	1.10 [1.02 - 1.18]	0.008	1.09 [1.02 - 1.17]
<i>SLC16A9</i>	rs1171614	T	C	0.86 [0.72 - 1.03]	0.102	0.82 [0.75 - 0.88]	5.899 $\times 10^{-7}$	0.83 [0.77 - 0.90]
<i>SLC17A3</i>	rs1165151	T	G	0.95 [0.82 - 1.06]	0.462	0.83 [0.77 - 0.88]	5.706 $\times 10^{-9}$	0.82 [0.77 - 0.87]
<i>SLC22A11</i>	rs2078267	C	T	1.18 [1.02 - 1.36]	0.024	1.16 [1.08 - 1.23]	6.868 $\times 10^{-6}$	1.17 [1.10 - 1.25]
<i>SLC22A12</i>	rs478607	G	A	1.15 [0.95 - 1.40]	0.152	1.27 [1.17 - 1.38]	2.110 $\times 10^{-8}$	1.27 [1.18 - 1.38]
<i>SLC24A9</i>	rs12498742	G	A	0.97 [0.46 - 0.69]	3.014 $\times 10^{-8}$	0.56 [0.51 - 0.61]	5.529 $\times 10^{-39}$	0.55 [0.51 - 0.60]
<i>STC1</i>	rs17786744	G	A	1.05 [0.91 - 1.22]	0.503	1.02 [0.95 - 1.08]	0.613	1.00 [0.94 - 1.06]
<i>TMEM171</i>	rs17632159	C	G	0.92 [0.78 - 1.08]	0.313	0.91 [0.85 - 0.98]	0.009	0.91 [0.84 - 0.97]
<i>VEGFA</i>	rs729761	T	G	1.03 [0.88 - 1.22]	0.682	0.95 [0.88 - 1.02]	0.175	0.95 [0.89 - 1.02]

OR [95% CI], P, ULT = Urate lowering therapy. The reported effect allele is A1. Effects are adjusted by age, sex, and BMI.

5.3.2 GWAS results

5.3.2.1 UK Biobank gout definitions GWAS

The total number of genome-wide significant SNPs was 397 for the hospital definition, 2044 for the self-reported gout definition, 1396 for the ULT use definition, 1411 for the Winnard definition, and 2310 for the self-reported gout or ULT use definition. There was considerable overlap between the definitions in the SNPs that met the nominal significance threshold of $P < 1 \times 10^{-5}$ (Figure 5.4). There were five main regions of the genome that had peaks of association with gout. These were at *GCKR* on chromosome 2, *SLC2A9* and *ABCG2* on chromosome 4, the *SLC17A1-SLC17A3* region on chromosome 6, and the *SLC22A11-SLC22A12* region on chromosome 11 (Figure 5.5). Only the peaks at *SLC2A9* and *ABCG2* reached genome-wide significance under the hospital definition.

There were a total of ten top SNPs from seven genes that were associated with gout at a genome-wide significance threshold, across the definitions used (Table 5.4). The top SNP for each locus differed between definitions for *GCKR* and *SLC2A9*, and *SLC22A12*. *GCKR* had the top SNP (rs780093) for self-report of gout, self-report of gout or ULT use, or ULT, whereas the top SNP for the Winnard definition was rs780094. The top SNP was rs9994216 in *SLC2A9* for the self-reported gout definition, and rs4697701 for the ULT usage, self-report of gout, and the Winnard definitions. *SLC22A12* had rs11231837 for the top SNP for the self-reported gout and self-reported gout or ULT use definitions, whereas rs7929627 was the top SNP for the ULT usage, and Winnard definitions. None of the top SNPs were the same as reported by Köttgen *et al.* (2013) in Table 1. This can largely be explained by the top SNPs reported here being associated with gout, whereas Köttgen *et al.* (2013) Table 1 was for urate.

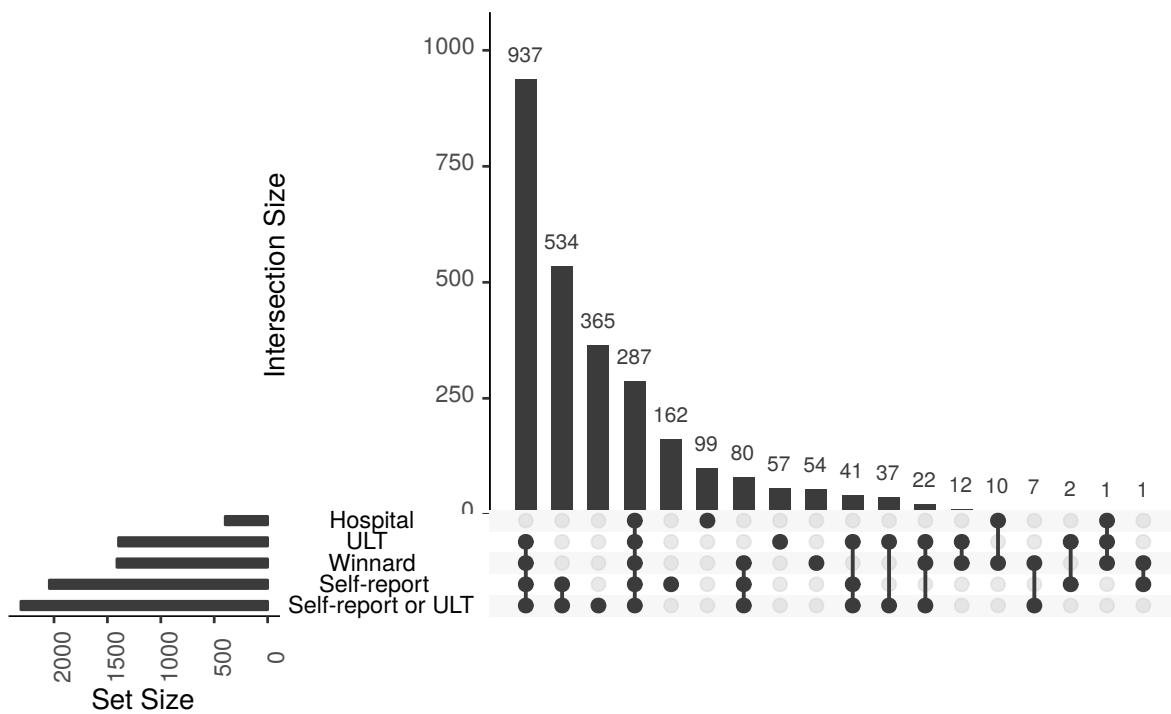


Figure 5.4: Overlap of markers that reached nominal genome-wide significance between the different gout classifications. The upper histogram indicates the size of the intersection and the left histogram indicates the total number of nominally significant SNPs by gout definition.

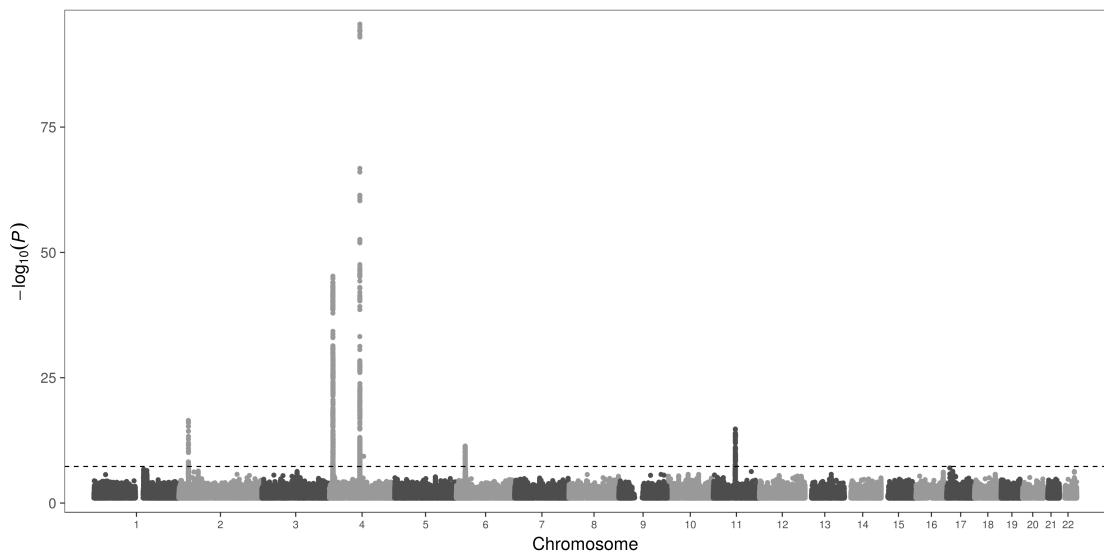


Figure 5.5: Manhattan plot for association with gout, adjusted for age, sex and BMI using the self-reported gout or ULT usage gout definition. The dotted line indicates the genome-wide significance threshold level of 5×10^{-8} . The genome-wide significant peaks in order are *GCKR* (chr2), *SLC2A9* and *ABGC2* (chr4), *SLC17A1-SLC17A3* (chr6), and *NRXN2-SLC22A12* (chr11).

Table 5.4: Top genome-wide significant SNPs by gout definition in the UK Biobank cohort.

Chromosome	Definition	Gene name	Marker	A1	A2	OR [95% CI], P
2	Self-report	<i>GCKR</i>	rs780093	T	C	1.316 [1.235-1.403], 1.990 x 10 ⁻¹⁷
2	Self-report or ULT	<i>GCKR</i>	rs780093	T	C	1.293 [1.217-1.374], 7.827 x 10 ⁻¹⁷
2	Winnard	<i>GCKR</i>	rs780094	T	C	1.276 [1.193-1.364], 1.057 x 10 ⁻¹²
2	ULT	<i>GCKR</i>	rs780093	T	C	1.286 [1.198-1.38], 3.666 x 10 ⁻¹²
4	Hospital	<i>ABCG2</i>	rs4148155	G	A	2.126 [1.777-2.543], 1.544 x 10 ⁻¹⁶
4	Self-report	<i>ABCG2</i>	rs4148155	G	A	2.279 [2.106-2.466], 2.844 x 10 ⁻⁹³
4	Self-report or ULT	<i>ABCG2</i>	rs4148155	G	A	2.234 [2.071-2.409], 2.970 x 10 ⁻⁹⁶
4	Winnard	<i>ABCG2</i>	rs4148155	G	A	2.185 [2.009-2.376], 1.125 x 10 ⁻⁷⁴
4	ULT	<i>ABCG2</i>	rs4148155	G	A	2.302 [2.109-2.513], 1.121 x 10 ⁻⁷⁷
4	Hospital	<i>SLC2A9</i>	rs734553	G	T	0.5603 [0.459-0.6838], 1.218 x 10 ⁻⁸
4	Self-report	<i>SLC2A9</i>	rs9994216	G	G	0.548 [0.5025-0.5975], 3.162 x 10 ⁻⁴²
4	Self-report or ULT	<i>SLC2A9</i>	rs4697701	A	G	0.5646 [0.5219-0.6108], 5.547 x 10 ⁻⁴⁶
4	Winnard	<i>SLC2A9</i>	rs4697701	A	G	0.5796 [0.5316-0.6319], 3.883 x 10 ⁻³⁵
4	ULT	<i>SLC2A9</i>	rs4697701	A	G	0.5725 [0.5223-0.6276], 1.207 x 10 ⁻³²
6	Self-report	<i>SLC17A1</i>	rs1165195	T	G	0.8093 [0.7585-0.8636], 1.683 x 10 ⁻¹⁰
6	Self-report or ULT	<i>SLC17A1</i>	rs1165195	T	G	0.8034 [0.7552-0.8547], 4.176 x 10 ⁻¹²
6	Winnard	<i>SLC17A1</i>	rs1165195	T	G	0.8175 [0.7634-0.8753], 7.601 x 10 ⁻⁹
6	ULT	<i>SLC17A1</i>	rs1165195	T	G	0.8162 [0.7592-0.8776], 3.926 x 10 ⁻⁸
6	Self-report	<i>SLC17A3</i>	rs1747550	A	G	0.8186 [0.7675-0.873], 1.090 x 10 ⁻⁹
6	Self-report or ULT	<i>SLC17A3</i>	rs1747550	A	G	0.8087 [0.7606-0.8599], 1.221 x 10 ⁻¹¹
6	Winnard	<i>SLC17A3</i>	rs1747550	A	G	0.8254 [0.7712-0.8833], 2.894 x 10 ⁻⁸
6	ULT	<i>SLC17A3</i>	rs1747550	A	G	0.8186 [0.7618-0.8796], 4.873 x 10 ⁻⁸
11	Self-report	<i>NRXN2</i>	rs10897521	T	C	1.337 [1.24-1.443], 5.729 x 10 ⁻¹⁴
11	Self-report or ULT	<i>NRXN2</i>	rs10897521	T	C	1.341 [1.247-1.441], 1.699 x 10 ⁻¹⁵
11	Winnard	<i>NRXN2</i>	rs10897521	T	C	1.31 [1.209-1.42], 4.063 x 10 ⁻¹¹
11	ULT	<i>NRXN2</i>	rs10897521	T	C	1.345 [1.236-1.464], 5.585 x 10 ⁻¹²
11	Self-report	<i>SLC22A12</i>	rs11231837	T	C	1.319 [1.223-1.423], 7.494 x 10 ⁻¹³
11	Self-report or ULT	<i>SLC22A12</i>	rs11231837	T	C	1.326 [1.234-1.425], 1.521 x 10 ⁻¹⁴
11	Winnard	<i>SLC22A12</i>	rs7929627	G	A	1.298 [1.198-1.406], 1.660 x 10 ⁻¹⁰
11	ULT	<i>SLC22A12</i>	rs7929627	G	A	1.333 [1.226-1.45], 2.022 x 10 ⁻¹¹

A1 is the effect allele.

5.3.2.2 Polynesian GWAS

The GWAS for gout in the Polynesian populations of New Zealand had a total of 2402 individuals, with a mean age of 48.6 (SD 14.9), was 60.5% male, and had a mean BMI of 35.0 (SD 8.35). A break down by gout affection is shown in Table 5.5. There were only two markers that met the threshold for genome-wide significance, rs2725215 (risk allele = T, OR = 1.939, 95% CI = 1.567-2.40, P = 1.162 x 10⁻⁹), located in *PKD2*, and rs2231142 (risk allele = T, OR = 2.306, 95% CI = 1.891-2.81, P = 1.570 x 10⁻¹⁶), located in *ABCG2*. The LD r² between these two markers was 0.58 (Polynesian Super Population (POL)). Two other markers reached the nominal significance threshold, rs2728108 (*ABCG2*) and rs11034401 (chr11 inter-genic). Other genes that had SNPs that were close to the nominal significance threshold, in the 10⁻⁵ < P < 10⁻⁴ range contained some previously associated genes/regions for urate and gout such as *SLC2A9* and the region containing *IBSP*, *MEPE*, *PKD2*, and *ABCG2*, all on chromosome 4 (Figure 5.6). Of the 39 SNPs in the 10⁻⁴ > P > 10⁻⁵ range, only rs3775948 in *SLC2A9*, rs7698623, rs2725220 and, rs1871744 in *ABCG2*, rs6481407 in *BICC1*, and rs6041522 in *LOC105372532*, had associations of P < 0.05 in the UK Biobank gout GWAS, with only rs3775948 being genome-wide significant (P = 9 x 10⁻⁴³).

5.3.3 Comparison of haplotypic selection with gout GWAS

In order to determine if additional loci (which did not meet the significance threshold in GWAS), had extra information from the selection tests that could prioritise the GWAS results, the gout GWAS and selection results were combined. The “European Ancestry” combination, combined the GWAS performed in the UK Biobank using the self-reported gout or ULT usage with the haplotypic selection results for the British in England and Scotland (GBR) population. The “Polynesian Ancestry” combination, combined the Polynesian gout GWAS with the haplotypic selection results for the Polynesian populations of Cook Island Māori in New Zealand (CIM), Māori in New Zealand (NZM), Samoans in New Zealand (SAM), and Tongans in New Zealand (TON). All of the analyses were restricted to the markers of the CoreExome SNP array. A new threshold for the combinations was used that required an |integrated haplotype homozygosity score (iHS)| > 2 or |number of segregating sites by length (nSL)| > 2 and a GWAS P < 2 x 10⁻⁴. The iHS and nSL threshold was selected because |iHS| or |nSL| > 2 is equivalent to P < 0.05, since the statistics are very similar to a Z-score. A “probability score” for the combined statistics was calculated using Score_{combined} = P_{GWAS} x P_{selection}, where P_{selection} was given by 1 - P(|Z|), and Z was the iHS or nSL value. With the Polynesian populations,

Table 5.5: Clinical details for participants of the Polynesian gout GWAS.

	Controls	Cases
n	1182	1222
Mean Age (SD)	44.14 (15.26)	52.95 (13.2)
Mean BMI (SD)	33.35 (8.12)	36.62 (8.26)
Sex (% Male)	43.9	81.4

Mean age is reported in years. BMI is reported in kg/m².

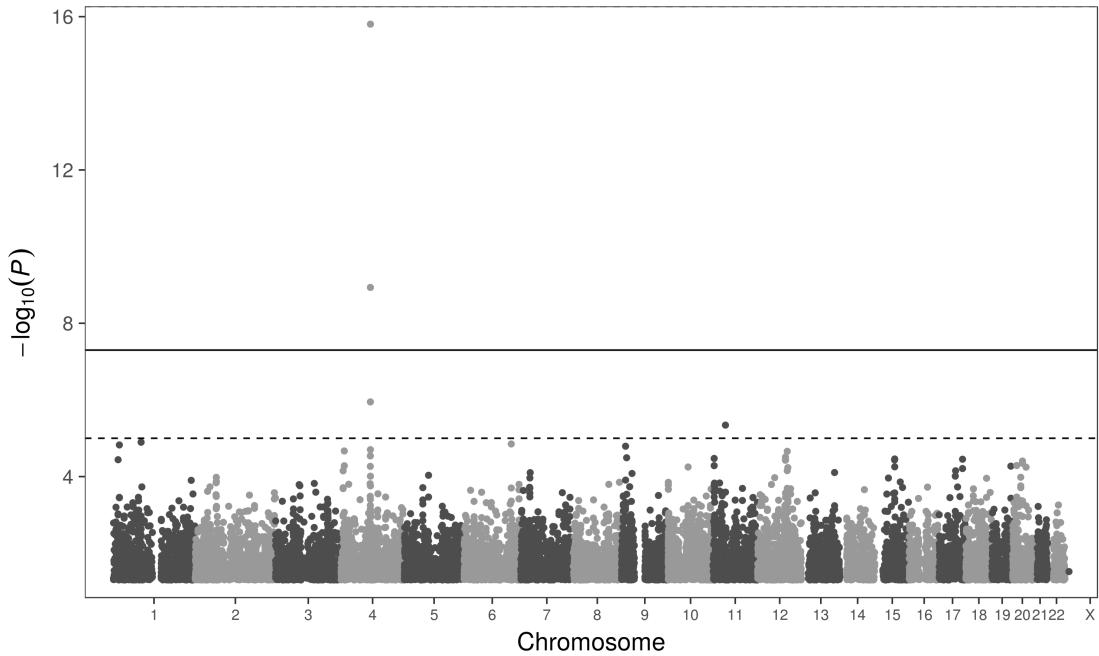


Figure 5.6: Manhattan plot for gout GWAS, adjusted for age and sex, in Polynesian ancestry individuals showing $-\log_{10}(P)$ for the association test by genomic position. The solid line indicates the genome-wide significance threshold of $P = 5 \times 10^{-8}$, the dashed line indicates the suggestive significance threshold of $P = 10^{-5}$.

the maximum $|value|$ was used. The combined threshold was set as $\text{Score}_{\text{combined}} < 5 \times 10^{-8}$, to mirror the GWAS significance threshold.

5.3.3.1 UK Biobank GWAS and haplotypic selection in GBR

When the results from the UK Biobank GWAS for gout (using the self-reported gout or ULT usage definition) and the iHS and nSL results for the GBR population were combined, there was only one SNP, rs12638016 (intergenic), with an $|iHS| > 2.6$ (equivalent to the most extreme 1%) that also had $P < 10^{-5}$ in the GWAS. If the significance threshold for iHS was reduced to $|iHS| > 2$ (approximately the 5% extreme most values), then there were three SNPs, all located at *SLC22A12* or within 1 kb, that were genome-wide significant in the GWAS (rs505802, rs9734313, and rs559946), and four SNPs that were nominally significant. There were 176 SNPs that had an $|iHS| > 2$ and a gout GWAS of $P < 0.01$. This decreased to ten SNPs, when the GWAS P threshold was reduced to 2×10^{-4} (Table 5.6). Only the variants in *SLC22A12* replicated in the UK Biobank replication cohort.

Using a threshold of $|nSL| > 2$ and a nominally significant $P < 10^{-5}$, two SNPs were identified, both on chromosome 6 (Table 5.6). The first was rs4712972, located in *SLC17A4*. The second was rs501220, located in *SLC17A3*. If the thresholds were relaxed, there were 158 SNPs that had a $|nSL| > 2$ and a gout GWAS $P < 0.01$. This decreased to seven, when a GWAS P threshold of 2×10^{-4} was used (Table 5.6). The only variants that replicated in the UK Biobank replication cohort were from *SLC17A3* and *SLC17A4*.

Table 5.6: SNPs that had |iHS| or |nSL| > 2 from the GBR, 1000 Genomes Project population and a gout association P > 2 × 10⁻⁴ in the UK Biobank self-reported gout or ULT definition.

Chromosome	Position	Marker	Gene name	Ref	Alt	Anc	Selection			GWAS			Score combined	UKBB Rep (OR [95% CI], P)
							GBR	A1	OR [95% CI], P	G	1.275 [1.160-1.402]	4.638 × 10 ⁻⁷		
iHS														6.230 × 10⁻⁹
2	27972833	rs12104449	<i>LOC105374378</i>	A	G	A	2.213	G	1.137 [1.071-1.208]	2.842 × 10 ⁻⁵				
3	166249372	rs12638016	<i>LOC105376360</i>	C	T	C	2.632	T	0.883 [0.830-0.939]	8.101 × 10 ⁻⁵	3.438 × 10 ⁻⁷			
10	3628517	rs11251954	<i>SLC22A12</i>	G	A	G	2.075	A	1.140 [1.067-1.218]	9.698 × 10 ⁻⁵	1.842 × 10 ⁻⁶			
11	64357072	rs505802	<i>SLC22A12</i>	T	C	C	2.220	C	1.246 [1.169-1.328]	1.255 × 10 ⁻¹¹	1.659 × 10⁻¹³			
11	64358311	rs9734313	<i>SLC22A12</i>	C	T	C	2.220	C	1.246 [1.169-1.328]	1.267 × 10 ⁻¹¹	1.674 × 10⁻¹³			
11	64358605	rs559946	<i>SLC22A12</i>	T	C	C	2.220	C	1.247 [1.170-1.329]	1.065 × 10 ⁻¹¹	1.407 × 10⁻¹³			
12	112871372	rs11066301	<i>PTPN11</i>	A	G	A	2.469	G	1.136 [1.070-1.206]	3.198 × 10 ⁻⁵	2.167 × 10 ⁻⁷			
15	58970805	rs7182060	<i>ADAM10</i>	T	G	T	-2.060	G	0.824 [0.752-0.903]	3.182 × 10 ⁻⁵	6.262 × 10 ⁻⁷			
16	79943738	rs7185008		G	A	G	-2.220	A	1.215 [1.111-1.328]	1.800 × 10 ⁻⁵	2.378 × 10 ⁻⁷			
21	37953187	rs432137		A	G	G	-2.840	A	1.123 [1.056-1.194]	1.992 × 10 ⁻⁴	4.494 × 10 ⁻⁷			
nSL														1.017 [0.976-1.059]
3	166249372	rs12638016	<i>SLC17A4</i>	C	T	C	2.450	T	0.883 [0.830-0.939]	8.101 × 10 ⁻⁵	5.784 × 10⁻⁷			
6	25772047	rs4712972	<i>SLC17A3</i>	A	G	G	2.306	A	1.202 [1.111-1.301]	5.202 × 10 ⁻⁶	5.489 × 10 ⁻⁸			
6	25873025	rs501220	<i>MUC5AC</i>	C	A	C	2.478	A	1.225 [1.130-1.327]	7.820 × 10 ⁻⁷	5.169 × 10⁻⁹			
11	1150353	rs17859811	<i>ADAM10</i>	G	A	g	-2.100	A	1.165 [1.076-1.261]	1.648 × 10 ⁻⁴	2.947 × 10 ⁻⁶			
15	58970805	rs7182060		T	G	T	-2.367	G	0.824 [0.752-0.903]	3.182 × 10 ⁻⁵	2.850 × 10 ⁻⁷			
16	79943738	rs7185008		G	A	G	-2.747	A	1.215 [1.111-1.328]	1.800 × 10 ⁻⁵	5.409 × 10 ⁻⁸			
21	37953187	rs432137		A	G	G	-2.984	A	1.123 [1.056-1.194]	1.992 × 10 ⁻⁴	2.838 × 10 ⁻⁷			

GWAS results are age, sex and BMI adjusted. Ref is the reference allele in GRCh37, Alt is the alternative allele, Anc is the ancestral allele, and A1 is the effect allele. Score combined < 5 × 10⁻⁸ is in bold.

5.3.3.2 Polynesian GWAS and haplotypic selection in Polynesian populations

Combining the results of the GWAS for gout in Polynesians with the results of the iHS analysis produced a single marker that met the threshold of an $|iHS| > 2.6$ and a GWAS $P < 5 \times 10^{-8}$; that SNP was rs2725215, located in *PKD2*. Rs2725215 also had a $|nSL| > 2.6$. Both iHS and nSL thresholds were only met for TON. There were 1604 SNPs that had a GWAS $P < 0.05$ and an $|iHS| > 2$; this reduced to 11 at the combined threshold of $|iHS| > 2$ and $P < 2 \times 10^{-4}$ (Table 5.7). Combining the nSL results with the GWAS, there were 1502 SNPs that had a GWAS $P < 0.05$ and a $|nSL| > 2$; this reduced to 14 at the combined threshold of $|nSL| > 2$ and $P < 2 \times 10^{-4}$ (Table 5.7).

At the combined threshold of $|iHS| > 2$ or $|nSL| > 2$ and a GWAS $P < 2 \times 10^{-4}$, there were seven SNPs, located within *IBSP*, *PKD2*, *ABCG2* and *PPCDC*, found with both iHS and nSL (Table 5.7). *CHN2* and *SLC39A11* only had SNPs that met the nSL threshold. Out of all of the SNPs that met the combined threshold, ten SNPs had evidence from both iHS and nSL. Rs12908919 (*PPCDC*) was the only SNP that showed significance in multiple populations (SAM and TON) and had support from both iHS and nSL (Table 5.7), whereas, there were six SNPs that had support from only the TON population, for both iHS and nSL. The loci represented were different to the SNPs in the European ancestry GWAS and selection analysis.

The region containing *IBSP*, *PKD2*, and *ABCG2* had multiple SNPs that met the suggestive significance threshold. Of those, six SNPs had evidence of both selection and association (Figure 5.7, Table 5.7). The selection signal came only from the Western Polynesian populations, with no indication in the Eastern Polynesian populations of selection in this region. The recombination rate in this region had intermittent small spikes, with three of the variants in *PKD2* having an LD $R^2 > 0.4$ with rs2231142 (calculated using East Asian Super Population (EAS)). These variants also had an association with gout that met the suggestive significance threshold. Only the variants from *PKD2* and *ABCG2* replicated their association from the UK Biobank replication cohort.

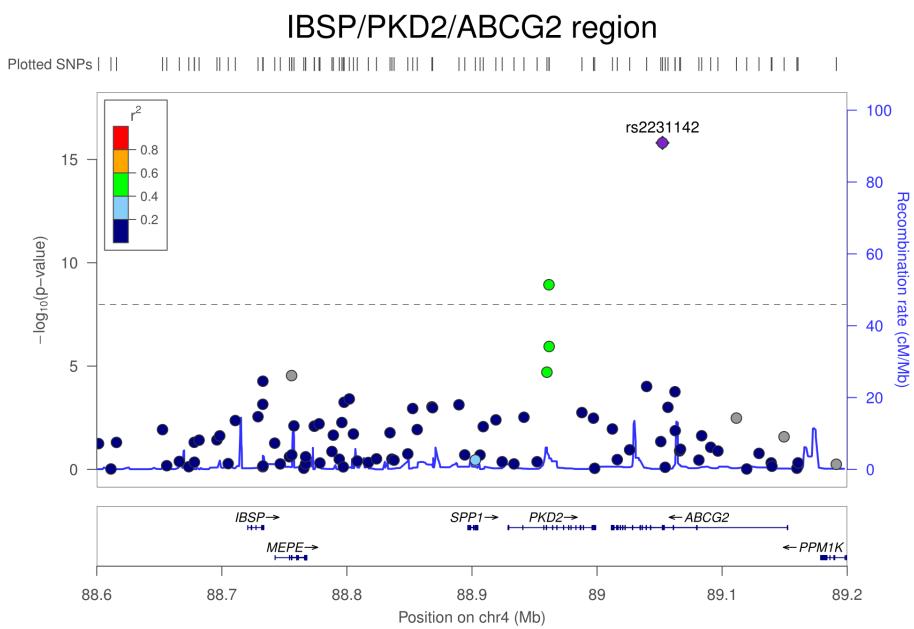


Figure 5.7: Locus zoom plot of Polynesian gout GWAS results for the region chr4:88,600,000-89,200,000, covering *IBSP*, *PKD2*, and *ABCG2*. Points indicate SNP association with gout by genomic position in the Polynesian gout GWAS. Linkage disequilibrium with rs2231142 is indicated by colour. Recombination rate in the East Asian super population of the 1000 Genomes Project is indicated by the blue line. The dashed line indicates the Genome-wide significance threshold $P = 5 \times 10^{-8}$.

Table 5.7: SNPs that had $|iHS|$ or $|nSL| > 2$ in the New Zealand Polynesian populations and a gout association $P > 2 \times 10^{-4}$ in the Polynesian GWAS.

Chromosome	Position	Marker	Gene name	Ref	Alt	Anc	Selection			GWAS			Score combined	UKBB Rep (OR [95% CI], P)		
							CIM	NZM	SAM	TON	A1	OR [95% CI], P				
iHS	4	88732874	rs17013182	<i>IBSP</i>	A	G	A	-	1.544	2.550	G	1.596 [1.272-2.002] 1.466 [1.230-1.748]	5.410 x 10 ⁻⁵ 1.988 x 10 ⁻⁵	2.910 x 10 ⁻⁷	1.079 [0.950-1.225] 1.142 [1.097-1.189]	
	4	88939922	rs2725220	<i>PKD2</i>	G	C	C	0.266	0.863	-1.800	-2.866	C	1.939 [1.567-2.400] 1.546 [1.297-1.843]	1.162 x 10 ⁻⁹ 1.127 x 10 ⁻⁶	4.138 x 10⁻⁸	1.142 [1.097-1.189] 1.150 [1.105-1.198]
	4	88961571	rs7273215	<i>PKD2</i>	C	T	T	-	-1.952	-2.866	T	1.939 [1.567-2.400] 1.546 [1.297-1.843]	1.162 x 10 ⁻⁹ 1.127 x 10 ⁻⁶	4.419 x 10⁻¹²	1.770 [1.671-1.875] 1.150 [1.105-1.198]	
	4	88961736	rs2728108	<i>PKD2</i>	A	C	C	0.270	0.728	-1.579	-2.738	C	1.939 [1.567-2.400] 1.546 [1.297-1.843]	1.162 x 10 ⁻⁹ 1.127 x 10 ⁻⁶	3.479 x 10⁻⁹	1.150 [1.105-1.198] 1.341 x 10 ⁻¹¹
	4	89039629	rs1871744	<i>ABCG2</i>	T	C	T	-0.663	-1.044	-1.723	-2.517	C	0.733 [0.627-0.857] 0.730 [1.891-2.813]	9.769 x 10 ⁻⁵ 1.570 x 10 ⁻⁶	5.789 x 10 ⁻⁷	0.840 [0.795-0.888] 2.098 [1.993-2.208]
	4	89052323	rs2231142	<i>ABCG2</i>	G	T	G	-0.252	-	1.424	2.136	T	0.733 [0.627-0.857] 0.730 [1.891-2.813]	9.769 x 10 ⁻⁵ 1.570 x 10 ⁻⁶	2.568 x 10⁻¹⁸	6.608 x 10 ⁻¹⁷⁷
	4	896567345	rs6914547	<i>G</i>	A	A	A	0.989	1.543	1.944	2.169	A	0.769 [0.671-0.881] 0.755 [0.652-0.874]	1.598 x 10 ⁻⁴ 1.395 x 10 ⁻⁴	2.404 x 10 ⁻⁶	1.005 [0.965-1.047] 1.013 [0.959-1.071]
	6	106507345	rs6914547	<i>G</i>	A	G	G	-2.658	-1.805	-0.747	-0.459	G	0.755 [0.652-0.874] 0.755 [0.652-0.874]	1.395 x 10 ⁻⁴ 1.395 x 10 ⁻⁴	6.261 x 10 ⁻⁷	0.639 [0.593-0.733] 0.982 [0.899-1.073]
	8	10790832	rs6469084	<i>G</i>	A	A	A	-2.170	-1.813	0.151	1.047	A	0.755 [0.652-0.874] 0.593 [0.463-0.758]	3.225 x 10 ⁻⁵ 1.457 [1.201-1.769]	4.844 x 10 ⁻⁷	0.639 [0.593-0.733] 1.057 [0.993-1.126]
	9	15521362	rs7856710	<i>G</i>	A	G	G	1.443	1.308	2.692	2.493	A	1.457 [1.201-1.769] 1.383 x 10 ⁻⁴	4.907 x 10 ⁻⁷	2.800 x 10 ⁻⁶	0.083 [0.997-1.082] 1.039 [0.997-1.082]
	15	75348712	rs12908919	<i>PPCDC</i>	G	A	T	-0.136	1.367	0.907	2.170	T	1.377 [1.164-1.628] 1.866 x 10 ⁻⁴	1.377 [1.164-1.628] 1.866 x 10 ⁻⁴	1.377 [1.164-1.628] 1.866 x 10 ⁻⁴	0.068 [0.997-1.082] 1.039 [0.997-1.082]
	17	55901677	rs2685501	<i>LOC105371839</i>	T	C	T	-	-	-	-	-	-	-	-	
nSL	4	88732874	rs17013182	<i>IBSP</i>	A	G	A	-	-	1.387	2.375	G	1.596 [1.272-2.002] 1.466 [1.230-1.748]	5.410 x 10 ⁻⁵ 1.988 x 10 ⁻⁵	4.750 x 10 ⁻⁷	1.079 [0.950-1.225] 1.142 [1.097-1.189]
	4	88939922	rs2725220	<i>PKD2</i>	G	C	C	0.309	0.705	-2.049	-3.129	C	1.939 [1.567-2.400] 1.546 [1.297-1.843]	1.162 x 10 ⁻⁹ 1.127 x 10 ⁻⁶	1.744 x 10⁻⁸	1.142 [1.097-1.189] 1.770 [1.671-1.875]
	4	88961571	rs2725215	<i>PKD2</i>	C	T	T	-	-	-2.198	-3.128	T	1.939 [1.567-2.400] 1.546 [1.297-1.843]	1.162 x 10 ⁻⁹ 1.127 x 10 ⁻⁶	1.024 x 10⁻¹²	1.022 x 10 ⁻⁸⁴
	4	88961736	rs2728108	<i>PKD2</i>	A	C	C	0.308	0.556	-1.754	-2.929	C	1.546 [1.297-1.843] 0.733 [0.627-0.857]	1.127 x 10 ⁻⁶ 9.769 x 10 ⁻⁵	1.919 x 10⁻⁹	1.150 [1.105-1.198] 0.840 [0.795-0.888]
	4	89039629	rs1871744	<i>ABCG2</i>	T	C	T	-0.691	-1.168	-1.724	-2.488	C	0.733 [0.627-0.857] 0.730 [1.891-2.813]	9.769 x 10 ⁻⁵ 1.570 x 10 ⁻⁶	6.277 x 10 ⁻⁷	1.057 [0.993-1.126] 2.098 [1.993-2.208]
	4	89052323	rs2231142	<i>ABCG2</i>	G	T	G	-0.489	-	1.440	2.146	T	0.733 [0.627-0.857] 0.730 [1.891-2.813]	9.769 x 10 ⁻⁵ 1.570 x 10 ⁻⁶	2.500 x 10⁻¹⁸	6.608 x 10 ⁻¹⁷⁷
	5	56390346	rs8331831	<i>A</i>	G	G	G	1.361	2.063	1.334	1.234	A	1.331 [1.145-1.547] 1.331 [1.145-1.547]	1.947 x 10 ⁻⁴ 1.947 x 10 ⁻⁴	3.812 x 10 ⁻⁶	0.995 [0.946-1.046] 0.995 [0.946-1.046]
	6	166567345	rs6914547	<i>G</i>	A	A	A	1.156	1.356	1.403	2.135	A	0.769 [0.671-0.881] 0.769 [0.671-0.881]	1.598 x 10 ⁻⁴ 1.598 x 10 ⁻⁴	2.618 x 10 ⁻⁶	0.841 [0.965-1.047] 0.841 [0.965-1.047]
	7	29213691	rs69144596	<i>CHN2</i>	C	T	C	-1.487	-1.601	-1.601	-0.764	T	1.324 [1.152-1.522] 0.755 [0.652-0.874]	7.809 x 10 ⁻⁵ 1.595 x 10 ⁻⁴	1.141 x 10 ⁻⁶	1.016 [0.938-1.101] 1.016 [0.938-1.101]
	8	107908032	rs6469084	<i>A</i>	G	G	G	-2.210	-1.767	-1.767	-0.851	-0.531	G	0.755 [0.652-0.874] 0.755 [0.652-0.874]	2.161 x 10 ⁻⁶	1.013 [0.959-1.071] 1.013 [0.959-1.071]
	15	75348712	rs12908919	<i>PPCDC</i>	G	A	G	1.429	1.419	2.340	2.198	A	1.457 [1.201-1.769] 1.383 x 10 ⁻⁴	1.334 x 10 ⁻⁶	1.057 [0.993-1.126] 1.039 [0.993-1.126]	
	17	55901677	rs2685501	<i>LOC105371839</i>	T	C	T	0.352	1.221	1.166	2.355	T	1.377 [1.164-1.628] 1.866 x 10 ⁻⁴	1.730 x 10 ⁻⁶	1.039 [0.997-1.082] 1.018 [0.977-1.060]	
	17	70897198	rs4969131	<i>SLC39A11</i>	C	T	T	-1.257	-0.913	-1.211	-1.963	C	0.732 [0.631-0.849] 5.981 x 10 ⁻⁵	3.527 x 10 ⁻⁵	5.981 x 10 ⁻⁷	1.018 [0.997-1.082] 1.018 [0.977-1.060]

GWAS results are age, sex, and PCA adjusted. Ref is the reference allele in GRCh37. Alt is the alternative allele. Anc is the ancestral allele, and A1 is the effect allele. Score combined < 5 x 10⁻⁸ is in bold.

5.4 Chapter Discussion

5.4.1 Performance of gout definition

The UK Biobank had a larger number of cases for the gout GWAS than Köttgen *et al.* (2013). This increase in cases reflects the increase in strength of association that was seen. A similar comparison of gout definitions was performed using data from the study for updated gout and classification criteria (SUGAR) with 983 rheumatology patients with known gout affection and definitions derived from epidemiology studies used in the Global Urate Genetics Consortium GWAS of hyperuricaemia and gout were tested for sensitivity and specificity (Dalbeth *et al.*, 2016). In the SUGAR paper, the definition that provided the highest specificity (82%) and sensitivity (72%) was self-reported gout or on ULT when tested against the mono-sodium urate crystal identification as the gold standard. Consistent with this, the work presented here found that the self-report of gout or ULT use definition also provided the highest precision from the definitions used, and supports the use of this definition in genetic studies when better gout classification methods are not available, such as ACR criteria or mono-sodium urate crystal identification.

The different definitions of gout used in this study may represent the different disease presentations or patient populations. Not all patients were captured by all criteria, and the Winnard definition was not as good as self-report of gout or ULT, despite a similarity in definitions, and had lower precision in the genetic association. The hospital diagnosis definition had the lowest prevalence and was restrictive, making it the least likely to capture most people with gout. One third of the people that met this definition did not meet any of the self-report definitions, for gout, and/or ULT usage. There is a number of reasons that might explain this. Firstly, the hospitalised population may have a different disease presentation from those in the community identified by self-report or ULT use. Secondly, a diagnosis of gout made in a hospital may subsequently be revised to a different diagnosis; this is not taken into account with the current methodology. Therefore, when self-report information is available it is recommended that the self-report of gout or ULT use be used as the way to define cases.

Each of the gout definitions tested had genome-wide significance for SNPs associated with gout in *ABCG2* and *SLC2A9*. These two genes encode proteins that transport urate in the gut (*ABCG2*) and proximal tubule of the kidney (*SLC2A9*). The large effect sizes in the association with gout, show similarity with their large effect sizes in the control of serum urate levels (Köttgen *et al.*, 2013).

Heritability estimates for the variance attributed to the additive effects of common SNPs for gout of 0.282-0.308 (excluding the hospital definition) were in line with those reported by Köttgen *et al.* (2013), with the same approach using the GCTA software of 0.27 to 0.41 for serum urate levels. There was also no statistically significant differences between the heritability estimates for the different definitions. The comparability between the heritability estimates for gout and serum urate by common genetic variants suggest the genetic heritabilities of serum urate levels and gout is similar. However, also contributing to the risk of gout are environmental factors, such as diet and medications. Because the estimates are constructed under an additive model, it does not account for the non-additive variance, such as gene-gene or gene-environment interactions, or rare variants, or the effect of structural variants

such as copy number variation.

5.4.2 GWAS and selection

The main peak of association in the Manhattan plots (Figures 5.5 and 5.6) for the GWAS analyses of both the UK Biobank, and the New Zealand Polynesians was *ABCG2*, this was consistent with previous GWAS analyses for gout, compared to GWAS for serum urate which has the strongest association with *SLC2A9* (Okada *et al.*, 2012b; Köttgen *et al.*, 2013; Nakayama *et al.*, 2017). Combining the haplotypic selection results for iHS and nSL from the GBR population with the UK Biobank gout GWAS showed additional evidence at *SLC22A12*, *PTPN11*, *SLC17A3*, *SLC17A4*, and *ADAM10*. However, from the GWAS those genes had already met the suggestive significance threshold of $P < 10^{-5}$. All except *ADAM10* had previously been reported as being significant at a genome-wide threshold for association with serum urate levels (Köttgen *et al.*, 2013). *ADAM10* had not been previously associated at a genome-wide threshold with either gout or serum urate levels, but had previously been reported as a strong candidate for positive selection (Deschamps *et al.*, 2016). *ADAM10* is involved in the innate immune system with the Notch signalling pathway, but also becomes down-regulated during interaction with the extracellular proteins PfSEL1/PfSEL2 of *P. falciparum* (Singh *et al.*, 2009). The strength of association weakened in the replication UK Biobank cohort for all loci that had been identified, with most showing no signs of significance except *SLC22A12*, which increased in strength of association.

The combination of the New Zealand Polynesian gout GWAS with the haplotypic selection statistics of iHS and nSL from the Polynesian populations of CIM, NZM, SAM, and TON did not have any corroboration with the equivalent combination of analyses in the European ancestry cohort. The TON population was the main source of the selection statistics that met the threshold, this could be due to the high proportion of gout patients ($> 50\%$) that were included in the sample. Between the iHS and nSL based results, there was a high degree of overlap in the SNPs that met the thresholds for both, although nSL had three SNPs that did not meet the threshold in iHS.

The *IBSP/PKD2/ABCG2* region has been associated with both gout and urate levels (Yang *et al.*, 2010a), but often just the lead SNP of rs2231142 is reported, as it has the strongest association and is one of the most likely causative variants due to the Q141K amino acid change that leads to a reduction in expression and a less efficient transporter (Woodward *et al.*, 2009). The *IBSP* locus encodes bone sialoprotein, which is involved with bone development (Kerr *et al.*, 1993). Loci nearby such as *MEPE* and *SPP1*, encode proteins with similar structure and functions, suggesting a shared evolutionary history (Rowe *et al.*, 2000). In the presence of urate crystals, as with gout, *IBSP* expression is reduced and affects osteoblast differentiation (Chhana *et al.*, 2011). *PKD2* is involved with kidney disease (Mochizuki *et al.*, 1996; Hildebrandt, 2010), and Polynesian populations have a 3.5 fold higher incidence of end-stage renal disease, compared to Europeans (Collins *et al.*, 2017). For all the variants that had both selection and GWAS signal at *PKD2*, the selection signal was for the ancestral allele, which was also the risk allele for gout. The selection signal was absent in the Eastern Polynesian populations, which displayed selection statistics in favour of the derived/protective gout alleles at *PKD2*. As previously discussed (section 3.4), many of the loci that have evidence suggesting

selection in Western Polynesian populations are calcium channels. The *IBSP/PKD2/ABCG2* region, is functionally involved with calcium, bone sialoprotein has a high affinity for calcium (Kerr *et al.*, 1993), and PKD2 has some homology with calcium channels, containing domains that are consistent with calcium binding (Mochizuki *et al.*, 1996), again demonstrating this calcium-related selection pattern.

5.4.3 Limitations

In the analysis of the different gout definitions, there was a difference between the the definitions in the number of SNPs that had a significant association with gout, with the hospital diagnosis definition having the fewest. One of the key differences between the definitions was the number of cases defined for each, ranging from 382 to 2295, which will have impacted on the power for each GWAS to detect associations. Similar to the GWAS results from the different gout definitions, the differences between heritability estimates of the different definitions may too be impacted by the differing numbers of cases. Due to computational reasons, only a sub-sample of the controls was used in the calculation of the heritability estimates. The use of all of the controls may have changed the results of the heritability analysis. Another limitation of the definitions was the lack of ‘gold standard’ to which the definitions could be compared, and therefore sensitivity and specificity were unable to be calculated for each definition. The ancestry of the UK Biobank samples included for the gout definitions were European, this means that results and conclusions from the definition and heritability analysis may not be applicable to other ancestral backgrounds.

With the combination of the selection statistics with the results of the gout GWAS analyses, the variants that were able to be compared were limited by the selection analysis, where the selection results were only from the markers that were available on the CoreExome SNP array. This is despite the UK Biobank GWAS being conducted on an imputed dataset, where there were results for 9.3 million markers after filtering. This compared to the total number of 236,868 markers for which there were iHS or nSL results. It is possible that if there were iHS and nSL results for a greater number of markers, more loci may have been prioritised. The Polynesian GWAS was limited by sample size, and also in the number of markers for which there were both association and selection results. The marker density of the CoreExome SNP array could have been improved through imputation, however, for Polynesian populations, the public haplotype reference panels lack representation of Polynesian populations. This means that Polynesian specific variants (such as rs373863828, Minster *et al.* (2016)) and therefore haplotypes, are not incorporated in the imputation, and as a result the imputation accuracy suffers (Howie *et al.*, 2011).

The combination of the Eastern and Western Polynesian populations into a single grouped population for the GWAS may also affect the results, as there are population specific effects, even between similar ancestral backgrounds such as the East and West Polynesian populations. Such differences with respect to gout have been seen with *ABCG2* (Phipps-Green *et al.*, 2010). This was similar to what was observed in the selection results at the extended locus including *IBSP*, *PKD2*, and *ABCG2* displaying a difference between Eastern and Western Polynesian populations. One method to take into account these differences between the Eastern and Western Polynesian populations, would be to perform the

GWAS separately in each and then meta-analyse the results.

5.4.4 Conclusions

The analyses in this chapter for performance of the case-definition for gout association studies showed that in the absence of the gold-standard of observing urate crystals in the synovial fluid, or the ACR criteria, it is recommended using the self-report or ULT definition, as this gave the best performance in estimating the cohort gout prevalence, out of the definitions tested in this analysis. It was also shown that incorporation of selection analyses with GWAS does provide evidence for additional variants associated with gout. Despite these associations already mostly being nominally significant from GWAS, the selection analyses do provide an avenue for prioritisation of GWAS results but only appeared to further enhance evidence at known loci, rather than aid in discovery. Further work such as incorporating multiple selection statistics, and the use of whole genome sequenced data or imputed genetic data would be useful to improve the prioritising of GWAS results.

Chapter 6

Summary and Conclusions

6.1 Summary

This thesis set out to investigate the role of genetic selection in the genome of modern Polynesian populations, and its effect on urate and metabolic disease. There were three main objectives that were covered. The first was to identify and characterise positive selection within Polynesian populations with regard to metabolic diseases such as gout, obesity, type 2 diabetes, kidney disease, and metabolic syndrome. The second was to investigate the shared ancestral history of Polynesian populations in genomic regions relevant to metabolic disease. The third was to investigate the impact of alternative gout definitions in association studies, and to incorporate the use of selection statistics into association analyses.

6.1.1 Evidence of selection in Polynesian populations

Chapter 3 utilised multiple selection and neutrality statistics, based on haplotypic and frequency spectrum methodologies, to establish regions of the genome that exhibited ‘signals of selection’. The findings of Chapter 3, specifically identified regions of the genome in Polynesian populations that had evidence suggesting that positive selection had played a role. The characterisation of these regions through pathway analysis indicated that metabolic functions dominated in the pathways with enrichment of genes that had evidence of possible selection. However, only fractions of the pathways were enriched. The majority of the genes in these pathways were enriched for significant markers from nSL. The Eastern Polynesian populations had a greater number of enriched pathways than the Western Polynesian populations. Many of the genes in the pathways enriched in the Eastern Polynesian populations were involved with signalling, specifically with calcium.

The enriched pathway in common between the Western Polynesian populations was ABC transporters. The loci, *ABCG2* and *ABCC4*, that encode for two ABC transporters, have previously had genetic variants identified that increase risk of gout in Western Polynesian populations (Phipps-Green *et al.*,

2010; Tanner *et al.*, 2017). *ABCC4* has also had a Western Polynesian specific SNP (rs972711951) identified that associated with gout (Tanner *et al.*, 2017), and there was evidence of possible selection at this locus, from CIM, NZM, and SAM.

The loci that exhibited ‘signatures of selection’ in Polynesian populations, that were also associated with urate were limited, and there was minimal evidence for the main effect loci for urate and gout of *SLC2A9* and *ABCG2*. Instead the loci that indicated possible selection included *RREB1*, *IGFR1* and *BCAS3*, none of which are urate transporters but instead are involved with more central metabolic pathways. The other metabolic diseases of obesity, type 2 diabetes, kidney disease, and metabolic syndrome all had a number of associated loci that also had evidence of possible selection. Some of these loci were associated with multiple traits, and had an immunological function. This points to a potential relationship of these loci being influenced by pathogenic challenge, that results in influence on metabolic diseases. One example of this potential relationship could be loci that were associated with obesity and type 2 diabetes, that were also associated with malarial infection. One of the genes, *DDC*, that was putatively associated with malaria, had one of the strongest iHS signals in NZM.

Due to urate having a central role in malaria infection and functioning as an adjuvant for the innate immune system (Ames *et al.*, 1981; Opitz *et al.*, 2009), as well as previously having been identified a selective pressure applied to the genome, malaria associated genes were investigated for signals of selection in Polynesian populations. Three loci that had associations with malaria, to varying degrees, showed possible evidence for selection through haplotypic statistics, these were *ABO*, *ATP2B4*, and *DDC*. The blood antigen locus of *ABO*, has a suggested mechanism for how a red blood cell displaying the type A or B antigen may increase cytoadherence with a malaria infected cell, increasing the infectivity of *Plasmodium falciparum* (Cserti and Dzik, 2015). However, this may be limited in benefit to regions that have a high frequency of O type such as Melanesia, and not for Polynesian populations (Simmons, 1962; Zerihun *et al.*, 2011).

Unfortunately, two previously identified loci (*PPARGC1A* and *CREBRF*) which had been posited as thrifty-gene candidates, were unable to be verified, or rejected as having under-gone selection due to the absence of the markers (or surrogates) that had previously been reported. The selection signal had been specific to those markers.

6.1.2 Shared ancestry of selected loci

Chapter 4 was an investigation into the genetic similarities of populations in the regions that lay in the extremes of the selection and neutrality statistic distributions. This chapter added to the evidence that the modern-day Polynesian populations had genetic similarities to modern day East Asian populations, and from the migration and settlement histories have a shared ancestry. There was also evidence from all the clustering methodologies that while the Polynesian populations were most similar to each other in a global context, there were in fact genetic differences between the East and West Polynesian populations, a finding also reported by Hudjashov *et al.* (2018).

The principal component analysis, used to partition the variance of the genetic data, where each subsequent principal component captures smaller amounts of variance, showed that the first four

components could be used to explain the genetic variation that separated the populations into their geographically based super population groups. The first component captured the difference of the African Super Population (AFR) populations and all the other populations. The second component captured the difference between the European Super Population (EUR) populations and the EAS, South Asian Super Population (SAS), and POL populations. The third component captured the variation responsible for separating the EAS from the POL populations. And the forth component captured the genetic variation that separated the SAS and American Super Population (AMR) populations from each other, and other populations.

The selection and neutrality statistics showed that populations within a super population were most similar to one another, with the greatest differences being between super populations. There was also evidence of a shared ancestry in both the frequency spectrum of variants, as well as with haplotypes that was consistent with the Out of Africa migration and subsequent population movements. There was not however a specific signal for selection that appeared in the Polynesian populations for urate or any of the metabolic disease associated loci, but instead a commonality between the frequency spectrum of similar geographic populations. The extremes of the frequency-based selection and neutrality statistics showed that there was similarity in the regions of the genome within a super population, but the regions that were in the extremes differed between super populations. This suggested that local adaptations were geographically restricted (Gravel *et al.*, 2011).

6.1.3 Incorporation of selection analyses into GWAS

Chapter 5 investigated the use of gout definitions that were common amongst multi-purpose cohorts and assessed the performance of multiple definitions. It was found that the best definition was that of self-reported gout or self-report of ULT usage, when the ACR criteria or observation of urate crystals in the synovial fluid is not available. The use of selection statistics in prioritisation of GWAS loci revealed that there were several loci that had possible evidence of selection that were “in the noise” of the GWAS signal, however, these loci were limited to previously identified regions, with the new suggestive associations failing to replicate.

6.2 Significance

Investigations of selection in different populations have yielded several population-specific genetic adaptations. Some examples of adaptations with evidence of selection include lactase in European populations (Bersaglieri *et al.*, 2004), pigmentation in South Asian populations (Jonnalagadda *et al.*, 2017), altitude adaptations in Tibetans (Huerta-Sánchez *et al.*, 2014), adaptations to climate in Greenlanders (Fumagalli *et al.*, 2015), and most recently adaptations for diving in the Bajau people (Ilardo *et al.*, 2018).

The health disparities that affect Polynesian populations, such as the high burden of metabolic diseases like obesity, type 2 diabetes, renal disease, and gout are important to understand and address.

Determining the origin of population genetic differences has the potential to lead to new insights for the biological model. Looking at genetic selection can help explain how these population genetic differences came to be. This thesis is the first to conduct a genome-wide scan for regions of genetic selection, and to identify and characterise selection through a range of selection statistics, in multiple Polynesian populations. The population sample sizes are also some of the largest for Polynesia compared to other studies (Friedlaender *et al.*, 2008; Kimura *et al.*, 2008; Skoglund *et al.*, 2016; Hudjashov *et al.*, 2018). From these genome-wide selection scans, there was evidence of metabolic pathways being enriched for genes displaying signals of selection. But importantly, the genes that were associated with urate and gout that displayed the most evidence of possible selection, were not urate transporters, but genes that also had associations with other metabolic diseases.

Research into the genetic ancestry of populations, and in particular Polynesian populations is still a current research focus (Hudjashov *et al.*, 2018; Matisoo-Smith and Gosling, 2018). The research in this thesis (in particular chapter 4) added to the current knowledge by comparing genome-wide SNP data and selection statistics, finding there was additional evidence of similarity in ancestry between modern-day Polynesian populations, and modern-day EAS populations. The regions that showed signs of possible selection had varying degrees of similarity between populations, some were only seen in East or West Polynesian populations, while others had signal that was shared with other populations, with sharing with EAS being the most common.

Recommendations around the use of gout definition in genetic studies, when the gold-standard gout diagnosis or other clinical based criteria (ACR criteria) are not available. When individual level genetic data is available, this can assist in pooling genetic data between studies, rather than performing meta-analysis. The heritability of gout was also confirmed to be similar to that of urate.

The incorporation of selection statistics provided another method to prioritise variants from GWAS. Exploring additional options such as these is important, especially when sample sizes for GWAS in Polynesian populations are unlikely to ever be large enough to have the power to detect all of the small effect loci that contribute to complex genetic diseases such as gout. The analysis of combining selection statistics with GWAS indicated it has potential by aiding in providing additional evidence for known loci but did not aid in the discovery of new gout-associated variants. Broadening the types of selection statistics used might provide different results.

6.3 Study limitations

One of the major limitations of this research project was the reliance on SNP array data. Ideally whole genome sequencing would have been used, however, Polynesian populations are generally under-represented in large sequencing projects. A clear demonstration of the benefit of using sequence data over SNP array data is with the *CREBRF* variant, where rs373863828 was specific to Polynesian populations (Minster *et al.*, 2016; Krishnan *et al.*, 2018) and not represented in my analysis. Using SNP array data impacted on all aspects of the selection statistics analysis. The marker density of the SNP array meant that the windowed statistics had a small number of markers per window, compared

to the HapMap data, with whole genome sequence data giving the highest density. This meant that the minimum number of markers of four was conveying signal for 100 kb regions of the genome. The data from the CoreExome SNP array could have been imputed, however, due to the Polynesian specific variant, and therefore haplotypes not being present in the reference panel haplotypes, the imputation quality would be affected. The impact of this, beyond the potential for false haplotypes to be introduced is still unknown. The trade-off benefit of having an increased density of markers versus the increased probability of incorrect haplotypes has not been quantified, as the specificity of haplotypes in Polynesian populations is unknown.

Another limiting factor for the analyses in chapters 3, 4 and 5, was the focus on positive selection. This focus came from the hypothesis that urate had been beneficial in the past. However, other types of selection are also relevant to complex genetic diseases, and influence the genome (Andrés *et al.*, 2009; Daub *et al.*, 2013). On top of selection, there is also the possibility of random genetic drift, population expansions, bottlenecks, and migrations that all play a role in shaping the genome.

The statistics used in this thesis have a wide range of time-frames they are powered for, however, many of the differences that are being looked at are in the < 10kyears ago (ya) time frame, so some of these methods might not be powered appropriately. In addition, the nature of the SNP array data means that some of the newer methods are not necessarily applicable, for instance singleton-density score (Field *et al.*, 2016b). This is especially true where the SNP array data are missing very low frequency variants due to an ascertainment bias in the markers on the SNP array, and from subsequent quality control procedures that might have been implemented.

Using the GWAS catalog provided a convenient method to incorporate many of the known genetic associations found from GWAS into the selection analysis. There were a few effects from this approach that will have influenced the results in the analyses involving the GWAS catalog derived gene lists. First, the associations were only for GWAS, and did not report on associations that were found in a non-GWAS setting, such as candidate genes, so some true genetic associations may have been missed. The impact of this on the conclusions is likely to be small, given that a trend being searched for was a systematic selection signal across pathways (section 3.3.3.1.3). Given the number of genes in a pathway that showed signs of selection and the pathway sizes involved, missing a few loci per pathway is unlikely to have made this systematic selection signal appear. Secondly, there is the fact that a large proportion of the GWAS that have been done and reported in the GWAS catalog are for European populations (Haga, 2010; Popejoy and Fullerton, 2016), and the focus of this research was in Polynesian populations. The result of this will be population-specific associations that are missed, so incorporation of other association information, such as candidate genes or curated gene lists would be beneficial. The use of the GWAS catalog, while extremely useful in terms of defining of gene lists, may have ‘cast a wide net’ for genes with associations, than perhaps an expertly curated candidate-gene list. By having an extended list, this may have increased the ‘noise’ in the selection signals which meant that trends in the loci that showed signs of selection for a trait were harder to distinguish.

The technicalities of annotating gene information onto SNPs and genomic regions can be challenging, especially with regard to annotation of intergenic regions where there might be no clear nearest gene. Consideration also needs to be given to the functional impact of variants on expression, or the effects

of regulatory elements such as DNase hypersensitivity sites. The incorporation of this information was limited, but with new annotation and analysis pipelines being developed (Ferrero, 2018), future work will benefit from including these features in the analysis.

Selection is not the only explanation for genetic differences between populations. Other causes, such as random genetic drift, migration, population expansions or bottlenecking can also contribute to these differences. In order to boost the confidence that the signal's that were seen were the result of selection, the analysis would have benefited from population simulations using models that may best explain the population history (Yuan *et al.*, 2012). Extensions to simulation techniques could include training machine learning algorithms or deep learning on simulated data and applying the model predictions to real data (Pybus *et al.*, 2015; Sheehan and Song, 2016; Schrider and Kern, 2018)

6.4 Future directions

Future work in this area would include the generation of a high-quality set of Polynesian whole-genome sequences that could be used to remove the ascertainment bias introduced through the SNP arrays. High quality sequences could also be used to supplement the reference haplotype panels that are currently available to include haplotypes that are specific to Polynesian populations which could then be used improved haplotype phasing, and for imputation of Polynesian specific variants.

If sequence data were available, then the use of some of the newer statistics such as the population branching statistic (Yi *et al.*, 2010), singleton density score (Field *et al.*, 2016b), and levels of exclusively shared differences (Librado and Orlando, 2018) could be made use of, which are designed for detecting selection on more recent time-scales.

Comparing the present day East Asian and Polynesian populations with ancient Polynesian DNA could prove useful in determining areas of the genome that differ, to refine the regions of the genome that have changed since population divergence, and subsequent changes in the Polynesian populations. Comparisons with Denisovan and Neanderthal data could also be used to investigate the impact introgression has had on the Polynesian populations, or if regions that showed signs of selection were also similar to regions of introgression.

Investigating the use of selection statistics with GWAS could still prove fruitful if other statistics and methodologies were incorporated that allowed for cross-population comparisons, especially in situations where there are differences in the disease prevalence between populations. This could be further improved by inclusion of expression quantitative trait loci information, and regulatory elements to assess if the markers and regions identified in the selection analyses have known impact on gene expression, or are tissue specific. Furthermore, replication of this analysis in other sample Polynesian populations, and the use of population simulations would be beneficial to increase the confidence in the regions that displayed evidence of possible selection.

6.5 Conclusion

Overall, it was shown that there was evidence suggesting positive selection in Polynesian populations with some of the regions including loci associated with urate and metabolic diseases. Further evidence was shown for the modern-day Polynesian populations being most similar to modern day East Asian populations, with there being a degree of similarity in the regions displaying signatures of selection. It was also shown that selection statistics could be used to group populations of similar ancestry.

Identification of additional gout-associated loci in GWAS was limited when selection statistics were incorporated, but the number of statistics used was not exhaustive, so expansion of the statistics used, particularly cross-populational statistics, may yield different results. It is also shown that the use of self-reported gout or ULT usage to define gout-cases in a case-control genetic association study gave the best performance, in the absence of the gold-standard gout diagnosis, or ACR criteria.

This thesis has provided identification and characterisation of regions of the genome that displayed “signatures of selection” in Polynesian populations, that was previously unavailable. It has also provided insight into the role of selection with respect to urate and metabolic disease.

References

- 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* **467**, 1061–1073
- 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65
- 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* **526**, 68–74
- Abraham, R., Moskvina, V., Sims, R., Hollingworth, P., Morgan, A., Georgieva, L., Dowzell, K., Cichon, S., Hillmer, A.M., O'Donovan, M.C., Williams, J., Owen, M.J. and Kirov, G. (2008) A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Medical Genomics* **1**, 44
- Achaz, G. (2008) Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–1424
- Ahmad, S., Zhao, W., Renström, F., Rasheed, A., Zaidi, M., Samuel, M., Shah, N. et al. (2016) A novel interaction between the *FLJ33534* locus and smoking in obesity: a genome-wide study of 14,131 Pakistani adults. *International Journal of Obesity* **40**, 186–190
- Ahmetov, I.I., Williams, A.G., Popov, D.V., Lyubaeva, E.V., Hakimullina, A.M., Fedotovskaya, O.N., Mozhayskaya, I.a., Vinogradova, O.L., Astratenkova, I.V., Montgomery, H.E. and Rogozkin, V.a. (2009) The combined impact of metabolic gene polymorphisms on elite endurance athlete status and related phenotypes. *Human Genetics* **126**, 751–761
- Aidoo, M., Terlouw, D.J., Kolczak, M.S., McElroy, P.D., ter Kuile, F.O., Kariuki, S., Nahlen, B.L., Lal, A.A. and Udhayakumar, V. (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *The Lancet* **359**, 1311–1312
- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723
- Akey, J.M. (2012) Parallel selection: Evolution's surprising predictability. *Current Biology* **22**, R407–R409
- Akey, J.M., Ruhe, A.L., Akey, D.T., Wong, A.K., Connelly, C.F., Madeoy, J., Nicholas, T.J. and Neff, M.W. (2010) Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1160–1165

- Alarcón-Riquelme, M.E., Ziegler, J.T., Molineros, J., Howard, T.D., Moreno-Estrada, A., Sánchez-Rodríguez, E., Ainsworth, H.C. *et al.* (2016) Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. *Arthritis & Rheumatology* **68**, 932–943
- Alberti, K.G. and Zimmet, P.Z. (1998) Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabetic Medicine: a Journal of the British Diabetic Association* **15**, 539–53
- Alberti, K.G.M.M., Eckel, R.H., Grundy, S.M., Zimmet, P.Z., Cleeman, J.I., Donato, K.a., Fruchart, J.C., James, W.P.T., Loria, C.M. and Smith, S.C. (2009) Harmonizing the metabolic syndrome: A joint interim statement of the international diabetes federation task force on epidemiology and prevention; National heart, lung, and blood institute; American heart association; World heart federation; International . *Circulation* **120**, 1640–1645
- Albrechtsen, A., Nielsen, F.C. and Nielsen, R. (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* **27**, 2534–2547
- Alexander, D.H. and Novembre, J. (2009) Fast model-based estimation of ancestry in unrelated individuals pp. 1655–1664
- Allison, A.C. (1956) The Sickle-cell and haemoglobin C genes in some African populations. *Annals of Human Genetics* **21**, 67–89
- Ames, B.N., Cathcart, R., Schwiers, E. and Hochstein, P. (1981) Uric acid provides an antioxidant defense in humans against oxidant- and radical-caused aging and cancer: a hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 6858–62
- Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C. *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* **43**, 246–252
- Andlauer, T.F.M., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., Chan, A. *et al.* (2016) Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science Advances* **2**, e1501678
- Andrés, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustamante, C.D., Clark, A.G. and Nielsen, R. (2009) Targets of balancing selection in the human genome. *Molecular Biology and Evolution* **26**, 2755–2764
- Antúnez, C., Boada, M., González-Pérez, A., Gayán, J., Ramírez-Lorca, R., Marín, J., Hernández, I. *et al.* (2011) The membrane-spanning 4-domains, subfamily A (*MS4A*) gene cluster contains a common variant associated with Alzheimer's disease. *Genome Medicine* **3**, 33
- Arking, D.E., Pulit, S.L., Crotti, L., Van Der Harst, P., Munroe, P.B., Koopmann, T.T., Sotoodehnia, N. *et al.* (2014) Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature Genetics* **46**, 826–836

- Armstrong, D.L., Zidovetzki, R., Alarcón-Riquelme, M.E., Tsao, B.P., Criswell, L.A., Kimberly, R.P., Harley, J.B., Sivils, K.L., Vyse, T.J., Gaffney, P.M., Langefeld, C.D. and Jacob, C.O. (2014) GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. *Genes and Immunity* **15**, 347–354
- Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T., Matsumoto, T. *et al.* (2009) A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nature Genetics* **41**, 1325–1329
- Aulchenko, Y.S., Hoppenbrouwers, I.A., Ramagopalan, S.V., Broer, L., Jafari, N., Hillert, J., Link, J., Lundström, W., Greiner, E., Sadovnick, A.D., Goossens, D., Van Broeckhoven, C., Del-Favero, J., Ebers, G.C., Oostra, B.A., van Duijn, C.M. and Hintzen, R.Q. (2008) Genetic variation in the *KIF1B* locus influences susceptibility to multiple sclerosis. *Nature Genetics* **40**, 1402–1403
- Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nature Genetics* **41**, 824–828
- Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S.S., Patterson, N. and Reich, D. (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *American Journal of Human Genetics* **81**, 234–242
- Ayub, Q., Moutsianas, L., Chen, Y., Panoutsopoulou, K., Colonna, V., Pagani, L., Prokopenko, I., Ritchie, G.R.S., Tyler-Smith, C., McCarthy, M.I., Zeggini, E. and Xue, Y. (2014) Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *American Journal of Human Genetics* **94**, 176–85
- Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F. *et al.* (2013) Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genetics* **9**, e1003509
- Barcham, M., Scheyvens, R. and Overton, J. (2009) New Polynesian triangle: Rethinking Polynesian migration and development in the Pacific. *Asia Pacific Viewpoint* **50**, 322–337
- Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**, 955–962
- Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D.J., Stevens, H., Todd, J.A., Walker, N.M. and Rich, S.S. (2009a) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41**, 703–707
- Barrett, J.C., Lee, J.C., Lees, C.W., Prescott, N.J., Anderson, C.A., Phillips, A., Wesley, E. *et al.* (2009b) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nature Genetics* **41**, 1330–1334

- Barton, N.H. (1998) The effect of hitch-hiking on neutral genealogies. *Genetical Research* **72**, 123–133
- Baurecht, H., Hotze, M., Brand, S., Büning, C., Cormican, P., Corvin, A., Ellinghaus, D. *et al.* (2015) Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. *American Journal of Human Genetics* **96**, 104–120
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300
- Benjelloun, B., Alberto, F.J., Streeter, I., Boyer, F., Coissac, E., Stucki, S., BenBati, M., Ibnelbachyr, M., Chentouf, M., Bechchari, A., Leempoel, K., Alberti, A., Engelen, S., Chikhi, A., Clarke, L., Flicek, P., Joost, S., Taberlet, P. and Pompanon, F. (2015) Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data. *Frontiers in Genetics* **6**, 1–14
- Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Martín, J., Fairfax, B.P., Knight, J.C., Chen, L., Replogle, J., Syvänen, A.C., Rönnblom, L., Graham, R.R., Wither, J.E., Rioux, J.D., Alarcón-Riquelme, M.E. and Vyse, T.J. (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* **47**, 1457–1464
- Berndt, S.I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E. *et al.* (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics* **45**, 501–512
- Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.a., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**, 1111–20
- Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S. *et al.* (2011) Genome-wide comparison of African-ancestry populations from CARE and other cohorts reveals signals of natural selection. *American Journal of Human Genetics* **89**, 368–81
- Bhatia, G., Patterson, N., Sankararaman, S. and Price, a.A.L. (2013) Estimating and interpreting FST: The impact of rare variants. *Genome Research* **23**, 1514–1521
- Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., Herráez, D.L., Brutsaert, T., Parra, E.J., Moore, L.G. and Shriver, M.D. (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genetics* **6**
- Billings, L.K. and Florez, J.C. (2010) The genetics of type 2 diabetes: what have we learned from GWAS? *Annals of the New York Academy of Sciences* **1212**, 59–77
- Bostrom, M.A., Lu, L., Chou, J., Hicks, P.J., Xu, J., Langefeld, C.D., Bowden, D.W. and Freedman, B.I. (2010) Candidate genes for non-diabetic ESRD in African Americans: a genome-wide association study using pooled DNA. *Human Genetics* **128**, 195–204

- Bradfield, J.P., Qu, H.Q., Wang, K., Zhang, H., Sleiman, P.M., Kim, C.E., Mentch, F.D., Qiu, H., Glessner, J.T., Thomas, K.A., Frackelton, E.C., Chiavacci, R.M., Imielinski, M., Monos, D.S., Pandey, R., Bakay, M., Grant, S.F.A., Polychronakos, C. and Hakonarson, H. (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genetics* **7**, e1002293
- Bradfield, J.P., Taal, H.R., Timpson, N.J., Scherag, A., Lecoeur, C., Warrington, N.M., Hypponen, E. et al. (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature Genetics* **44**, 526–531
- Browning, B.L. and Browning, S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223
- Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P. et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678
- Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-wide association studies. *PLOS Computational Biology* **8**, 1–11.
Available: <https://doi.org/10.1371/journal.pcbi.1002822>
- Cadzow, M., Boocock, J., Nguyen, H.T., Wilcox, P., Merriman, T.R. and Black, M.A. (2014) A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics* **5**, 1–8
- Cadzow, M., Merriman, T.R., Boocock, J., Dalbeth, N., Stamp, L.K., Black, M.A., Visscher, P.M. and Wilcox, P.L. (2016) Lack of direct evidence for natural selection at the candidate thrifty gene locus, *PPARGC1A*. *BMC Medical Genetics* **17**, 80
- Cadzow, M., Merriman, T.R. and Dalbeth, N. (2017) Performance of gout definitions for genetic epidemiological studies: analysis of UK Biobank. *Arthritis Research & Therapy* **19**, 181
- Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J. et al. (2002) A Human Genome Diversity Cell Line Panel. *Science* **296**, 261–262
- Capon, F., Bijlmakers, M.J., Wolf, N., Quaranta, M., Huffmeier, U., Allen, M., Timms, K. et al. (2008) Identification of *ZNF313/RNF114* as a novel psoriasis susceptibility gene. *Human Molecular Genetics* **17**, 1938–1945
- Cappellini, M.D. and Fiorelli, G. (2008) Glucose-6-phosphate dehydrogenase deficiency. *The Lancet* **371**, 64–74
- Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J. and Nickerson, D.A. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research* **15**, 1553–65

- Casto, A.M. and Feldman, M.W. (2011) Genome-wide association study SNPs in the human genome diversity project populations: Does selection affect unlinked SNPs with shared trait associations? *PLoS Genetics* **7**
- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16
- Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303
- Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N.R. and Ma'ayan, A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128
- Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Research* **20**, 393–402
- Chhana, A., Callon, K.E., Pool, B., Naot, D., Watson, M., Gamble, G.D., McQueen, F.M., Cornish, J. and Dalbeth, N. (2011) Monosodium urate monohydrate crystals inhibit osteoblast viability and function: Implications for development of bone erosion in gout. *Annals of the Rheumatic Diseases* **70**, 1684–1691
- Chittoor, G., Kent, J.W., Almeida, M., Puppala, S., Farook, V.S., Cole, S.A., Haack, K. et al. (2016) GWAS and transcriptional analysis prioritize ITPR1 and CNTN4 for a serum uric acid 3p26 QTL in Mexican Americans. *BMC Genomics* **17**, 276
- Cho, Y.S., Chen, C.H., Hu, C., Long, J., Ong, R.T.H., Sim, X., Takeuchi, F. et al. (2011) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature Genetics* **44**, 67–72
- Choi, H., Mount, D. and Reginato, A. (2005) Pathogenesis of gout. *Annals of Internal Medicine* **143**, 499–516
- Choi, H.K. and Ford, E.S. (2007) Prevalence of the Metabolic Syndrome in Individuals with Hyperuricemia. *American Journal of Medicine* **120**, 442–447
- Choi, H.K., Ford, E.S., Li, C. and Curhan, G. (2007) Prevalence of the metabolic syndrome in patients with gout: the Third National Health and Nutrition Examination Survey. *Arthritis and Rheumatism* **57**, 109–115
- Chung, S.A., Taylor, K.E., Graham, R.R., Nititham, J., Lee, A.T., Ortmann, W.A., Jacob, C.O. et al. (2011) Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genetics* **7**, e1001323
- Clark, J.T. and Kelly, K.M. (1993) Human genetics, paleoenvironments, and malaria: Relationships and implications for the settlement of Oceania. *American Anthropologist* **95**, 612–630

- Cleophas, M.C., Joosten, L.A., Stamp, L.K., Dalbeth, N., Woodward, O.M. and Merriman, T.R. (2017) *ABCG2* polymorphisms in gout: Insights into disease susceptibility and treatment approaches. *Pharmacogenomics and Personalized Medicine* **10**, 129–142
- Cockerham, C.C. (1969) Variance of Gene Frequencies. *Evolution* **23**, 72–84
- Cockerham, C.C. (1973) Analyses of gene frequencies. *Genetics* **74**, 679–700
- Colhoun, H.M., McKeigue, P.M. and Smith, G.D. (2003) Problems of reporting genetic associations with complex outcomes. *The Lancet* **361**, 865–872
- Collins, J.F., Tutone, V. and Walker, C. (2017) *Kidney Disease in Maori and Pacific people in New Zealand*. Elsevier Inc.
- Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y. and Tyler-Smith, C. (2014) Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology* **15**, R88
- Comabella, M., Craig, D.W., Camiña-Tato, M., Morcillo, C., Lopez, C., Navarro, A., Rio, J., Montalban, X. and Martin, R. (2008) Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms. *PloS One* **3**, e3490
- Comuzzie, A.G., Cole, S.A., Laston, S.L., Voruganti, V.S., Haack, K., Gibbs, R.A. and Butte, N.F. (2012) Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS One* **7**, e51954
- Cook, J.P. and Morris, A.P. (2016) Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *European Journal of Human Genetics* **24**, 1175–1180
- Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. and Pritchard, J.K. (2009) The role of geography in human adaptation. *PLoS Genetics* **5**, e1000500
- Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C., Warram, J.H. and Todd, J.A. (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics* **40**, 1399–1401
- Cordell, H.J., Han, Y., Mells, G.F., Li, Y., Hirschfield, G.M., Greene, C.S., Xie, G. et al. (2015) International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature Communications* **6**, 8019
- Cotsapas, C., Speliotes, E.K., Hatoum, I.J., Greenawalt, D.M., Dobrin, R., Lum, P.Y., Suver, C., Chudin, E., Kemp, D., Reitman, M., Voight, B.F., Neale, B.M., Schadt, E.E., Hirschhorn, J.N., Kaplan, L.M. and Daly, M.J. (2009) Common body mass index-associated variants confer risk of extreme obesity. *Human Molecular Genetics* **18**, 3502–3507

- Cserti, C.M. and Dzik, W.H. (2015) Review article The ABO blood group system and *Plasmodium falciparum* malaria. *Blood* **110**, 2250–2259
- Cui, B., Zhu, X., Xu, M., Guo, T., Zhu, D., Chen, G., Li, X., Xu, L., Bi, Y., Chen, Y., Xu, Y., Li, X., Wang, W., Wang, H., Huang, W. and Ning, G. (2011) A genome-wide association study confirms previously reported loci for type 2 diabetes in Han Chinese. *PLoS One* **6**, e22353
- Dalbeth, N., Schumacher, H.R., Fransen, J., Neogi, T., Jansen, T.L., Brown, M., Louthrenoo, W. et al. (2016) Survey definitions of gout for epidemiologic studies: Comparison with crystal identification as the gold standard. *Arthritis Care and Research* **68**, 1894–1898
- Darwin, C. (1909) The Origin of Species, volume 11 of The Harvard Classics. *PF Collier and Son* p. 94
- Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. and Excoffier, L. (2013) Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution* **30**, 1544–1558
- De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I.W., Ottoboni, L., Aggarwal, N.T., Piccio, L. et al. (2009) Meta-analysis of genome scans and replication identify *CD6*, *IRF8* and *TNFRSF1A* as new multiple sclerosis susceptibility loci. *Nature Genetics* **41**, 776–782
- de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L. et al. (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* **49**, 256–261
- Dehghan, A., Köttgen, A., Yang, Q., Hwang, S.J., Kao, W.L., Rivadeneira, F., Boerwinkle, E., Levy, D., Hofman, A., Astor, B.C., Benjamin, E.J., van Duijn, C.M., Witteman, J.C., Coresh, J. and Fox, C.S. (2008) Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *The Lancet* **372**, 1953–1961
- Delaneau, O., Marchini, J. and Zagury, J.F. (2012) A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–81
- Delaneau, O., Zagury, J.F. and Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6
- Delaneau, O., Marchini, J., McVean, G.a., Donnelly, P., Lunter, G., Marchini, J.L., Myers, S. et al. (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5**, 1–9
- Demirci, F.Y., Wang, X., Kelly, J.A., Morris, D.L., Barmada, M.M., Feingold, E., Kao, A.H., Sivils, K.L., Bernatsky, S., Pineau, C., Clarke, A.E., Ramsey-Goldman, R., Vyse, T.J., Gaffney, P.M., Manzi, S. and Kamboh, M.I. (2016) Identification of a new susceptibility locus for systemic lupus erythematosus on chromosome 12 in individuals of European ancestry. *Arthritis & Rheumatology* **68**, 174–83
- Depaulis, F., Mousset, S. and Veuille, M. (2003) Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution* **57**, 190–200

- Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E. and Quintana-Murci, L. (2016) Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *American Journal of Human Genetics* **98**, 5–21
- Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W., Wojcicki, A. and Eriksson, N. (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics* **7**, e1002141
- Döring, A., Gieger, C., Mehta, D., Gohlke, H., Prokisch, H., Coassini, S., Fischer, G., Henke, K., Klopp, N., Kronenberg, F., Paulweber, B., Pfeufer, A., Rosskopf, D., Völzke, H., Illig, T., Meitinger, T., Wichmann, H.E. and Meisinger, C. (2008) *SLC2A9* influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics* **40**, 430–436
- Dubois, P.C.A., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A. et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics* **42**, 295–302
- Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H. et al. (2006) A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463
- Duggan, A.T. and Stoneking, M. (2014) Recent developments in the genetic history of East Asia and Oceania. *Current Opinion in Genetics & Development* **29**, 9–14
- Duggan, A.T., Evans, B., Friedlaender, F.R., Friedlaender, J.S., Koki, G., Merriwether, D.A., Kayser, M. and Stoneking, M. (2014) Maternal history of Oceania from complete mtDNA genomes: Contrasting ancient diversity with recent homogenization due to the Austronesian expansion. *American Journal of Human Genetics* **94**, 721–733
- Ellinghaus, E., Ellinghaus, D., Stuart, P.E., Nair, R.P., Debrus, S., Raelson, J.V., Belouchi, M. et al. (2010) Genome-wide association study identifies a psoriasis susceptibility locus at *TRAF3IP2*. *Nature Genetics* **42**, 991–995
- Euser, S.M., Hofman, A., Westendorp, R.G.J. and Breteler, M.M.B. (2009) Serum uric acid and cognitive function and dementia. *Brain* **132**, 377–82
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R. et al. (2018) The reactome pathway knowledgebase. *Nucleic Acids Research* **46**, D649–D655
- Fadista, J., Manning, A.K., Florez, J.C. and Groop, L. (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* **24**, 1202–1205
- Fay, J.C. and Wu, C.I. (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution* **16**, 1003–1005

- Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413
- Feller, W. (1951) Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* pp. 227–246
- Felsenstein, J. (1965) The effect of linkage on directional selection. *Genetics* **52**, 349–363
- Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R. (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution* **31**, 1275–91
- Ferrero, E. (2018) Using regulatory genomics data to interpret the function of disease variants and prioritise genes from expression studies [version 2; referees: 2 approved]. *F1000Research* **7**
- Feulner, T.M., Laws, S.M., Friedrich, P., Wagenpfeil, S., Wurst, S.H.R., Riehle, C., Kuhn, K.A., Krawczak, M., Schreiber, S., Nikolaus, S., Förstl, H., Kurz, A. and Riemenschneider, M. (2010) Examination of the current top candidate genes for AD in a genome-wide association study. *Molecular Psychiatry* **15**, 756–766
- Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. and Pritchard, J.K. (2016a) Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764
- Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. and Pritchard, J.K. (2016b) Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764
- Fisher, R.A. (1919) XV.—The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433
- Fisher, R.A. (1922) On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**, 321–341
- Fisher, R.A. (1930) *The genetical theory of natural selection*. Oxford Univ Press
- Flint, J., Hill, A.V., Bowden, D.K., Oppenheimer, S.J., Sill, P.R., Serjeantson, S.W., Bana-Koiri, J., Bhatia, K., Alpers, M.P. and Boyce, A.J. (1986) High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750
- Foo, J.N., Tan, L.C., Irwan, I.D., Au, W.L., Low, H.Q., Prakash, K.M., Ahmad-Annuar, A. *et al.* (2017) Genome-wide association study of Parkinson's disease in East Asians. *Human Molecular Genetics* **26**, 226–232
- Franke, A., Hampe, J., Rosenstiel, P., Becker, C., Wagner, F., Hässler, R., Little, R.D. *et al.* (2007) Systematic association mapping identifies *NELL1* as a novel IBD disease gene. *PloS One* **2**, e691
- Franke, A., Balschun, T., Karlsen, T.H., Sventoraityte, J., Nikolaus, S., Mayr, G., Domingues, F.S. *et al.* (2008) Sequence variants in *IL10*, *ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. *Nature Genetics* **40**, 1319–1323

- Franke, A., Balschun, T., Sina, C., Ellinghaus, D., Häsler, R., Mayr, G., Albrecht, M. *et al.* (2010a) Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (*IL17REL*). *Nature Genetics* **42**, 292–294
- Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W. *et al.* (2010b) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42**, 1118–1125
- Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B. *et al.* (2007) A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894
- Freudenberg, J., Lee, H.S., Han, B.G., Shin, H.D., Kang, Y.M., Sung, Y.K., Shim, S.C., Choi, C.B., Lee, A.T., Gregersen, P.K. and Bae, S.C. (2011) Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis and Rheumatism* **63**, 884–893
- Friedlaender, J., Friedlaender, F. and Reed, F. (2008) The genetic structure of Pacific Islanders. *PLoS Genetics* **4**, 174–190
- Fu, W., O'Connor, T.D. and Akey, J.M. (2013) Genetic architecture of quantitative traits and complex diseases. *Current Opinion in Genetics & Development* **23**, 678–83
- Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709
- Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jorgensen, M.E., Korneliussen, T.S., Gerbault, P., Skotte, L., Linneberg, A., Christensen, C., Brandslund, I., Jorgensen, T., Huerta-Sanchez, E., Schmidt, E.B., Pedersen, O., Hansen, T., Albrechtsen, A. and Nielsen, R. (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347
- Gallego-Delgado, J., Ty, M., Orengo, J.M., van de Hoef, D. and Rodriguez, A. (2014) A surprising role for uric acid: The inflammatory malaria response. *Current Rheumatology Reports* **16**, 401
- Garner, C., Ahn, R., Ding, Y.C., Steele, L., Stoven, S., Green, P.H., Fasano, A., Murray, J.A. and Neuhausen, S.L. (2014) Genome-wide association study of celiac disease in North America confirms *FRMD4B* as new celiac locus. *PloS One* **9**, e101428
- Ghassibe-Sabbagh, M., Haber, M., Salloum, A.K., Al-Sarraj, Y., Akle, Y., Hirbli, K., Romanos, J., Mouzaya, F., Gauguier, D., Platt, D.E., El-Shanti, H. and Zalloua, P.A. (2014) T2DM GWAS in the Lebanese population confirms the role of *TCF7L2* and *CDKAL1* in disease susceptibility. *Scientific Reports* **4**, 7351
- Glantzounis, G., Tsimogiannis, E., Kappas, A. and Galaris, D. (2005) Uric acid and oxidative stress. *Current Pharmaceutical Design* **11**, 4145–4151
- Gorski, M., Tin, A., Garnaas, M., McMahon, G.M., Chu, A.Y., Tayo, B.O., Pattaro, C. *et al.* (2015) Genome-wide association study of kidney function decline in individuals of European descent. *Kidney International* **87**, 1017–1029

- Gosling, A.L., Matisoo-Smith, E. and Merriman, T.R. (2014) Hyperuricaemia in the Pacific: why the elevated serum urate levels? *Rheumatology international* **34**, 743–57
- Gosling, A.L., Buckley, H.R., Matisoo-Smith, E. and Merriman, T.R. (2015) Pacific populations, metabolic disease and ‘just-so stories’: A critique of the ‘thrifty genotype’ hypothesis in Oceania. *Annals of Human Genetics* **79**, 470–480
- Gourraud, P.A., Sdika, M., Khankhanian, P., Henry, R.G., Beheshtian, A., Matthews, P.M., Hauser, S.L., Oksenberg, J.R., Pelletier, D. and Baranzini, S.E. (2013) A genome-wide association study of brain lesion distribution in multiple sclerosis. *Brain* **136**, 1012–1024
- Graff, M., Ngwa, J.S., Workalemahu, T., Homuth, G., Schipf, S., Teumer, A., Völzke, H. *et al.* (2013) Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Human Molecular Genetics* **22**, 3597–3607
- Graham, R.R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C.J., Leon, J.M., Burtt, N.P. *et al.* (2008) Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nature Genetics* **40**, 1059–1061
- Grant, S.F.A., Qu, H.Q., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R. *et al.* (2009) Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. *Diabetes* **58**, 290–295
- Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F. *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11983–11988
- Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guillet, W., Dutil, J., Via, M., Sandoval, K., Bedoya, G., Oleksyk, T.K., Ruiz-Linares, A., Burchard, E.G., Martinez-Cruzado, J.C. and Bustamante, C.D. (2013) Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLoS Genetics* **9**
- Gregersen, P.K., Amos, C.I., Lee, A.T., Lu, Y., Remmers, E.F., Kastner, D.L., Seldin, M.F., Criswell, L.A., Plenge, R.M., Holers, V.M., Mikuls, T.R., Sokka, T., Moreland, L.W., Bridges, S.L., Xie, G., Begovich, A.B. and Siminovitch, K.A. (2009) REL, encoding a member of the NF- κ B family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nature Genetics* **41**, 820–823
- Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E.S., Schaffner, S.F. and Sabeti, P.C. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–6
- Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., Cabili, M., Adegbola, R.a., Bamezai, R.N.K., Hill, A.V.S., Vannberg, F.O., Rinn, J.L., Lander, E.S., Schaffner, S.F. and Sabeti, P.C. (2013) Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–13

- Groucott, H.S., Petraglia, M.D., Bailey, G., Scerri, E.M.L., Parton, A., Clark-Balzan, L., Jennings, R.P., Lewis, L., Blinkhorn, J., Drake, N.A., Breeze, P.S., Inglis, R.H., Devès, M.H., Meredith-Williams, M., Boivin, N., Thomas, M.G. and Scally, A. (2015) Rethinking the dispersal of *Homo sapiens* out of Africa. *Evolutionary Anthropology: Issues, News, and Reviews* **24**, 149–164
- Gudbjartsson, D.F., Holm, H., Indridason, O.S., Thorleifsson, G., Edvardsson, V., Sulem, P., de Vegt, F. et al. (2010) Association of variants at *UMOD* with chronic kidney disease and kidney stones-role of age and comorbid diseases. *PLoS Genetics* **6**, e1001039
- Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.a., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S. et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **47**, 435–444
- Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37**, 549–554
- Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., Samuels, D.C., Shyr, Y. and Long, J. (2014) Illumina human exome genotyping array clustering and quality control. *Nature Protocols* **9**, 2643–2662
- Haasl, R.J. and Payseur, B.A. (2016) Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology* **25**, 5–23
- Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I.W., Gabriel, S.B., Mirel, D.B., Ivinson, A.J., Pericak-Vance, M.A., Gregory, S.G., Rioux, J.D., McCauley, J.L., Haines, J.L., Barcellos, L.F., Cree, B., Oksenberg, J.R. and Hauser, S.L. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *The New England Journal of Medicine* **357**, 851–862
- Haga, S.B. (2010) Impact of limited population diversity of genome-wide association studies. *Genetics in Medicine* **12**, 81–84
- Hahsler, M. and Piekenbrock, M. (2017) *dbSCAN: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*. Available: <https://CRAN.R-project.org/package=dbSCAN>. R package version 1.1-1
- Hakonarson, H., Grant, S.F.A., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R. et al. (2007) A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. *Nature* **448**, 591–594
- Hakonarson, H., Qu, H.Q., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R. et al. (2008) A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **57**, 1143–1146
- Hamza, T.H., Zabetian, C.P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D.M., Doheny, K.F., Paschall, J., Pugh, E., Kusel, V.I., Collura, R., Roberts, J., Griffith, A., Samii, A., Scott, W.K., Nutt, J., Factor, S.A. and Payami, H. (2010) Common genetic variation in the *HLA* region is associated with late-onset sporadic Parkinson's disease. *Nature Genetics* **42**, 781–785

- Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H. *et al.* (2009) Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genetics* **41**, 1234–1237
- Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G. and Di Rienzo, A. (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics* **4**, 1–13
- Hancock, A.M., Witonsky, D.B., Alkorta-Aranburu, G., Beall, C.M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J.K., Coop, G. and Di Rienzo, A. (2011) Adaptations to climate-mediated selective pressures in humans. *PLoS Genetics* **7**, 1–16
- Hanson, R.L., Muller, Y.L., Kobes, S., Guo, T., Bian, L., Ossowski, V., Wiedrich, K., Sutherland, J., Wiedrich, C., Mahkee, D., Huang, K., Abdussamad, M., Traurig, M., Weil, E.J., Nelson, R.G., Bennett, P.H., Knowler, W.C., Bogardus, C. and Baier, L.J. (2014) A genome-wide association study in American Indians implicates *DNER* as a susceptibility locus for type 2 diabetes. *Diabetes* **63**, 369–376
- Hara, K., Fujita, H., Johnson, T.A., Yamauchi, T., Yasuda, K., Horikoshi, M., Peng, C. *et al.* (2014) Genome-wide association study identifies three novel loci for type 2 diabetes. *Human Molecular Genetics* **23**, 239–246
- Haritunians, T., Taylor, K.D., Targan, S.R., Dubinsky, M., Ippoliti, A., Kwon, S., Guo, X., Melmed, G.Y., Berel, D., Mengesha, E., Psaty, B.M., Glazer, N.L., Vasiliauskas, E.A., Rotter, J.I., Fleshner, P.R. and McGovern, D.P.B. (2010) Genetic predictors of medically refractory ulcerative colitis. *Inflammatory Bowel Diseases* **16**, 1830–1840
- Harley, J.B., Alarcón-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P. *et al.* (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nature Genetics* **40**, 204–210
- Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S. *et al.* (2009) Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. *Nature Genetics* **41**, 1088–1093
- Haslam, D.W. and James, W.P.T. (2005) Obesity. *The Lancet* **366**, 1197–1209
- Hawley, N.L. and McGarvey, S.T. (2015) Obesity and diabetes in Pacific Islanders: the current burden and the need for urgent action. *Current Diabetes Reports* **15**
- Heinzen, E.L., Need, A.C., Hayden, K.M., Chiba-Falek, O., Roses, A.D., Strittmatter, W.J., Burke, J.R., Hulette, C.M., Welsh-Bohmer, K.A. and Goldstein, D.B. (2010) Genome-wide scan of copy number variation in late-onset Alzheimer's disease. *Journal of Alzheimer's Disease* **19**, 69–77
- Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014) A genetic atlas of human admixture history. *Science* **343**, 747–751

- Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S. and Przeworski, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics* **72**, 1527–1535
- Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J.E., Willemsen, G., Hottenga, J.J., Abdellaoui, A. et al. (2013) Inference of the genetic architecture underlying bmi and height with the use of 20,240 sibling pairs. *American Journal of Human Genetics* **93**, 865–875
- Hermission, J. and Pennings, P.S. (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–52
- Hermission, J. and Pennings, P.S. (2017) Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution* **8**, 700–716
- Hider, J.L., Gittelman, R.M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J.M. and Parra, E.J. (2013) Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evolutionary Biology* **13**, 150
- Hildebrandt, F. (2010) Genetic kidney diseases. *The Lancet* **375**, 1287–1295
- Hill, A.V., Bowden, D.K., Trent, R.J., Higgs, D.R., Oppenheimer, S.J., Thein, S.L., Mickleson, K.N., Weatherall, D.J. and Clegg, J.B. (1985) Melanesians and Polynesians share a unique alpha-thalassemia mutation. *American Journal of Human Genetics* **37**, 571–80
- Hill, W.G., Goddard, M.E. and Visscher, P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**, 1–10
- Hill-Burns, E.M., Wissemann, W.T., Hamza, T.H., Factor, S.A., Zabetian, C.P. and Payami, H. (2014) Identification of a novel Parkinson's disease locus via stratified genome-wide association study. *BMC Genomics* **15**, 118
- Hirschfield, G.M., Liu, X., Xu, C., Lu, Y., Xie, G., Lu, Y., Gu, X. et al. (2009) Primary biliary cirrhosis associated with *HLA*, *IL12A*, and *IL12RB2* variants. *The New England Journal of Medicine* **360**, 2544–2555
- Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R. et al. (2011) Common variants at *ABCA7*, *MS4A6A/MS4A4E*, *EPHA1*, *CD33* and *CD2AP* are associated with Alzheimer's disease. *Nature Genetics* **43**, 429–435
- Hollis-Moffatt, J., Phipps-Green, A., Chapman, B., Jones, G., van Rij, A., Gow, P., Harrison, A., Highton, J., Jones, P., Montgomery, G., Stamp, L., Dalbeth, N. and Merriman, T. (2012) The renal urate transporter *SLC17A1* locus: confirmation of association with gout. *Arthritis Research & Therapy* **14**, R92
- Hollis-Moffatt, J.E., Xu, X., Dalbeth, N., Merriman, M.E., Topless, R., Waddell, C., Gow, P.J., Harrison, A.A., Highton, J., Jones, P.B.B., Stamp, L.K. and Merriman, T.R. (2009) Role of the

urate transporter *SLC2A9* gene in susceptibility to gout in New Zealand Māori, Pacific Island, and Caucasian case-control sample sets. *Arthritis & Rheumatism* **60**, 3485–3492

Hollis-Moffatt, J.E., Gow, P.J., Harrison, A.a., Highton, J., Jones, P.B.B., Stamp, L.K., Dalbeth, N. and Merriman, T.R. (2011) The *SLC2A9* nonsynonymous Arg265His variant and gout: evidence for a population-specific effect on severity. *Arthritis Research & Therapy* **13**, R85

Hom, G., Graham, R.R., Modrek, B., Taylor, K.E., Ortmann, W., Garnier, S., Lee, A.T. et al. (2008) Association of systemic lupus erythematosus with *C8orf13-BLK* and *ITGAM-ITGAX*. *The New England Journal of Medicine* **358**, 900–909

Howie, B., Marchini, J. and Stephens, M. (2011) Genotype imputation with thousands of genomes. *G3* **1**, 457–70

Hu, H.J., Jin, E.H., Yim, S.H., Yang, S.Y., Jung, S.H., Shin, S.H., Kim, W.U., Shim, S.C., Kim, T.G. and Chung, Y.J. (2011) Common variants at the promoter region of the *APOM* confer a risk of rheumatoid arthritis. *Experimental & Molecular Medicine* **43**, 613–621

Huang, J., Ellinghaus, D., Franke, A., Howie, B. and Li, Y. (2012) 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *European Journal of Human Genetics* **20**, 801–805

Hudjashov, G., Karafet, T.M., Lawson, D.J., Downey, S., Savina, O., Sudoyo, H., Lansing, J.S., Hammer, M.F. and Cox, M.P. (2017) Complex patterns of admixture across the Indonesian Archipelago. *Molecular Biology and Evolution* **34**, 2439–2452

Hudjashov, G., Endicott, P., Post, H., Nagle, N., Ho, S.Y.W., Lawson, D.J., Reidla, M., Karmin, M., Rootsi, S., Metspalu, E., Saag, L., Villems, R., Cox, M.P., Mitchell, R.J., Garcia-Bertrand, R.L., Metspalu, M. and Herrera, R.J. (2018) Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. *Scientific Reports* **8**, 1–12

Hudson, R.R., Kreitman, M. and Aguadé, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159

Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X. et al. (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197

Huffman, J.E., Albrecht, E., Teumer, A., Mangino, M., Kapur, K., Johnson, T., Kutalik, Z. et al. (2015) Modulation of genetic associations with serum urate levels by body-mass-index in humans. *PloS One* **10**, e0119752

Hughes, a.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170

Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J. et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics* **40**, 395–402

- Ilardo, M.A., Moltke, I., Korneliussen, T.S., Cheng, J., Stern, A.J., Racimo, F., de Barros Damgaard, P., Sikora, M., Seguin-Orlando, A., Rasmussen, S., van den Munckhof, I.C., ter Horst, R., Joosten, L.A., Netea, M.G., Salingkat, S., Nielsen, R. and Willerslev, E. (2018) Physiological and genetic adaptations to diving in sea nomads. *Cell* **173**, 569–580
- Imamura, M., Maeda, S., Yamauchi, T., Hara, K., Yasuda, K., Morizono, T., Takahashi, A. *et al.* (2012) A single-nucleotide polymorphism in *ANK1* is associated with susceptibility to type 2 diabetes in Japanese populations. *Human Molecular Genetics* **21**, 3042–3049
- Imamura, M., Takahashi, A., Yamauchi, T., Hara, K., Yasuda, K., Grarup, N., Zhao, W. *et al.* (2016) Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nature Communications* **7**, 10531
- Investigators, T.A. (1989) The Atherosclerosis Risk in Community (ARIC) study: design and objectives. *American Journal of Epidemiology* **129**, 687–702
- Jakkula, E., Leppä, V., Sulonen, A.M., Varilo, T., Kallio, S., Kemppinen, A., Purcell, S. *et al.* (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in *STAT3* gene. *American Journal of Human Genetics* **86**, 285–291
- Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K. *et al.* (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics* **41**, 657–665
- Jensen, J.D., Kim, Y., DuMont, V.B., Aquadro, C.F. and Bustamante, C.D. (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401–1410
- Jia, P., Wang, L., Meltzer, H.Y. and Zhao, Z. (2011) Pathway-based analysis of GWAS datasets: effective but caution required. *The International Journal of Neuropsychopharmacology* **14**, 567–572
- Jiang, L., Yin, J., Ye, L., Yang, J., Hemani, G., Liu, A.J., Zou, H. *et al.* (2014) Novel risk loci for rheumatoid arthritis in Han Chinese and congruence with risk variants in Europeans. *Arthritis & Rheumatology* **66**, 1121–1132
- Jiao, H., Arner, P., Hoffstedt, J., Brodin, D., Dubern, B., Czernichow, S., van't Hooft, F., Axelsson, T., Pedersen, O., Hansen, T., Sørensen, T.I.A., Hebebrand, J., Kere, J., Dahlman-Wright, K., Hamsten, A., Clement, K. and Dahlman, I. (2011) Genome wide association study identifies *KCNMA1* contributing to human obesity. *BMC Medical Genomics* **4**, 51
- Jin, Y., Birlea, S.A., Fain, P.R., Gowan, K., Riccardi, S.L., Holland, P.J., Mailloux, C.M. *et al.* (2010) Variant of *TYR* and autoimmunity susceptibility loci in generalized vitiligo. *The New England Journal of Medicine* **362**, 1686–1697
- Jin, Y., Birlea, S.A., Fain, P.R., Gowan, K., Riccardi, S.L., Holland, P.J., Bennett, D.C. *et al.* (2011) Genome-wide analysis identifies a quantitative trait locus in the MHC class II region associated with generalized vitiligo age of onset. *The Journal of Investigative Dermatology* **131**, 1308–1312

- Jin, Y., Birlea, S.A., Fain, P.R., Ferrara, T.M., Ben, S., Riccardi, S.L., Cole, J.B. *et al.* (2012) Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature Genetics* **44**, 676–680
- Jonnalagadda, M., Bharti, N., Patil, Y., Ozarkar, S., K, S.M., Joshi, R. and Norton, H. (2017) Identifying signatures of positive selection in pigmentation genes in two South Asian populations. *American Journal of Human Biology* pp. 1–10
- Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P.V., Snaedal, J., Bjornsson, S. *et al.* (2013) Variant of *TREM2* associated with the risk of Alzheimer's disease. *The New England Journal of Medicine* **368**, 107–116
- Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124
- Julia, A., Ballina, J., Cañete, J.D., Balsa, A., Tornero-Molina, J., Naranjo, A., Alperi-López, M., Erra, A., Pascual-Salcedo, D., Barceló, P., Camps, J. and Marsal, S. (2008) Genome-wide association study of rheumatoid arthritis in the Spanish population: *KLF12* as a risk locus for rheumatoid arthritis susceptibility. *Arthritis and Rheumatism* **58**, 2275–2286
- Julia, A., Domènech, E., Ricart, E., Tortosa, R., García-Sánchez, V., Gisbert, J.P., Mateu, P.N. *et al.* (2013) A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at *RBX1-EP300*. *Gut* **62**, 1440–1445
- Julia, A., Domènech, E., Chaparro, M., García-Sánchez, V., Gomollón, F., Panés, J., Mañosa, M. *et al.* (2014) A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis. *Human Molecular Genetics* **23**, 6927–6934
- Jung, E.S., Cheon, J.H., Lee, J.H., Park, S.J., Jang, H.W., Chung, S.H., Park, M.H., Kim, T.G., Oh, H.B., Yang, S.K., Park, S.H., Han, J.Y., Hong, S.P., Kim, T.I., Kim, W.H. and Lee, M.G. (2016) HLA-C*01 is a Risk Factor for Crohn's Disease. *Inflammatory Bowel Diseases* **22**, 796–806
- Juyal, G., Negi, S., Sood, A., Gupta, A., Prasad, P., Senapati, S., Zaneveld, J., Singh, S., Midha, V., van Sommeren, S., Weersma, R.K., Ott, J., Jain, S., Juyal, R.C. and Thelma, B.K. (2015) Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* **64**, 571–579
- Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y. and Kamatani, N. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genetics* **42**, 210–215
- Kamboh, M.I., Demirci, F.Y., Wang, X., Minster, R.L., Carrasquillo, M.M., Pankratz, V.S., Younkin, S.G. *et al.* (2012) Genome-wide association study of Alzheimer's disease. *Translational Psychiatry* **2**, e117
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361

- Kaplan, N.L., R., Hudson and Langley, C.H. (1989) The "hitch-hiking effect" revisited. *Genetics* **13**, 887–899
- Karlsson, E.K., Kwiatkowski, D.P. and Sabeti, P.C. (2014) Natural selection and infectious disease in human populations. *Nature Reviews Genetics* **15**, 379–93
- Karns, R., Viali, S., Tuitele, J., Sun, G., Cheng, H., Weeks, D.E., Mcgarvey, S.T. and Deka, R. (2012) Common variants in *FTO* are not significantly associated with obesity-related phenotypes among Samoans of Polynesia. *Annals of Human Genetics* **76**, 17–24
- Kawashima, M., Hitomi, Y., Aiba, Y., Nishida, N., Kojima, K., Kawai, Y., Nakamura, H. et al. (2017) Genome-wide association studies identify *PRKCB* as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Human Molecular Genetics* **26**, 650–659
- Kayser, M. (2010) The human genetic history of Oceania: near and remote views of dispersal. *Current Biology* **20**, R194–R201
- Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L.A., Moyse-Faurie, C., Rutledge, R.B., Schiefenhoevel, W., Gil, D., Lin, A.a., Underhill, P.a., Oefner, P.J., Trent, R.J. and Stoneking, M. (2006) Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Molecular Biology and Evolution* **23**, 2234–2244
- Kayser, M., Lao, O., Saar, K., Brauer, S., Wang, X., Nürnberg, P., Trent, R.J. and Stoneking, M. (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics* **82**, 194–8
- Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. and Akey, J.M. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research* **16**, 980–989
- Kenny, E.E., Pe'er, I., Karban, A., Ozelius, L., Mitchell, A.A., Ng, S.M., Erazo, M. et al. (2012) A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genetics* **8**, e1002559
- Kerr, J.M., Fisher, L.W., Termine, J.D., Wang, M.G., McBride, O.W. and Young, M.F. (1993) The human bone sialoprotein gene (IBSP): genomic localization and characterization. *Genomics* **17**, 408–415
- Khalili, H., Sull, A., Sarin, S., Boivin, F.J., Halabi, R., Svajger, B., Li, A., Cui, V.W., Drysdale, T. and Bridgewater, D. (2016) Developmental origins for kidney disease due to Shroom3 deficiency. *Journal of the American Society of Nephrology* **27**, 2965–2973
- Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777
- Kimura, M. (1955) Random Genetic Drift in Multi-Allelic Locus. *Evolution* **9**, 419–435
- Kimura, M. (1968) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269

- Kimura, M. (1979a) Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 3440–4
- Kimura, M. (1979b) The neutral theory of molecular evolution. *Scientific American* **241**, 98–126
- Kimura, M. and Crow, J. (1963) The measurement of effective population number. *Evolution* **17**, 279–288
- Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M. and Tokunaga, K. (2008) Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Molecular Biology and Evolution* **25**, 1750–61
- Kingman, J.F.C. (1982) On the genealogy of large populations. *Journal of Applied Probability* **19**, 27–43
- Kivisild, T., Bamshad, M., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W., Dixon, M., Papiha, S., Mastana, S., Mir, M., Ferak, V. and Villems, R. (1999) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Current Biology* **9**, 1331–1334
- Kochi, Y., Okada, Y., Suzuki, A., Ikari, K., Terao, C., Takahashi, A., Yamazaki, K. et al. (2010) A regulatory variant in *CCR6* is associated with rheumatoid arthritis susceptibility. *Nature Genetics* **42**, 515–519
- Koh, X.H., Liu, X. and Teo, Y.Y. (2014) Can evidence from genome-wide association studies and positive natural selection surveys be used to evaluate the thrifty gene hypothesis in East Asians? *PloS One* **9**, e110974
- Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F. et al. (2011) Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nature Genetics* **43**, 984–989
- Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 29
- Köttgen, A., Glazer, N.L., Dehghan, A., Hwang, S.J., Katz, R., Li, M., Yang, Q. et al. (2009) Multiple loci associated with indices of renal function and chronic kidney disease. *Nature Genetics* **41**, 712–717
- Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A. et al. (2010) New loci associated with kidney function and chronic kidney disease. *Nature Genetics* **42**, 376–384
- Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis, G. et al. (2013) Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics* **45**, 145–54
- Kozyrev, S.V., Abelson, A.K., Wojcik, J., Zaghlool, A., Reddy, M.V.P.L., Sanchez, E., Gunnarsson, I. et al. (2008) Functional variants in the B-cell gene *BANK1* are associated with systemic lupus erythematosus. *Nature Genetics* **40**, 211–216

- Kraja, A.T., Vaidya, D., Pankow, J.S., Goodarzi, M.O., Assimes, T.L., Kullo, I.J., Sovio, U. *et al.* (2011) A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. *Diabetes* **60**, 1329–1339
- Kratzer, J.T., Lanaspa, M.a., Murphy, M.N., Cicerchi, C., Graves, C.L., Tipton, P.A., Ortlund, E.A., Johnson, R.J. and Gaucher, E.a. (2014) Evolutionary history and metabolic insights of ancient mammalian uricases. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 3763–8
- Kreitman, M. (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**, 539–559
- Krishnan, M., Major, T.J., Topless, R.K., Dewes, O., Yu, L., Thompson, J.M.D., McCowan, L. *et al.* (2018) Discordant association of the *CREBRF* rs373863828 A allele with increased BMI and protection from type 2 diabetes in Māori and Pacific (Polynesian) people living in Aotearoa/New Zealand. *Diabetologia* **61**, 1603–1613
- Kristiansson, K., Perola, M., Tikkanen, E., Kettunen, J., Surakka, I., Havulinna, A.S., Stancáková, A. *et al.* (2012) Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circulation. Cardiovascular Genetics* **5**, 242–249
- Kugathasan, S., Baldassano, R.N., Bradfield, J.P., Sleiman, P.M.A., Imielinski, M., Guthery, S.L., Cucchiara, S. *et al.* (2008) Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nature Genetics* **40**, 1211–1215
- Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W. and Ma'ayan, A. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90–W97
- Kutzing, M.K. and Firestein, B.L. (2008) Altered uric acid levels and disease states. *Pharmacology and Experimental Therapeutics* **324**, 1–7
- Lai, H.M., Chen, C.J., Su, B.Y.J., Chen, Y.C., Yu, S.F., Yen, J.H., Hsieh, M.C., Cheng, T.T. and Chang, S.J. (2012) Gout and type 2 diabetes have a mutual inter-dependent effect on genetic risk factors and higher incidences. *Rheumatology* **51**, 715–720
- Lambert, J.C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O. *et al.* (2009) Genome-wide association study identifies variants at *CLU* and *CR1* associated with Alzheimer's disease. *Nature Genetics* **41**, 1094–1099
- Lambert, J.C., Grenier-Boley, B., Harold, D., Zelenika, D., Chouraki, V., Kamatani, Y., Sleegers, K. *et al.* (2013) Genome-wide haplotype association study identifies the *FRMD4A* gene as a risk locus for Alzheimer's disease. *Molecular Psychiatry* **18**, 461–470
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921

- Lee, H.J., Kim, J., Lee, T., Son, J.K., Yoon, H.B., Baek, K.S., Jeong, J.Y., Cho, Y.M., Lee, K.T., Yang, B.C., Lim, H.J., Cho, K., Kim, T.H., Kwon, E.G., Nam, J., Kwak, W., Cho, S. and Kim, H. (2014) Deciphering the genetic blueprint behind Holstein milk proteins and production. *Genome Biology and Evolution* **6**, 1366–1374
- Lee, S.H., Wray, N.R., Goddard, M.E. and Visscher, P.M. (2011) Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**, 294–305
- Lee, Y.H., Bae, S.C., Choi, S.J., Ji, J.D. and Song, G.G. (2012) Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. *Molecular Biology Reports* **39**, 10627–10635
- Lessard, C.J., Sajuthi, S., Zhao, J., Kim, K., Ice, J.A., Li, H., Ainsworth, H. et al. (2016) Identification of a Systemic Lupus Erythematosus Risk Locus Spanning *ATG16L2*, *FCHSD2*, and *P2RY2* in Koreans. *Arthritis & Rheumatology* **68**, 1197–1209
- Li, C., Li, Z., Liu, S., Wang, C., Han, L., Cui, L., Zhou, J. et al. (2015) Genome-wide association analysis identifies three new risk loci for gout arthritis in Han Chinese. *Nature Communications* **6**, 7041
- Li, H., Ruan, J. and Durbin, R. (2008a) Mapping short DNA sequencing reads and calling variants using mapping. *Genome Research* pp. 1851–1858
- Li, H., Wetten, S., Li, L., Jean, P.L.S., Upmanyu, R., Surh, L., Hosford, D. et al. (2008b) Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Archives of Neurology* **65**, 45–53
- Li, H., Gan, W., Lu, L., Dong, X., Han, X., Hu, C., Yang, Z. et al. (2013) A genome-wide association study identifies *GRK5* and *RASGRP1* as type 2 diabetes loci in Chinese Hans. *Diabetes* **62**, 291–298
- Li, S., Sanna, S., Maschio, A., Busonero, F., Usala, G., Mulas, A., Lai, S. et al. (2007) The *GLUT9* gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genetics* **3**, e194
- Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S. et al. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genetics* **3**, e58
- Librado, P. and Orlando, L. (2018) Detecting signatures of positive selection along defined branches of a population tree using LSD. *Molecular Biology and Evolution* **35**, 1520–1535
- Lill, C.M., Roehr, J.T., McQueen, M.B., Kavvoura, F.K., Bagade, S., Schjeide, B.M.M., Schjeide, L.M. et al. (2012) Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS Genetics* **8**, e1002548
- Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S. et al. (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **47**, 979–986

- Liu, X., Invernizzi, P., Lu, Y., Kosoy, R., Lu, Y., Bianchi, I., Podda, M. *et al.* (2010) Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nature Genetics* **42**, 658–660
- Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C. *et al.* (2008) A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genetics* **4**, e1000041
- Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C. *et al.* (2015a) Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206
- Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C. *et al.* (2015b) Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206
- Loos, R.J.F., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M. *et al.* (2008) Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nature Genetics* **40**, 768–775
- Lu, X., Wang, L., Lin, X., Huang, J., Charles gu, C., He, M., Shen, H. *et al.* (2015) Genome-wide association study in Chinese identifies novel loci for blood pressure and hypertension. *Human Molecular Genetics* **24**, 865–874
- Ma, R.C.W., Hu, C., Tam, C.H., Zhang, R., Kwan, P., Leung, T.F., Thomas, G.N. *et al.* (2013) Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near *PAX4*. *Diabetologia* **56**, 1291–1305
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flück, P., Cunningham, F. and Parkinson, H. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896–D901
- Machiela, M.J. and Chanock, S.J. (2015) LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557
- Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M. *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* **46**, 234–244
- Maiga, B., Dolo, A., Touré, O., Dara, V., Tapily, A., Campino, S., Sepulveda, N., Risley, P., Silva, N., Corran, P., Rockett, K.A., Kwiatkowski, D., Clark, T.G., Troye-Blomberg, M. and Doumbo, O.K. (2013) Human candidate polymorphisms in sympatric ethnic groups differing in malaria susceptibility in Mali. *PLoS One* **8**, 1–11
- Major, T.J., Dalbeth, N., Stahl, E.A. and Merriman, T.R. (2018) An update on the genetics of hyperuricaemia and gout. *Nature Reviews Rheumatology* **2011**, 1–8

- Maliepaard, M., Scheffer, G.L., Faneyte, I.F., van Gastelen, M.A., Pijnenborg, A.C., Schinkel, A.H., van De Vijver, M.J., Schepers, R.J. and Schellens, J.H. (2001) Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. *Cancer Research* **61**, 3458–64
- Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M. et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206
- Mandal, A.K. and Mount, D.B. (2015) The molecular physiology of uric acid homeostasis. *Annual Review of Physiology* **77**, 323–345
- Martinelli-Boneschi, F., Esposito, F., Brambilla, P., Lindström, E., Lavorgna, G., Stankovich, J., Rodegher, M. et al. (2012) A genome-wide association study in progressive multiple sclerosis. *Multiple Sclerosis* **18**, 1384–1394
- Matesanz, F., González-Pérez, A., Lucas, M., Sanna, S., Gayán, J., Urcelay, E., Zara, I. et al. (2012) Genome-wide association study of multiple sclerosis confirms a novel locus at 5p13.1. *PloS One* **7**, e36140
- Matisoo-Smith, E. (2012) *On the great blue highway: Human migration in the Pacific*. Cambridge University Press, New York
- Matisoo-Smith, E. (2015) Ancient DNA and the human settlement of the Pacific: A review. *Journal of Human Evolution* **79**, 93–104
- Matisoo-Smith, E. and Gosling, A.L. (2018) Walking backwards into the future: the need for a holistic evolutionary approach in Pacific health research. *Annals of Human Biology* **45**, 175–187
- Matisoo-Smith, E. and Robins, J.H. (2004) Origins and dispersals of Pacific peoples: evidence from mtDNA phylogenies of the Pacific rat. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9167–72
- Matsuo, H., Yamamoto, K., Nakaoka, H., Nakayama, A., Sakiyama, M., Chiba, T., Takahashi, A. et al. (2016) Genome-wide association study of clinically defined gout identifies multiple risk loci and its association with clinical subtypes. *Annals of the Rheumatic Diseases* **75**, 652–659
- Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M. et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283
- McGovern, D.P.B., Gardet, A., Törkvist, L., Goyette, P., Essers, J., Taylor, K.D., Neale, B.M. et al. (2010a) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nature Genetics* **42**, 332–337

- McGovern, D.P.B., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A. *et al.* (2010b) Fucosyltransferase 2 (*FUT2*) non-secretor status is associated with Crohn's disease. *Human Molecular Genetics* **19**, 3468–3476
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., Others and DePristo, M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303
- McVean, G. (2007) The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**, 1395–1406
- Meda, S.A., Narayanan, B., Liu, J., Perrone-Bizzozero, N.I., Stevens, M.C., Calhoun, V.D., Glahn, D.C., Shen, L., Risacher, S.L., Saykin, A.J. and Pearlson, G.D. (2012) A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *NeuroImage* **60**, 1608–1621
- Melka, M.G., Bernard, M., Mahboubi, A., Abrahamowicz, M., Paterson, A.D., Syme, C., Lourdusamy, A., Schumann, G., Leonard, G.T., Perron, M., Richer, L., Veillette, S., Gaudet, D., Paus, T. and Pausova, Z. (2012) Genome-wide scan for loci of adolescent obesity and their relationship with blood pressure. *The Journal of Clinical Endocrinology and Metabolism* **97**, E145–150
- Mells, G.F., Floyd, J.A.B., Morley, K.I., Cordell, H.J., Franklin, C.S., Shin, S.Y., Heneghan, M.A. *et al.* (2011) Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nature Genetics* **43**, 329–332
- Merriman, T.R. and Dalbeth, N. (2011) The genetic basis of hyperuricaemia and gout. *Joint, Bone, Spine : Revue Du Rhumatisme* **78**, 35–40
- Metzker, M.L. (2010) Sequencing technologies the next generation. *Nature Reviews Genetics* **11**, 31–46
- Meyre, D., Delplanque, J., Chèvre, J.C., Lecoeur, C., Lobbens, S., Gallina, S., Durand, E. *et al.* (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics* **41**, 157–159
- Minster, R.L., Hawley, N.L., Su, C.T., Sun, G., Kershaw, E.E., Cheng, H., Buhule, O.D., Lin, J., Reupena, S., Viali, I., Tuitele, J., Naseri, T., Urban, Z., Deka, R., Weeks, D.E. and McGarvey, S.T. (2016) A thrifty variant in *CREBRF* strongly influences body mass index in Samoans. *Nature Publishing Group* **48**, 1–9
- Mochizuki, T., Wu, G., Hayashi, T., Xenophontos, S.L., Veldhuisen, B., Saris, J.J., Reynolds, D.M., Cai, Y., Gabow, P.A., Pierides, A., Kimberling, W.J., Breuning, M.H., Deltas, C.C., Peters, D.J.M. and Somlo, S. (1996) *PKD2*, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* **272**, 1339–1342
- Monda, K.L., Chen, G.K., Taylor, K.C., Palmer, C., Edwards, T.L., Lange, L.A., Ng, M.C.Y. *et al.* (2013) A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nature Genetics* **45**, 690–696

- Morris, A.P. (2011) Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology* **35**, 809–822
- Morris, D.L., Sheng, Y., Zhang, Y., Wang, Y.F., Zhu, Z., Tombleson, P., Chen, L. *et al.* (2016) Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nature Genetics* **48**, 940–946
- Myles, S., Hradetzky, E., Engelken, J., Lao, O., Nürnberg, P., Trent, R.J., Wang, X., Kayser, M. and Stoneking, M. (2007) Identification of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. *European Journal of Human Genetics* **15**, 584–589
- Myles, S., Lea, R.a., Ohashi, J., Chambers, G.K., Weiss, J.G., Hardouin, E., Engelken, J., Macartney-Coxson, D.P., Eccles, D.a., Naka, I., Kimura, R., Inaoka, T., Matsumura, Y. and Stoneking, M. (2011) Testing the thrifty gene hypothesis: the Gly482Ser variant in *PPARGC1A* is associated with BMI in Tongans. *BMC Medical Genetics* **12**, 10
- Myouzen, K., Kochi, Y., Okada, Y., Terao, C., Suzuki, A., Ikari, K., Tsunoda, T., Takahashi, A., Kubo, M., Taniguchi, A., Matsuda, F., Ohmura, K., Momohara, S., Mimori, T., Yamanaka, H., Kamatani, N., Yamada, R., Nakamura, Y. and Yamamoto, K. (2012) Functional variants in *NFKBIE* and *RTKN2* involved in activation of the NF- κ B pathway are associated with rheumatoid arthritis in Japanese. *PLoS Genetics* **8**, e1002949
- Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E. *et al.* (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF- κ B pathways. *Nature Genetics* **41**, 199–204
- Nakamura, M., Nishida, N., Kawashima, M., Aiba, Y., Tanaka, A., Yasunami, M., Nakamura, H. *et al.* (2012) Genome-wide association study identifies *TNFSF15* and *POU2AF1* as susceptibility loci for primary biliary cirrhosis in the Japanese population. *American Journal of Human Genetics* **91**, 721–728
- Nakayama, A., Nakaoka, H., Yamamoto, K., Sakiyama, M., Shaukat, A., Toyoda, Y., Okada, Y. *et al.* (2017) GWAS of clinically defined gout and subtypes identifies multiple susceptibility loci that include urate transporter genes. *Annals of the Rheumatic Diseases* **76**, 869–877
- Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., Sveinbjörnsdóttir, S., Stefánsson, K., Martinez, M., Hardy, J., Heutink, P., Brice, A., Gasser, T., Singleton, A.B. and Wood, N.W. (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet* **377**, 641–9
- Nalls, M.A., Pankratz, N., Lill, C.M., Do, C.B., Hernandez, D.G., Saad, M., DeStefano, A.L. *et al.* (2014) Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics* **46**, 989–993
- Namjou, B., Keddache, M., Marsolo, K., Wagner, M., Lingren, T., Cobb, B., Perry, C., Kennebeck, S., Holm, I.A., Li, R., Crimmins, N.A., Martin, L., Solti, I., Kohane, I.S. and Harley, J.B. (2013)

EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. *Frontiers in Genetics* **4**, 268

Nanayakkara, S., Senevirathna, S.T.M.L.D., Abeysekera, T., Chandrajith, R., Ratnatunga, N., Gunarathne, E.D.L., Yan, J. et al. (2014) An integrative study of the genetic, social and environmental determinants of chronic kidney disease characterized by tubulointerstitial damages in the North Central Region of Sri Lanka. *Journal of Occupational Health* **56**, 28–38

Nath, S.D., Voruganti, V.S., Arar, N.H., Thameem, F., Lopez-Alvarenga, J.C., Bauer, R., Blangero, J., MacCluer, J.W., Comuzzie, A.G. and Abboud, H.E. (2007) Genome scan for determinants of serum uric acid variability. *Journal of the American Society of Nephrology* **18**, 3156–3163

Neel, J.V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American Journal of Human Genetics* **14**, 353–62

Negi, S., Juyal, G., Senapati, S., Prasad, P., Gupta, A., Singh, S., Kashyap, S., Kumar, A., Kumar, U., Gupta, R., Kaur, S., Agrawal, S., Aggarwal, A., Ott, J., Jain, S., Juyal, R.C. and Thelma, B.K. (2013) A genome-wide association study reveals *ARL15*, a novel non-HLA susceptibility gene for rheumatoid arthritis in North Indians. *Arthritis and Rheumatism* **65**, 3026–3035

Nelson, P.T., Estus, S., Abner, E.L., Parikh, I., Malik, M., Neltner, J.H., Ighodaro, E. et al. (2014) *ABCC9* gene polymorphism is associated with hippocampal sclerosis of aging pathology. *Acta Neuropathologica* **127**, 825–843

New Zealand Ministry of Health (2012) *Mortality and Demographic Data 2012*. New Zealand Ministry of Health.

Available: <http://www.health.govt.nz/publication/mortality-and-demographic-data-2012> [2015-11-18]

New Zealand Ministry of Health (2013) New Zealand Health Survey: annual update of key findings 2012/2013.

Available: <http://www.health.govt.nz/system/files/documents/publications/new-zealand-health-survey-annual-update-2012-13-dec13-v2.pdf> [2015-04-09]

New Zealand Ministry of Health (2016) *Annual Update of Key Results 2015/16: New Zealand Health Survey*. New Zealand Ministry of Health

Ng, M.C.Y., Shriner, D., Chen, B.H., Li, J., Chen, W.M., Guo, X., Liu, J. et al. (2014) Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genetics* **10**, e1004517

Nielsen, R. (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197–218

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566–1575

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A.G. (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics* **8**, 857–868

- Nielsen, R., Hubisz, M.J., Hellmann, I., Torgerson, D., Andres, A.M., Albrechtsen, A., Gutenkunst, R., Adams, M.D., Cargill, M., Boyko, A., Indap, A., Bustamante, C.D. and Clark, A.G. (2009) Darwinian and demographic forces affecting human protein coding genes. *Genome Research* **19**, 838–849
- Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E. (2017) Tracing the peopling of the world through genomics. *Nature* **541**, 302–310
- Nischwitz, S., Cepok, S., Kröner, A., Wolf, C., Knop, M., Müller-Sarnowski, F., Pfister, H., Roeske, D., Rieckmann, P., Hemmer, B., Ising, M., Uhr, M., Bettecken, T., Holsboer, F., Müller-Myhsok, B. and Weber, F. (2010) Evidence for *VAV2* and *ZNF433* as susceptibility genes for multiple sclerosis. *Journal of Neuroimmunology* **227**, 162–166
- Okada, Y., Shimane, K., Kochi, Y., Tahira, T., Suzuki, A., Higasa, K., Takahashi, A. et al. (2012a) A genome-wide association study identified *AFF1* as a susceptibility locus for systemic lupus erythematosus in Japanese. *PLoS Genetics* **8**, e1002455
- Okada, Y., Sim, X., Go, M.J., Wu, J.Y., Gu, D., Takeuchi, F., Takahashi, A. et al. (2012b) Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nature Genetics* **44**, 904–909
- Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Kawaguchi, T. et al. (2012c) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nature Genetics* **44**, 511–516
- Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y. et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381
- Olijhoek, J.K., van der Graaf, Y., Haffner, S.M. and Visseren, F.L.J. (2007) Defining the metabolic syndrome: Resolving unresolved issues? *European Journal of Internal Medicine* **18**, 309–313
- Opitz, B., Eitel, J., Meixenberger, K. and Suttorp, N. (2009) Role of Toll-like receptors, NOD-like receptors and RIG-I-like receptors in endothelial cells and systemic infections. *Thrombosis and Haemostasis* **102**, 1103–1109
- Orengo, J.M., Leliwa-Sytek, A., Evans, J.E., Evans, B., Hoef, D.v.d., Nyako, M., Day, K. and Rodriguez, A. (2009) Uric acid is a mediator of the *Plasmodium falciparum*-induced inflammatory response. *PLoS One* **4**, e5194
- Orozco, G., Viatte, S., Bowes, J., Martin, P., Wilson, A.G., Morgan, A.W., Steer, S., Wordsworth, P., Hocking, L.J., Barton, A., Worthington, J. and Eyre, S. (2014) Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended uk genome-wide association study. *Arthritis & Rheumatology* **66**, 24–30
- Osborn, O. and Olefsky, J.M. (2012) The cellular and signaling networks linking the immune system and metabolism in disease. *Nature Medicine* **18**, 363–374

- Östensson, M., Montén, C., Bacelis, J., Gudjonsdottir, A.H., Adamovic, S., Ek, J., Ascher, H., Pollak, E., Arnell, H., Browaldh, L., Agardh, D., Wahlström, J., Nilsson, S. and Torinsson-Naluai, Å. (2013) A possible mechanism behind autoimmune disorders discovered by genome-wide linkage and association analysis in celiac disease. *PloS One* **8**, e70174
- Ostrowski, J., Paziewska, A., Lazowska, I., Ambroziewicz, F., Goryca, K., Kulecka, M., Rawa, T. et al. (2016) Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. *Scientific Reports* **6**, 39831
- Padyukov, L., Seielstad, M., Ong, R.T.H., Ding, B., Rönnelid, J., Seddighzadeh, M., Alfredsson, L. and Klareskog, L. (2011) A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Annals of the Rheumatic Diseases* **70**, 259–265
- Pagani, F., Raponi, M. and Baralle, F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6368–6372
- Pagani, L., Lawson, J., Jagoda, E., Mörseburg, A., Clemente, F., Hudjashov, G., DeGiorgio, M. et al. (2016) Geographical barriers, environmental challenges, and complex migration events during the peopling of Eurasia. *Nature* pp. 238–242
- Palmer, N.D., McDonough, C.W., Hicks, P.J., Roh, B.H., Wing, M.R., An, S.S., Hester, J.M. et al. (2012) A genome-wide association search for type 2 diabetes genes in African Americans. *PloS One* **7**, e29202
- Pankratz, N., Beecham, G.W., DeStefano, A.L., Dawson, T.M., Doheny, K.F., Factor, S.A., Hamza, T.H. et al. (2012) Meta-analysis of Parkinson's disease: identification of a novel locus, *RIT2*. *Annals of Neurology* **71**, 370–384
- Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., Roberts, R.G. et al. (2007) Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics* **39**, 830–832
- Parmar, M.S. (2009) Uric acid and cardiovascular risk. *The New England Journal of Medicine* **360**, 539; author reply 540–1
- Parra, E.J., Below, J.E., Krithika, S., Valladares, A., Barta, J.L., Cox, N.J., Hanis, C.L., Wacher, N., Garcia-Mena, J., Hu, P., Shriver, M.D., Kumate, J., McKeigue, P.M., Escobedo, J. and Cruz, M. (2011) Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia* **54**, 2038–2046
- Paternoster, L., Evans, D.M., Nohr, E.A., Holst, C., Gaborieau, V., Brennan, P., Gjesing, A.P. et al. (2011) Genome-wide population-based association study of extremely overweight young adults—the GOYA study. *PloS one* **6**, e24303

- Patsopoulos, N.A., Esposito, F., Reischl, J., Lehr, S., Bauer, D., Heubach, J., Sandbrink, R. *et al.* (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Annals of Neurology* **70**, 897–912
- Pattaro, C., Köttgen, A., Teumer, A., Garnaas, M., Böger, C.A., Fuchsberger, C., Olden, M. *et al.* (2012) Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genetics* **8**, e1002584
- Pattaro, C., Teumer, A., Gorski, M., Chu, A.Y., Li, M., Mijatovic, V., Garnaas, M. *et al.* (2016) Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nature Communications* **7**, 10023
- Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, e190
- Pei, Y.F., Zhang, L., Liu, Y., Li, J., Shen, H., Liu, Y.Z., Tian, Q., He, H., Wu, S., Ran, S., Han, Y., Hai, R., Lin, Y., Zhu, J., Zhu, X.Z., Papasian, C.J. and Deng, H.W. (2014) Meta-analysis of genome-wide association data identifies novel susceptibility loci for obesity. *Human Molecular Genetics* **23**, 820–830
- Peng, B. and Amos, C.I. (2010) Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics* **11**, 442
- Pérez-Palma, E., Bustos, B.I., Villamán, C.F., Alarcón, M.A., Avila, M.E., Ugarte, G.D., Reyes, A.E., Opazo, C. and De Ferrari, G.V. (2014) Overrepresentation of glutamate signaling in Alzheimer's disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. *PloS One* **9**, e95413
- Perry, J.R.B., Voight, B.F., Yengo, L., Amin, N., Dupuis, J., Ganser, M., Grallert, H. *et al.* (2012) Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in *LAMA1* and enrichment for risk variants in lean compared to obese cases. *PLoS Genetics* **8**, e1002741
- Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. and Lercher, M.J. (2014) PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*
- Phipps-Green, A.J., Hollis-Moffatt, J.E., Dalbeth, N., Merriman, M.E., Topless, R., Gow, P.J., Harrison, A.A., Highton, J., Jones, P.B.B., Stamp, L.K. and Merriman, T.R. (2010) A strong role for the *ABCG2* gene in susceptibility to gout in New Zealand Pacific Island and Caucasian, but not Māori, case and control sample sets. *Human Molecular Genetics* **19**, 4813–4819
- Phipps-Green, A.J., Merriman, M.E., Topless, R., Altaf, S., Montgomery, G.W., Franklin, C., Jones, G.T., van Rij, A.M., White, D., Stamp, L.K., Dalbeth, N. and Merriman, T.R. (2016) Twenty-eight loci that influence serum urate levels: analysis of association with gout. *Annals of the Rheumatic Diseases* **75**, 124–30
- Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. and Others (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826–837

- Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y. and Hinds, D.A. (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* **48**, 709–717
- Plenge, R.M., Cotsapas, C., Davies, L., Price, A.L., de Bakker, P.I.W., Maller, J., Pe'er, I. et al. (2007a) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nature Genetics* **39**, 1477–1482
- Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A. et al. (2007b) *TRAF1-C5* as a risk locus for rheumatoid arthritis—a genomewide study. *The New England Journal of Medicine* **357**, 1199–1209
- Popejoy, A.B. and Fullerton, S.M. (2016) Genomics is failing on diversity. *Nature* **538**, 161–164
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909
- Price, A.L., Butler, J., Patterson, N., Capelli, C., Pascali, V.L., Scarnicci, F., Ruiz-Linares, A. et al. (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS Genetics* **4**, 9–17
- Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59
- Przeworski, M. (2002) The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189
- Purcell, S. and Chang, C. (2015) PLINK 1.9.
Available: www.cog-genomics.org/plink/1.9/
- Pybus, M., Dall'olio, G.M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetti, J. and Engelken, J. (2014) 1000 Genomes Selection Brower 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research* **42**, D903–909
- Pybus, M., Luisi, P., Dall'Olio, G.M., Uzkudun, M., Laayouni, H., Bertranpetti, J. and Engelken, J. (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* **31**, 3946–3952
- Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G. and Simianer, H. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* **12**, 318
- Qanbari, S., Strom, T.M., Haberer, G., Weigend, S., Gheyas, A.a., Turner, F., Burt, D.W., Preisinger, R., Gianola, D. and Simianer, H. (2012) A high resolution genome-wide scan for significant selective sweeps: An application to pooled sequence data in laying chickens. *PLoS One* **7**, 1–12
- Qi, L., Cornelis, M.C., Kraft, P., Stanya, K.J., Kao, W.H.L., Pankow, J.S., Dupuis, J. et al. (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human Molecular Genetics* **19**, 2706–2715

Quan, C., Ren, Y.Q., Xiang, L.H., Sun, L.D., Xu, A.E., Gao, X.H., Chen, H.D. *et al.* (2010) Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the *MHC*. *Nature Genetics* **42**, 614–618

Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K. and Santachiara-Benerecetti, a.S. (1999) Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nature Genetics* **23**, 437–441

R Core Team (2017) R: A Language and Environment for Statistical Computing

Raelson, J.V., Little, R.D., Ruether, A., Fournier, H., Paquin, B., Van Eerdewegh, P., Bradley, W.E.C. *et al.* (2007) Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14747–14752

Ramírez-Soriano, A. and Nielsen, R. (2009) Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* **181**, 701–710

Ramírez-Soriano, A., Ramos-Onsins, S.E., Rozas, J., Calafell, F. and Navarro, A. (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179**, 555–567

Ramos, P.S. (2017) Population genetics and natural selection in rheumatic disease. *Rheumatic Disease Clinics of North America* **43**, 313–326

Rana, J.S., Nieuwdorp, M., Jukema, J.W. and Kastelein, J.J.P. (2007) Cardiovascular metabolic syndrome - an interplay of, obesity, inflammation, diabetes and coronary heart disease. *Diabetes, Obesity & Metabolism* **9**, 218–32

Randhawa, I.A.S., Khatkar, M.S., Thomson, P.C. and Raadsma, H.W. (2014) Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics* **15**, 34

Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L. *et al.* (2008) Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nature Genetics* **40**, 1216–1223

Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genetics* **39**, 596–604

Roeder, K., Bacanu, S.A., Wasserman, L. and Devlin, B. (2006) Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics* **78**, 243–252

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. *Science* **298**, 2381–2385

- Rowe, P.S., De Zoysa, P.A., Dong, R., Wang, H.R., White, K.E., Econis, M.J. and Oudet, C.L. (2000) *MEPE*, a new gene expressed in bone marrow and tumors causing osteomalacia. *Genomics* **67**, 54–68
- Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proen  a, C., Bacot, F. et al. (2009) Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genetics* **41**, 1110–1115
- Saad, M., Lesage, S., Saint-Pierre, A., Corvol, J.C., Zelenika, D., Lambert, J.C., Vidailhet, M. et al. (2011) Genome-wide association study confirms *BST1* and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. *Human Molecular Genetics* **20**, 615–627
- Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, a., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive natural selection in the human lineage. *Science* **312**, 1614–1620
- Sankar, P., Cho, M.K., Wolpe, P.R. and Schairer, C. (2006) What is in a cause? Exploring the relationship between genetic cause and felt stigma. *Genetics in Medicine* **8**, 33–42
- Sanna, S., Pitzalis, M., Zoledziewska, M., Zara, I., Sidore, C., Murru, R., Whalen, M.B. et al. (2010) Variants within the immunoregulatory *CBLB* gene are associated with multiple sclerosis. *Nature Genetics* **42**, 495–497
- Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T. et al. (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nature Genetics* **41**, 1303–1307
- Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A., Moutsianas, L., Dilthey, A. et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219
- Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I.W., Chen, H., Roix, J.J. et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336
- Saxena, R., Saleheen, D., Been, L.F., Garavito, M.L., Braun, T., Bjonne, A., Young, R. et al. (2013) Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes* **62**, 1746–1755
- Saxena, R., Plenge, R.M., Bjonne, A.C., Dashti, H.S., Okada, Y., Haq, W.G.E., Hammoudeh, M. et al. (2017) A Multinational Arab Genome-Wide Association Study Identifies New Genetic Associations for Rheumatoid Arthritis. *Arthritis & Rheumatology* **69**, 976–985
- Scherag, A., Dina, C., Hinney, A., Vatin, V., Scherag, S., Vogel, C.I.G., M  ller, T.D. et al. (2010) Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS Genetics* **6**, e1000916
- Schrider, D.R. and Kern, A.D. (2018) Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics* **34**, 301–312

- Schrider, D.R., Mendes, F.K., Hahn, M.W. and Kern, a.D. (2015) Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**, 267–284
- Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345
- Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S. *et al.* (2007) Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genetics* **3**, e115
- Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C. *et al.* (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. *Journal of the American Medical Association* **303**, 1832–1840
- Sheehan, S. and Song, Y.S. (2016) Deep learning for population genetic inference. *PLoS Computational Biology* **12**, 1–28
- Shen, R., Fan, J.B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K. and Olliphant, A. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **573**, 70–82
- Shu, X.O., Long, J., Cai, Q., Qi, L., Xiang, Y.B., Cho, Y.S., Tai, E.S. *et al.* (2010) Identification of new genetic risk variants for type 2 diabetes. *PLoS Genetics* **6**, e1001127
- Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Mägi, R., Strawbridge, R.J. *et al.* (2015) New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196
- Silverberg, M.S., Cho, J.H., Rioux, J.D., McGovern, D.P.B., Wu, J., Annese, V., Achkar, J.P. *et al.* (2009) Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nature Genetics* **41**, 216–220
- Sim, X., Ong, R.T.H., Suo, C., Tay, W.T., Liu, J., Ng, D.P.K., Boehnke, M., Chia, K.S., Wong, T.Y., Seielstad, M., Teo, Y.Y. and Tai, E.S. (2011) Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genetics* **7**, e1001363
- Simmons, R.T. (1962) Blood group genes in Polynesians and comparisons with other Pacific peoples. *Oceania* **32**, 198–210
- Simón-Sánchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genetics* **41**, 1308–1312
- Simonsen, K.L., Churchill, G.a. and Aquadro, C.F. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429

- Singh, M., Mukherjee, P., Narayanasamy, K., Arora, R., Sen, S.D., Gupta, S., Natarajan, K. and Malhotra, P. (2009) Proteome analysis of *Plasmodium falciparum* extracellular secretory antigens at asexual blood stages reveals a cohort of proteins with possible roles in immune modulation and signaling. *Molecular & Cellular Proteomics* **8**, 2102–2118
- Siva, N. (2008) 1000 genomes project. *Nature biotechnology* **26**, 256
- Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G.R. *et al.* (2016) Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513
- Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885
- Snaith, M.L. (2004) Gout pp. 1–9
- Soares, P., Alshamali, F., Pereira, J.B., Fernandes, V., Silva, N.M., Afonso, C., Costa, M.D., Musilová, E., MacAulay, V., Richards, M.B., Černý, V. and Pereira, L. (2012) The expansion of mtDNA haplogroup L3 within and out of Africa. *Molecular Biology and Evolution* **29**, 915–927
- Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L. *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937–948
- Spencer, C.C.A., Su, Z., Donnelly, P. and Marchini, J. (2009) Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**
- Spencer, C.C.A., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., Barker, R.A. *et al.* (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Human Molecular Genetics* **20**, 345–353
- Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., Murabito, J.M., O'Donnell, C.J., Vasan, R.S., Wolf, P.a. and Levy, D. (2007) The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American Journal of Epidemiology* **165**, 1328–35
- Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y. *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42**, 508–514
- Stajich, J.E. and Hahn, M.W. (2005) Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution* **22**, 63–73
- Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U. *et al.* (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genetics* **39**, 770–775

- Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. and Williams Jr, R.M. (1949) The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1
- Strange, A., Capon, F., Spencer, C.C.A., Knight, J., Weale, M.E., Allen, M.H., Barton, A. *et al.* (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*. *Nature Genetics* **42**, 985–990
- Stuart, P.E., Nair, R.P., Ellinghaus, E., Ding, J., Tejasvi, T., Gudjonsson, J.E., Li, Y. *et al.* (2010) Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nature Genetics* **42**, 1000–1004
- Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadottir, H.T., Helgason, A., Gudjonsson, S.A., Zanon, C. *et al.* (2011) Identification of low-frequency variants associated with gout and serum uric acid levels. *Nature Genetics* **43**, 1127–1130
- Szpiech, Z.A. and Hernandez, R.D. (2014) selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution* **31**, 2824–2827
- Tabassum, R., Chauhan, G., Dwivedi, O.P., Mahajan, A., Jaiswal, A., Kaur, I., Bandesh, K. *et al.* (2013) Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* **62**, 977–986
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595
- Tajima, F. (1996) Infinite allele model and infinite site model in population genetics. *Journal of Genetics* **75**, 27–31
- Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., Ikegami, H., Sugiyama, T., Katsuya, T., Miyagishi, M., Nakashima, N., Nawata, H., Nakamura, J., Kono, S., Takayanagi, R. and Kato, N. (2009) Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes* **58**, 1690–1699
- Tang, K., Thornton, K.R. and Stoneking, M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology* **5**, e171
- Tang, X.F., Zhang, Z., Hu, D.Y., Xu, A.E., Zhou, H.S., Sun, L.D., Gao, M. *et al.* (2013) Association analyses identify three susceptibility Loci for vitiligo in the Chinese Han population. *The Journal of Investigative Dermatology* **133**, 403–410
- Tanner, C., Boocock, J., Stahl, E.A., Dobbyn, A., Mandal, A.K., Cadzow, M., Phipps-Green, A.J., Topless, R.K., Hindmarsh, J.H., Stamp, L.K., Dalbeth, N., Choi, H.K., Mount, D.B. and Merriman, T.R. (2017) Population-Specific Resequencing Associates the ATP-Binding Cassette Subfamily C Member 4 Gene With Gout in New Zealand Māori and Pacific Men. *Arthritis and Rheumatology* **69**, 1461–1469
- Tennessen, J.a. and Akey, J.M. (2011) Parallel adaptive divergence among geographically diverse human populations. *PLoS Genetics* **7**, e1002127

- Terao, C., Yamada, R., Ohmura, K., Takahashi, M., Kawaguchi, T., Kochi, Y., Okada, Y., Nakamura, Y., Yamamoto, K., Melchers, I., Lathrop, M., Mimori, T. and Matsuda, F. (2011) The human *AIRE* gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Human Molecular Genetics* **20**, 2680–2685
- Teshima, K.M., Coop, G. and Przeworski, M. (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**, 702–712
- The International Hapmap Consortium (2005) A haplotype map of the human genome. *Nature* **437**, 1299–320
- Thomas, P.D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15398–15403
- Thorleifsson, G., Walters, G.B., Gudbjartsson, D.F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., Styrkarsdottir, U. et al. (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics* **41**, 18–24
- Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J. et al. (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* **489**, 443–446
- Timpson, N.J., Lindgren, C.M., Weedon, M.N., Randall, J., Ouwehand, W.H., Strachan, D.P., Rayner, N.W., Walker, M., Hitman, G.A., Doney, A.S.F., Palmer, C.N.A., Morris, A.D., Hattersley, A.T., Zeggini, E., Frayling, T.M. and McCarthy, M.I. (2009) Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes* **58**, 505–510
- Tin, A., Woodward, O.M., Kao, W.H.L., Liu, C.T., Lu, X., Nalls, M.A., Shriner, D. et al. (2011) Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel *URAT1* loss-of-function allele. *Human Molecular Genetics* **20**, 4056–4068
- Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R. et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics* **39**, 857–864
- Tsai, F.J., Yang, C.F., Chen, C.C., Chuang, L.M., Lu, C.H., Chang, C.T., Wang, T.Y., Chen, R.H., Shiu, C.F., Liu, Y.M., Chang, C.C., Chen, P., Chen, C.H., Fann, C.S.J., Chen, Y.T. and Wu, J.Y. (2010) A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genetics* **6**, e1000847
- Tsoi, L.C., Spain, S.L., Ellinghaus, E., Stuart, P.E., Capon, F., Knight, J., Tejasvi, T. et al. (2015) Enhanced meta-analysis and replication studies identify five new psoriasis susceptibility loci. *Nature Communications* **6**, 7001

- Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P.K. *et al.* (2008) SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature Genetics* **40**, 1098–1102
- Utsunomiya, Y.T., Pérez O'Brien, A.M., Sonstegard, T.S., Van Tassell, C.P., do Carmo, A.S., Mészáros, G., Sölkner, J. and Garcia, J.F. (2013) Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One* **8**, 1–11
- Utsunomiya, Y.T., Perez O'Brien, A.M., Sonstegard, T.S., Sölkner, J. and Garcia, J.F. (2015) Genomic data as the "hitchhiker's guide" to cattle adaptation: tracking the milestones of past selection in the bovine genome. *Frontiers in Genetics* **6**, 1–13
- Vacic, V., Ozelius, L.J., Clark, L.N., Bar-Shira, A., Gana-Weisz, M., Gurevich, T., Gusev, A. *et al.* (2014) Genome-wide mapping of IBD segments in an Ashkenazi PD cohort identifies associated haplotypes. *Human Molecular Genetics* **23**, 4693–4702
- van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C. *et al.* (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*. *Nature Genetics* **39**, 827–829
- Visscher, P.M., Hill, W.G. and Wray, N.R. (2008) Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–266
- Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) Five years of GWAS discovery. *American Journal of Human Genetics* **90**, 7–24
- Vitart, V., Rudan, I., Hayward, C., Gray, N.K., Floyd, J., Palmer, C.N.A., Knott, S.A. *et al.* (2008) *SLC2A9* is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nature Genetics* **40**, 437–442
- Vitti, J.J., Grossman, S.R. and Sabeti, P.C. (2013) Detecting natural selection in genomic data. *Annual Review of Genetics* **47**, 97–120
- Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biology* **4**, e72
- Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**, 579–589
- Wakeley, J. and Aliacar, N. (2001) Gene genealogies in a metapopulation. *Genetics* **159**, 893–905
- Waldner, H. (2009) The role of innate immune responses in autoimmune disease development. *Autoimmunity Reviews* **8**, 400–404
- Wall, J., Cox, M., Mendez, F., Woerner, A., Severson, T. and Hammer, M. (2008) A novel DNA sequence database for analyzing human demographic history. *Genome Research* **18**, 1354

- Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K. *et al.* (2008) Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American Journal of Human Genetics* **82**, 139–149
- Wallace, C., Smyth, D.J., Maisuria-Armer, M., Walker, N.M., Todd, J.A. and Clayton, D.G. (2010) The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nature Genetics* **42**, 68–71
- Wallace, S.L., Robinson, H., Masi, A.T., Decker, J.L., McCarty, D.J. and Yü, T.F.F. (1977) Preliminary criteria for the classification of the acute arthritis of primary gout. *Arthritis and Rheumatism* **20**, 895–900
- Wang, D.G., Fan, J.b., Siao, C.j., Berno, A., Young, P., Sapolsky, R., Ghandour, G. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome **280**, 1077–1082
- Wang, J.H., Pappas, D., De Jager, P.L., Pelletier, D., de Bakker, P.I., Kappos, L., Polman, C.H., Chibnik, L.B., Hafler, D.A., Matthews, P.M., Hauser, S.L., Baranzini, S.E. and Oksenberg, J.R. (2011a) Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data. *Genome Medicine* **3**, 3
- Wang, K., Li, W.D., Zhang, C.K., Wang, Z., Glessner, J.T., Grant, S.F.A., Zhao, H., Hakonarson, H. and Price, R.A. (2011b) A genome-wide association study on obesity and obesity-related traits. *PloS One* **6**, e18939
- Wang, M., Huang, X., Li, R., Xu, H., Jin, L. and He, Y. (2014) Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. *Molecular Biology and Evolution* **31**, 3068–3080
- Wang, Y.C., McPherson, K., Marsh, T., Gortmaker, S.L. and Brown, M. (2011c) Health and economic burden of the projected obesity trends in the USA and the UK. *The Lancet* **378**, 815–825
- Warren, H.R., Evangelou, E., Cabrera, C.P., Gao, H., Ren, M., Mifsud, B., Ntalla, I. *et al.* (2017) Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics* **49**, 403–415
- Warrington, N.M., Howe, L.D., Paternoster, L., Kaakinen, M., Herrala, S., Huikari, V., Wu, Y.Y., Kemp, J.P., Timpson, N.J., Pourcain, B.S., Smith, G.D., Tilling, K., Jarvelin, M.R., Pennell, C.E., Evans, D.M., Lawlor, D.A., Briollais, L. and Palmer, L.J. (2015) A genome-wide association study of body mass index across early life and childhood. *International Journal of Epidemiology* **44**, 700–712
- Watanabe, S., Kang, D.H., Feng, L., Nakagawa, T., Kanellis, J., Lan, H., Mazzali, M. and Johnson, R.J. (2002) Uric acid, hominoid evolution, and the pathogenesis of salt-sensitivity. *Hypertension* **40**, 355–360
- Watterson, G.a. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276

- Webster, J.A., Myers, A.J., Pearson, J.V., Craig, D.W., Hu-Lince, D., Coon, K.D., Zismann, V.L. *et al.* (2008) *SORL1* as an Alzheimer's disease predisposition gene? *Neuro-degenerative Diseases* **5**, 60–64
- Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., Liu, Z., Liu, R., Zhao, F., Zhang, L., Lu, J., Liu, C. and Du, L. (2015) Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genomics* **16**, 194
- Weir, B.S. and Cockerham, C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370
- Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. and Hill, W.G. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468–76
- Weissglas-Volkov, D., Aguilar-Salinas, C.A., Nikkola, E., Deere, K.A., Cruz-Bautista, I., Arellano-Campos, O., Muñoz-Hernandez, L.L., Gomez-Munguia, L., Ordoñez-Sánchez, M.L., Linga Reddy, P.M., Lusis, A.J., Matikainen, N., Taskinen, M.R., Riba, L., Cantor, R.M., Sinsheimer, J.S., Tusie-Luna, T. and Pajukanta, P. (2013) Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. *Journal of Medical Genetics* **50**, 298–308
- Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678
- Wen, W., Cho, Y.S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., Chen, C.H. *et al.* (2012) Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature Genetics* **44**, 307–311
- Wen, W., Zheng, W., Okada, Y., Takeuchi, F., Tabara, Y., Hwang, J.Y., Dorajoo, R. *et al.* (2014) Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Human Molecular Genetics* **23**, 5492–5504
- Wen, W., Kato, N., Hwang, J.Y., Guo, X., Tabara, Y., Li, H., Dorajoo, R. *et al.* (2016) Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Scientific Reports* **6**, 17958
- Wen, Y.J., Zhang, H., Ni, Y.L., Huang, B., Zhang, J., Feng, J.Y., Wang, S.B., Dunwell, J.M., Zhang, Y.M. and Wu, R. (2017) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings in Bioinformatics* **19**, 1–13
- Wheeler, E., Huang, N., Bochukova, E.G., Keogh, J.M., Lindsay, S., Garg, S., Henning, E., Blackburn, H., Loos, R.J.F., Wareham, N.J., O'Rahilly, S., Hurles, M.E., Barroso, I. and Farooqi, I.S. (2013) Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature Genetics* **45**, 513–517
- Wigginton, J.E., Cutler, D.J. and Abecasis, G.R. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893

- Willer, C.J., Speliotes, E.K., Loos, R.J.F., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I. *et al.* (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics* **41**, 25–34
- Williams, A.L., Jacobs, S.B.R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., Gómez-Vázquez, M.J., Burtt, N.P., Aguilar-Salinas, C.A., González-Villalpando, C., Florez, J.C., Orozco, L., Haiman, C.A., Tusié-Luna, T. and Altshuler, D. (2014) Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101
- Wilmshurst, J.M., Hunt, T.L., Lipo, C.P. and Anderson, A.J. (2011) High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1815–1820
- Winnard, D., Wright, C., Taylor, W.J., Jackson, G., Te Karu, L., Gow, P.J., Arroll, B., Thornley, S., Gribben, B. and Dalbeth, N. (2012) National prevalence of gout derived from administrative health data in Aotearoa New Zealand. *Rheumatology (Oxford, England)* **51**, 901–909
- Winnard, D., Wright, C., Jackson, G., Gow, P., Kerr, A., McLachlan, A., Orr-Walker, B. and Dalbeth, N. (2013) Gout, diabetes and cardiovascular disease in the Aotearoa New Zealand adult population: co-prevalence and implications for clinical practice. *New Zealand Medical Journal* **126**, 53–64
- Wiuf, C. and Donnelly, P. (1999) Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology* **56**, 183–201
- Wollstein, A., Lao, O., Becker, C., Brauer, S., Trent, R.J., Nürnberg, P., Stoneking, M. and Kayser, M. (2010) Demographic history of Oceania inferred from genome-wide data. *Current Biology* **20**, 1983–1992
- Woodward, O.M., Köttgen, A., Coresh, J., Boerwinkle, E., Guggino, W.B. and Köttgen, M. (2009) Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 10338–42
- Wright, S. (1931) Evolution in Mendelian Populations. *Genetics* **16**, 97–159
- Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* **15**, 323–354
- Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M. *et al.* (2010) A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at *UBE2E2* and *C2CD4A-C2CD4B*. *Nature Genetics* **42**, 864–868
- Yamazaki, K., Umeno, J., Takahashi, A., Hirano, A., Johnson, T.A., Kumashiro, N., Morizono, T. *et al.* (2013) A genome-wide association study identifies 2 susceptibility loci for Crohn's disease in a Japanese population. *Gastroenterology* **144**, 781–788
- Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011a) GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76–82

- Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M. *et al.* (2011b) Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519–525
- Yang, J., Yang, W., Hirankarn, N., Ye, D.Q., Zhang, Y., Pan, H.F., Mok, C.C. *et al.* (2011c) *ELF1* is associated with systemic lupus erythematosus in Asian populations. *Human Molecular Genetics* **20**, 601–607
- Yang, J., Loos, R.J.F., Powell, J.E., Medland, S.E., Speliotes, E.K., Chasman, D.I., Rose, L.M. *et al.* (2012) *FTO* genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272
- Yang, J., Zeng, J., Goddard, M.E., Wray, N.R. and Visscher, P.M. (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* **49**, 1304–1311
- Yang, Q., Guo, C.Y., Cupples, L.A., Levy, D., Wilson, P.W.F. and Fox, C.S. (2005) Genome-wide search for genes affecting serum uric acid levels: the Framingham Heart Study. *Metabolism: clinical and experimental* **54**, 1435–1441
- Yang, Q., Köttgen, A., Dehghan, A., Smith, A.V., Glazer, N.L., Chen, M.H., Chasman, D.I. *et al.* (2010b) Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circulation: Cardiovascular Genetics* **3**, 523–530
- Yang, Q., Köttgen, A., Dehghan, A., Smith, A.V., Glazer, N.L., Chen, M.H., Chasman, D.I. *et al.* (2010a) Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. *Circulation: Cardiovascular Genetics* **3**, 523–530
- Yang, S.K., Hong, M., Zhao, W., Jung, Y., Tayebi, N., Ye, B.D., Kim, K.J., Park, S.H., Lee, I., Shin, H.D., Cheong, H.S., Kim, L.H., Kim, H.J., Jung, S.A., Kang, D., Youn, H.S., Liu, J. and Song, K. (2013a) Genome-wide association study of ulcerative colitis in Koreans suggests extensive overlapping of genetic susceptibility with Caucasians. *Inflammatory Bowel Diseases* **19**, 954–966
- Yang, S.K., Hong, M., Zhao, W., Jung, Y., Baek, J., Tayebi, N., Kim, K.M. *et al.* (2014) Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut* **63**, 80–87
- Yang, S.K., Hong, M., Oh, H., Low, H.Q., Jung, S., Ahn, S., Kim, Y. *et al.* (2016) Identification of loci at 1q21 and 16q23 that affect susceptibility to inflammatory bowel disease in Koreans. *Gastroenterology* **151**, 1096–1099
- Yang, W., Shen, N., Ye, D.Q., Liu, Q., Zhang, Y., Qian, X.X., Hirankarn, N. *et al.* (2010c) Genome-wide association study in Asian populations identifies variants in *ETS1* and *WDFY4* associated with systemic lupus erythematosus. *PLoS Genetics* **6**, e1000841
- Yang, W., Tang, H., Zhang, Y., Tang, X., Zhang, J., Sun, L., Yang, J. *et al.* (2013b) Meta-analysis followed by replication identifies loci in or near *CDKN1B*, *TET3*, *CD80*, *DRAM1*, and *ARID5B* as associated with systemic lupus erythematosus in Asians. *American Journal of Human Genetics* **92**, 41–51

- Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y. *et al.* (2008) Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nature Genetics* **40**, 1092–1097
- Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X. *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78
- Yin, X., Low, H.Q., Wang, L., Li, Y., Ellinghaus, E., Han, J., Estivill, X. *et al.* (2015) Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. *Nature Communications* **6**, 6916
- Yuan, X., Miller, D.J., Zhang, J., Herrington, D. and Wang, Y. (2012) An Overview of Population Genetic Data Simulation. *Journal of Computational Biology* **19**, 42–54
- Zabaneh, D. and Balding, D.J. (2010) A genome-wide association study of the metabolic syndrome in Indian Asian men. *PloS One* **5**, e11961
- Zaykin, D.V. (2011) Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology* **24**, 1836–1841
- Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341
- Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I.W. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**, 638–645
- Zeng, K., Fu, Y.X., Shi, S. and Wu, C.I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–1439
- Zerihun, T., Degarege, A. and Erko, B. (2011) Association of ABO blood group and *Plasmodium falciparum* malaria in Dore Bafeno Area, Southern Ethiopia. *Asian Pacific Journal of Tropical Biomedicine* **1**, 289–294
- Zhai, W., Nielsen, R. and Slatkin, M. (2009) An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution* **26**, 273–283
- Zhang, G., Muglia, L.J., Chakraborty, R., Akey, J.M. and Williams, S.M. (2013) Signatures of natural selection on genetic variants affecting complex human traits. *Applied and Translational Genomics* **2**, 77–93
- Zhang, X.J., Huang, W., Yang, S., Sun, L.D., Zhang, F.Y., Zhu, Q.X., Zhang, F.R. *et al.* (2009) Psoriasis genome-wide association study identifies susceptibility variants within *LCE* gene cluster at 1q21. *Nature Genetics* **41**, 205–210
- Zhang, Y., Liu, K., Ma, L., Liu, K., Shi, X., Zhang, Y., He, N., Zhao, Y., Zhu, X., Jin, T. and Others (2016) Associations of gout with polymorphisms in *SLC2A9*, *WDR1*, *CLNK*, *PKD2*, and *ABCG2* in

Chinese Han and Tibetan populations. *International Journal of Clinical and Experimental Pathology* **9**, 7503–7517

Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1193–1198

Appendix A

Supplemental tables

A1 Chapter 3 tables

A1.1 Hider Regions

This table contains the regions from Hider *et al.* (2013) that were used in section 3.3.2.

Table S1: Regions from Hider et al 2013 that were replicated for selection.

Chrom	Start	End	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fu and Li's F Region	Fu and Li's F Present	iHS Present
chr1	12700001	12725000	-	-	-	TRUE	FALSE	-
chr1	27050001	27075000	TRUE	FALSE	-	-	-	-
chr1	27075001	27100000	TRUE	FALSE	-	-	-	-
chr1	27100001	27125000	TRUE	FALSE	-	TRUE	TRUE	-
chr1	27250001	27275000	TRUE	TRUE	-	TRUE	FALSE	-
chr1	27350001	27375000	TRUE	FALSE	-	TRUE	FALSE	-
chr1	33000001	33025000	-	-	-	-	-	-
chr1	37450001	37475000	TRUE	FALSE	-	-	-	-
chr1	46775001	46800000	-	-	-	TRUE	FALSE	-
chr1	46800001	46825000	-	-	TRUE	FALSE	FALSE	-
chr1	49150001	49175000	-	-	-	TRUE	FALSE	-
chr1	51825001	51850000	TRUE	FALSE	-	-	-	-
chr1	64350001	64375000	TRUE	TRUE	-	TRUE	TRUE	-
chr1	69350001	69375000	TRUE	TRUE	-	-	-	-
chr1	75625001	75650000	TRUE	FALSE	-	-	-	-
chr1	75650001	75675000	-	-	TRUE	FALSE	TRUE	-
chr1	80825001	80850000	-	-	TRUE	FALSE	-	-
chr1	92650001	92675000	TRUE	FALSE	-	-	-	-
chr1	102125001	102150000	TRUE	FALSE	-	-	-	-
chr1	119350001	119375000	-	TRUE	FALSE	TRUE	FALSE	-
chr1	150000001	150025000	-	-	-	TRUE	FALSE	-
chr1	150075001	150100000	-	-	-	TRUE	FALSE	-
chr1	154950001	154975000	TRUE	TRUE	-	TRUE	TRUE	-
chr1	157025001	157050000	-	-	TRUE	FALSE	-	-
chr1	162975001	163000000	-	-	-	TRUE	TRUE	-
chr1	171000001	171025000	-	-	-	TRUE	TRUE	-
chr1	175225001	175250000	-	TRUE	FALSE	-	-	-
chr1	177025001	177050000	TRUE	FALSE	-	TRUE	TRUE	-
chr1	185725001	185750000	TRUE	TRUE	-	TRUE	TRUE	-
chr1	185750001	185775000	TRUE	TRUE	-	-	-	-
chr1	185875001	185900000	TRUE	TRUE	-	TRUE	TRUE	-
chr1	188850001	188875000	-	-	TRUE	TRUE	-	-
chr1	195800001	195825000	-	-	TRUE	FALSE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr1	196250001	196275000	TRUE	FALSE	-	-	-	TRUE	FALSE	-	-
chr1	206300001	206325000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr1	238225001	238250000	-	-	-	-	-	TRUE	TRUE	-	-
chr1	245000001	245025000	TRUE	FALSE	-	-	-	TRUE	FALSE	-	-
chr1	247150001	247175000	TRUE	FALSE	-	-	-	TRUE	TRUE	-	-
chr1	247875001	247900000	-	-	-	-	-	TRUE	TRUE	-	-
chr1	247925001	247950000	-	-	-	-	-	TRUE	TRUE	-	-
chr1	247950001	247975000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr1	247975001	248000000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr2	625001	650000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	950001	975000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	1600001	1625000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	5650001	5675000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	9875001	9900000	-	-	-	-	-	-	-	TRUE	FALSE
chr2	15175001	15200000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	17975001	18000000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	21375001	21400000	-	-	TRUE	TRUE	TRUE	FALSE	-	-	-
chr2	21950001	21975000	TRUE	TRUE	-	-	-	-	-	-	-
chr2	28425001	28450000	-	-	-	-	-	TRUE	TRUE	-	-
chr2	28450001	28475000	TRUE	FALSE	-	-	-	TRUE	FALSE	-	-
chr2	65425001	65450000	TRUE	FALSE	-	-	-	TRUE	TRUE	-	-
chr2	72125001	72150000	TRUE	FALSE	-	-	-	-	-	-	-
chr2	72425001	72450000	TRUE	TRUE	-	-	-	-	-	-	-
chr2	72475001	72500000	TRUE	TRUE	-	-	-	-	-	-	-
chr2	72500001	72525000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr2	72525001	72550000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr2	72550001	72575000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr2	72650001	72675000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-
chr2	72675001	72700000	TRUE	TRUE	-	-	-	-	-	-	-
chr2	72725001	72750000	TRUE	TRUE	-	-	-	-	-	-	-
chr2	81100001	81125000	-	-	TRUE	TRUE	TRUE	-	-	TRUE	FALSE
chr2	82500001	82525000	TRUE	TRUE	-	-	-	TRUE	TRUE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr2	82625001	82650000	TRUE	TRUE	-	-	-	-	-	-
chr2	82725001	82750000	TRUE	TRUE	-	-	-	-	-	-
chr2	82775001	82800000	TRUE	TRUE	-	-	-	-	-	-
chr2	84750001	84775000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr2	84825001	84850000	-	-	-	-	TRUE	TRUE	-	-
chr2	84850001	84875000	-	-	-	-	TRUE	TRUE	-	-
chr2	108950001	108975000	TRUE	TRUE	-	-	-	-	-	-
chr2	108975001	109000000	TRUE	TRUE	-	-	-	-	-	-
chr2	109550001	109575000	TRUE	TRUE	-	-	-	-	-	-
chr2	117575001	117600000	-	-	-	-	-	-	TRUE	FALSE
chr2	118975001	119000000	TRUE	FALSE	-	-	-	-	-	-
chr2	121675001	121700000	-	-	-	-	TRUE	FALSE	-	-
chr2	125450001	125475000	-	-	-	-	-	-	TRUE	FALSE
chr2	129400001	129425000	TRUE	FALSE	-	-	-	-	-	-
chr2	138050001	138075000	-	-	TRUE	FALSE	-	-	-	-
chr2	141200001	141225000	-	-	-	-	TRUE	FALSE	-	-
chr2	141225001	141250000	-	-	-	-	TRUE	TRUE	-	-
chr2	149600001	149625000	-	-	-	-	TRUE	TRUE	-	-
chr2	151125001	151150000	-	-	-	-	TRUE	TRUE	-	-
chr2	158325001	158350000	-	-	TRUE	FALSE	-	-	TRUE	TRUE
chr2	168025001	168050000	-	-	TRUE	FALSE	-	-	-	-
chr2	177575001	177600000	TRUE	FALSE	-	-	TRUE	FALSE	-	-
chr2	180900001	180925000	-	-	TRUE	FALSE	-	-	-	-
chr2	197475001	197500000	TRUE	TRUE	-	-	-	-	-	-
chr2	201800001	201825000	TRUE	FALSE	-	-	-	-	-	-
chr2	206275001	206300000	TRUE	FALSE	-	-	-	-	-	-
chr2	206300001	206325000	TRUE	FALSE	-	-	-	-	-	-
chr2	210800001	210825000	-	-	-	-	TRUE	TRUE	-	-
chr2	214175001	214200000	-	-	-	-	TRUE	TRUE	-	-
chr2	214250001	214275000	-	-	-	-	TRUE	TRUE	-	-
chr2	214275001	214300000	-	-	-	-	TRUE	FALSE	-	-
chr2	214300001	214325000	-	-	-	-	TRUE	FALSE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D	Fay and	Fay and	Fu and Li's	Fu and Li's	iHS Region	iHS
			Region	Present	Wu's H	Wu's H	F Region	F Present		Present
chr2	218500001	218525000	TRUE	-	TRUE	-	-	-	-	-
chr2	223925001	223950000	-	-	-	-	TRUE	TRUE	TRUE	FALSE
chr2	232200001	232225000	-	-	-	-	-	-	TRUE	FALSE
chr2	236625001	236650000	-	-	FALSE	-	-	-	TRUE	FALSE
chr2	238025001	238050000	TRUE	-	FALSE	-	-	-	-	-
chr2	238050001	238075000	TRUE	-	FALSE	-	-	-	-	-
chr3	650001	675000	TRUE	-	FALSE	-	-	-	-	-
chr3	17325001	17350000	TRUE	-	TRUE	-	-	-	-	-
chr3	17350001	17375000	TRUE	-	TRUE	-	-	-	-	-
chr3	17375001	17400000	TRUE	-	FALSE	-	-	-	-	-
chr3	17400001	17425000	TRUE	-	TRUE	-	-	-	-	-
chr3	17475001	17500000	TRUE	-	FALSE	-	-	-	-	-
chr3	17525001	17550000	TRUE	-	FALSE	-	-	-	-	-
chr3	17750001	17775000	TRUE	-	TRUE	-	-	-	-	-
chr3	17825001	17850000	TRUE	-	TRUE	-	TRUE	TRUE	-	-
chr3	19050001	19075000	-	-	-	-	TRUE	TRUE	TRUE	FALSE
chr3	25900001	25925000	TRUE	-	TRUE	-	-	-	-	-
chr3	26000001	26025000	TRUE	-	TRUE	-	-	-	-	-
chr3	26050001	26075000	TRUE	-	TRUE	-	-	-	-	-
chr3	26100001	26125000	TRUE	-	TRUE	-	TRUE	TRUE	-	-
chr3	36450001	36475000	-	-	-	-	TRUE	FALSE	-	-
chr3	36900001	36925000	-	-	TRUE	-	-	-	TRUE	TRUE
chr3	37925001	37950000	TRUE	-	TRUE	-	-	-	TRUE	FALSE
chr3	37950001	37975000	TRUE	-	TRUE	-	-	-	TRUE	TRUE
chr3	41425001	41450000	-	-	TRUE	-	TRUE	TRUE	-	-
chr3	52350001	52375000	TRUE	-	TRUE	-	-	-	TRUE	-
chr3	62675001	62700000	-	-	TRUE	-	TRUE	TRUE	-	-
chr3	71500001	71525000	TRUE	-	TRUE	-	-	-	TRUE	-
chr3	85575001	85600000	-	-	TRUE	-	TRUE	TRUE	-	-
chr3	87375001	87400000	-	-	TRUE	-	TRUE	TRUE	-	-
chr3	101950001	101975000	TRUE	-	TRUE	-	-	-	TRUE	TRUE
chr3	102050001	102075000	TRUE	-	TRUE	-	-	-	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr3	114525001	114550000	TRUE	TRUE	-	-	-	-	-	-
chr3	114575001	114600000	TRUE	FALSE	-	-	-	-	-	-
chr3	135375001	135400000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr3	135575001	135600000	-	-	TRUE	FALSE	-	-	TRUE	-
chr3	143625001	143650000	-	-	TRUE	FALSE	-	-	-	-
chr3	154750001	154775000	TRUE	FALSE	-	-	TRUE	FALSE	-	TRUE
chr3	165725001	165750000	-	-	-	-	TRUE	FALSE	-	FALSE
chr3	167800001	167825000	-	-	-	-	TRUE	FALSE	-	-
chr3	167825001	167850000	-	-	-	-	TRUE	FALSE	-	-
chr3	167850001	167875000	-	-	-	-	TRUE	FALSE	-	-
chr3	170775001	170800000	TRUE	TRUE	-	-	-	-	TRUE	-
chr3	170800001	170825000	-	-	-	-	TRUE	FALSE	-	-
chr3	173050001	173075000	-	-	-	-	TRUE	TRUE	-	-
chr3	175275001	175300000	TRUE	TRUE	-	-	TRUE	FALSE	-	-
chr3	183000001	183025000	-	-	TRUE	TRUE	TRUE	FALSE	-	-
chr3	190050001	190075000	-	-	TRUE	TRUE	TRUE	FALSE	-	-
chr3	195825001	195850000	-	-	-	-	-	-	-	-
chr4	3825001	3850000	-	-	-	-	-	-	TRUE	TRUE
chr4	11800001	11825000	TRUE	FALSE	-	-	-	-	-	-
chr4	13675001	13700000	TRUE	FALSE	-	-	-	-	TRUE	TRUE
chr4	21100001	21125000	-	-	TRUE	TRUE	-	-	-	-
chr4	34750001	34775000	-	-	TRUE	TRUE	-	-	-	-
chr4	35550001	35575000	-	-	TRUE	FALSE	-	-	-	-
chr4	35575001	35600000	-	-	-	-	TRUE	TRUE	-	-
chr4	35600001	35625000	-	-	TRUE	FALSE	-	-	TRUE	TRUE
chr4	36950001	36975000	-	-	TRUE	-	-	-	-	-
chr4	41900001	41925000	TRUE	TRUE	-	-	-	-	-	-
chr4	41925001	41950000	TRUE	TRUE	-	-	-	-	-	-
chr4	41950001	41975000	TRUE	FALSE	-	-	-	-	-	-
chr4	42025001	42050000	-	-	-	-	-	-	TRUE	TRUE
chr4	45575001	45600000	-	-	-	-	TRUE	FALSE	-	-
chr4	45675001	45700000	TRUE	FALSE	-	-	TRUE	TRUE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr4	58425001	58450000	-	-	-	-	TRUE	FALSE	-	-
chr4	64525001	64550000	-	-	-	-	TRUE	FALSE	-	-
chr4	64550001	64575000	-	-	-	-	TRUE	FALSE	-	-
chr4	65525001	65550000	-	-	-	-	-	-	TRUE	FALSE
chr4	70400001	70425000	-	-	-	-	-	-	TRUE	FALSE
chr4	76800001	76825000	-	-	TRUE	FALSE	-	-	-	-
chr4	76950001	76975000	-	-	TRUE	FALSE	-	-	-	-
chr4	78250001	78275000	-	-	-	-	-	-	TRUE	FALSE
chr4	86450001	86475000	-	-	-	-	TRUE	FALSE	-	-
chr4	90325001	90350000	TRUE	FALSE	-	-	TRUE	FALSE	-	-
chr4	91175001	91200000	-	-	-	-	TRUE	FALSE	-	-
chr4	91325001	91350000	-	-	-	-	TRUE	FALSE	-	-
chr4	91375001	91400000	-	-	-	-	TRUE	FALSE	-	-
chr4	91400001	91425000	-	-	-	-	TRUE	FALSE	-	-
chr4	101775001	101800000	TRUE	TRUE	-	-	-	-	-	-
chr4	109250001	109275000	-	-	-	-	TRUE	FALSE	-	-
chr4	116625001	116650000	-	-	-	-	TRUE	TRUE	-	-
chr4	132875001	132900000	-	-	-	-	-	-	TRUE	FALSE
chr4	133100001	133125000	-	-	-	-	TRUE	FALSE	-	-
chr4	133425001	133450000	-	TRUE	FALSE	-	-	-	-	-
chr4	135425001	135450000	-	-	-	TRUE	TRUE	TRUE	-	-
chr4	143500001	143525000	-	TRUE	FALSE	TRUE	TRUE	TRUE	-	-
chr4	144875001	144900000	-	TRUE	TRUE	TRUE	TRUE	TRUE	-	-
chr4	144900001	144925000	-	TRUE	FALSE	-	TRUE	TRUE	-	-
chr4	153175001	153200000	-	-	-	TRUE	TRUE	TRUE	-	-
chr4	156450001	156475000	-	-	-	-	-	-	TRUE	FALSE
chr4	157975001	158000000	-	TRUE	FALSE	-	-	-	TRUE	FALSE
chr4	159375001	159400000	-	-	-	-	-	-	TRUE	FALSE
chr4	161775001	161800000	-	TRUE	-	TRUE	FALSE	-	-	-
chr4	164150001	164175000	TRUE	TRUE	-	-	-	-	-	-
chr4	165650001	165675000	-	TRUE	FALSE	-	-	-	-	-
chr4	166075001	166100000	-	TRUE	TRUE	-	-	-	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr4	167375001	167400000	-	-	-	-	-	-	TRUE	FALSE
chr4	171225001	171250000	-	-	-	-	-	-	TRUE	FALSE
chr4	172575001	172600000	-	-	-	-	-	-	TRUE	-
chr4	177725001	177750000	TRUE	TRUE	-	-	-	-	-	-
chr4	179850001	179875000	-	-	TRUE	FALSE	-	-	TRUE	-
chr4	180200001	180225000	-	-	-	-	-	-	FALSE	-
chr4	184050001	184075000	-	-	-	-	-	-	TRUE	FALSE
chr4	184075001	184100000	-	-	-	-	-	-	TRUE	FALSE
chr4	184275001	184300000	-	-	TRUE	FALSE	-	-	-	-
chr4	187400001	187425000	-	-	TRUE	FALSE	-	-	-	-
chr4	189825001	189850000	-	-	-	-	-	-	TRUE	FALSE
chr5	150001	175000	-	-	-	-	-	-	TRUE	FALSE
chr5	3400001	3425000	-	-	-	-	-	-	TRUE	FALSE
chr5	3625001	3650000	-	-	-	-	-	-	TRUE	FALSE
chr5	4175001	4200000	-	-	-	-	-	-	TRUE	-
chr5	4200001	4225000	-	-	-	-	-	-	TRUE	-
chr5	4450001	4475000	-	-	-	-	-	-	TRUE	-
chr5	8725001	8750000	-	-	-	-	-	-	TRUE	TRUE
chr5	12925001	12950000	-	-	TRUE	FALSE	-	-	-	-
chr5	28975001	29000000	-	-	TRUE	FALSE	-	-	-	-
chr5	29000001	29025000	-	-	TRUE	FALSE	-	-	-	-
chr5	29625001	29650000	-	-	TRUE	FALSE	-	-	-	-
chr5	34425001	34450000	-	-	TRUE	FALSE	-	-	-	-
chr5	38650001	38675000	-	-	-	TRUE	-	-	TRUE	-
chr5	41600001	41625000	-	TRUE	TRUE	FALSE	-	-	TRUE	-
chr5	41650001	41675000	TRUE	TRUE	-	-	-	-	TRUE	-
chr5	41675001	41700000	TRUE	TRUE	-	-	-	-	TRUE	-
chr5	41700001	41725000	-	TRUE	-	-	-	-	TRUE	-
chr5	41925001	41950000	TRUE	TRUE	-	-	-	-	TRUE	-
chr5	60575001	60600000	-	-	-	-	-	-	TRUE	FALSE
chr5	64200001	64225000	-	-	-	TRUE	-	-	TRUE	-
chr5	79950001	79975000	-	TRUE	TRUE	FALSE	-	-	TRUE	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr5	87225001	87250000	-	-	-	-	TRUE	TRUE	-	-
chr5	87250001	87275000	-	-	-	-	TRUE	TRUE	-	-
chr5	113600001	113625000	-	-	-	-	-	-	TRUE	TRUE
chr5	116675001	116700000	TRUE	117375000	FALSE	-	-	-	-	FALSE
chr5	117350001	117375000	TRUE	117400000	TRUE	-	-	-	-	-
chr5	117375001	117400000	TRUE	117475000	TRUE	-	-	-	-	-
chr5	117450001	117475000	TRUE	117525000	TRUE	-	-	-	-	-
chr5	117500001	117525000	TRUE	117550000	TRUE	TRUE	-	TRUE	TRUE	-
chr5	117525001	117550000	TRUE	117600000	TRUE	TRUE	-	TRUE	TRUE	-
chr5	117575001	117600000	TRUE	118150001	TRUE	TRUE	-	-	-	-
chr5	118150001	118175000	TRUE	119700001	TRUE	FALSE	-	TRUE	TRUE	-
chr5	119700001	119725000	TRUE	119800000	TRUE	TRUE	-	-	FALSE	-
chr5	119775001	119800000	-	-	-	-	-	TRUE	TRUE	-
chr5	124775001	124800000	-	-	-	-	-	-	FALSE	-
chr5	128275001	128300000	-	-	-	-	-	TRUE	TRUE	-
chr5	128300001	128325000	-	-	-	-	-	-	FALSE	-
chr5	135575001	135600000	-	-	-	-	-	TRUE	TRUE	-
chr5	137025001	137050000	TRUE	-	FALSE	-	-	-	-	-
chr5	141350001	141375000	-	-	-	-	-	TRUE	TRUE	-
chr5	153525001	153550000	TRUE	-	TRUE	-	-	-	-	TRUE
chr5	159100001	159125000	-	-	-	-	-	-	-	-
chr5	166650001	166675000	-	-	TRUE	-	-	-	-	-
chr5	175150001	175175000	TRUE	-	FALSE	-	-	-	-	-
chr6	3700001	3725000	-	-	TRUE	-	-	-	-	-
chr6	8450001	8475000	-	-	TRUE	-	-	-	-	-
chr6	12775001	12800000	TRUE	-	TRUE	-	-	TRUE	TRUE	-
chr6	12875001	12900000	TRUE	-	TRUE	-	-	-	-	-
chr6	29375001	29400000	-	-	TRUE	-	-	FALSE	-	-
chr6	30375001	30400000	-	-	TRUE	-	-	FALSE	-	-
chr6	32350001	32375000	-	-	TRUE	-	-	FALSE	-	-
chr6	32375001	32400000	-	-	TRUE	-	-	TRUE	TRUE	-
chr6	32725001	32750000	-	-	TRUE	-	-	FALSE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr6	32750001	32775000	-	-	TRUE	FALSE	-	-	-	-
chr6	41050001	41075000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	41075001	41100000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	41100001	41125000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-
chr6	50025001	50050000	TRUE	FALSE	-	-	-	-	-	-
chr6	55750001	55775000	-	-	TRUE	FALSE	-	-	-	-
chr6	69150001	69175000	-	-	-	-	-	-	TRUE	FALSE
chr6	70325001	70350000	-	-	TRUE	FALSE	-	-	-	-
chr6	71850001	71875000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	78875001	78900000	-	-	TRUE	FALSE	-	-	-	-
chr6	79350001	79375000	-	-	TRUE	FALSE	TRUE	TRUE	FALSE	-
chr6	82675001	82700000	-	-	TRUE	FALSE	-	-	-	-
chr6	85900001	85925000	-	-	TRUE	FALSE	-	-	-	-
chr6	101750001	101775000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	103925001	103950000	-	-	-	-	-	-	TRUE	FALSE
chr6	105900001	105925000	TRUE	FALSE	-	-	-	-	-	-
chr6	105925001	105950000	TRUE	FALSE	-	-	-	-	-	-
chr6	110325001	110350000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-
chr6	113825001	113850000	TRUE	FALSE	-	-	TRUE	TRUE	FALSE	-
chr6	119625001	119650000	-	-	TRUE	FALSE	-	-	-	-
chr6	126875001	126900000	TRUE	TRUE	-	-	-	-	-	-
chr6	127350001	127375000	TRUE	FALSE	-	-	-	-	-	-
chr6	128625001	128650000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	129050001	129075000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	136025001	136050000	-	-	-	-	TRUE	TRUE	FALSE	-
chr6	159800001	159825000	-	-	TRUE	FALSE	-	-	-	-
chr6	167625001	167650000	-	-	TRUE	FALSE	-	-	-	-
chr7	8975001	9000000	-	-	TRUE	FALSE	-	-	TRUE	FALSE
chr7	13400001	13425000	-	-	TRUE	FALSE	-	-	TRUE	FALSE
chr7	19100001	19125000	TRUE	TRUE	-	-	-	-	-	-
chr7	19850001	19875000	-	-	-	-	-	-	TRUE	FALSE
chr7	42575001	42600000	-	-	-	-	TRUE	TRUE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr7	44400001	44425000	TRUE	FALSE	-	-	TRUE	FALSE	-	-
chr7	44425001	44450000	-	-	-	-	TRUE	FALSE	-	-
chr7	44450001	44475000	-	-	-	-	TRUE	FALSE	-	-
chr7	44475001	44500000	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	-	-
chr7	54125001	54150000	-	-	-	-	-	-	-	TRUE
chr7	62475001	62500000	-	-	TRUE	TRUE	-	-	-	FALSE
chr7	66600001	66625000	-	-	TRUE	TRUE	TRUE	TRUE	-	-
chr7	66625001	66650000	-	-	TRUE	TRUE	TRUE	TRUE	-	-
chr7	67425001	67450000	-	-	TRUE	TRUE	FALSE	-	-	-
chr7	81100001	81125000	-	-	TRUE	TRUE	FALSE	-	-	-
chr7	89150001	89175000	-	-	TRUE	TRUE	FALSE	-	-	-
chr7	10170001	101725000	TRUE	TRUE	-	-	-	-	-	-
chr7	117375001	117400000	TRUE	TRUE	-	-	-	-	-	-
chr7	117425001	117450000	TRUE	FALSE	-	-	-	-	-	-
chr7	119775001	119775000	-	TRUE	TRUE	FALSE	-	-	-	-
chr7	136475001	136500000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr7	136525001	136550000	-	-	-	-	TRUE	TRUE	-	-
chr7	139300001	139325000	TRUE	FALSE	-	-	-	-	-	-
chr7	144350001	144375000	-	-	-	-	TRUE	FALSE	-	-
chr7	144400001	144425000	-	-	-	-	TRUE	FALSE	-	-
chr7	147325001	147350000	TRUE	FALSE	-	-	TRUE	TRUE	-	-
chr7	149925001	149950000	-	-	-	-	-	-	TRUE	TRUE
chr7	149975001	150000000	-	-	-	-	-	-	TRUE	TRUE
chr8	1200001	1225000	-	-	TRUE	FALSE	-	-	-	-
chr8	2075001	2100000	-	-	TRUE	FALSE	-	-	-	-
chr8	3550001	3575000	-	-	TRUE	FALSE	-	-	-	-
chr8	5000001	5025000	-	-	TRUE	TRUE	-	-	-	-
chr8	5900001	5925000	-	-	TRUE	FALSE	-	-	-	-
chr8	5950001	5975000	-	-	-	-	-	-	TRUE	FALSE
chr8	11775001	11800000	TRUE	FALSE	TRUE	TRUE	-	-	-	-
chr8	13650001	13675000	-	-	TRUE	TRUE	-	-	-	-
chr8	14250001	14275000	-	-	TRUE	TRUE	-	-	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr8	15275001	15300000	-	TRUE	FALSE	-	TRUE	-	-	-
chr8	37475001	37500000	TRUE	FALSE	-	-	FALSE	-	-	-
chr8	41275001	41300000	TRUE	FALSE	-	-	-	-	-	-
chr8	47500001	47525000	-	-	-	-	TRUE	TRUE	TRUE	FALSE
chr8	67500001	67525000	TRUE	FALSE	-	-	TRUE	-	-	-
chr8	67525001	67550000	TRUE	FALSE	-	-	TRUE	FALSE	-	-
chr8	79075001	79100000	-	TRUE	FALSE	-	TRUE	TRUE	TRUE	-
chr8	81525001	81550000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	81600001	81625000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	84975001	85000000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	93000001	93025000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	93025001	93050000	TRUE	TRUE	-	-	TRUE	TRUE	TRUE	-
chr8	93050001	93075000	TRUE	TRUE	-	-	TRUE	TRUE	TRUE	-
chr8	93075001	93100000	TRUE	TRUE	-	-	TRUE	TRUE	TRUE	-
chr8	94950001	94975000	-	-	TRUE	FALSE	-	-	-	-
chr8	100800001	100825000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	111800001	111825000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	111825001	111850000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	120900001	120925000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	127100001	127125000	TRUE	TRUE	TRUE	-	TRUE	TRUE	TRUE	-
chr8	127125001	127150000	TRUE	TRUE	TRUE	-	TRUE	TRUE	TRUE	-
chr8	129550001	129575000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-
chr8	130950001	130975000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	133175001	133200000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	136250001	136275000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	136400001	136425000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	138900001	138925000	-	-	TRUE	FALSE	-	-	-	-
chr8	138950001	138975000	TRUE	FALSE	-	-	-	-	-	-
chr8	138975001	139000000	TRUE	TRUE	-	-	-	-	-	-
chr8	145100001	145125000	-	-	-	-	TRUE	TRUE	TRUE	-
chr8	145125001	145150000	TRUE	TRUE	-	-	TRUE	TRUE	TRUE	-
chr8	145775001	145800000	-	-	-	-	TRUE	TRUE	TRUE	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D	Fay and	Fay and	Fu and Li's	Fu and Li's	iHS Region	iHS
			Region	Present	Wu's H	Wu's H	F Region	F Present		Present
chr8	145800001	145825000	-	-	-	-	TRUE	TRUE	-	-
chr9	500001	525000	-	-	-	-	TRUE	FALSE	-	-
chr9	650001	675000	-	-	TRUE	FALSE	-	-	TRUE	FALSE
chr9	8750001	8775000	-	-	TRUE	TRUE	-	-	-	-
chr9	10450001	10475000	-	-	TRUE	TRUE	-	-	-	-
chr9	13850001	13875000	TRUE	TRUE	-	-	-	-	-	-
chr9	13875001	13900000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr9	17875001	17900000	TRUE	FALSE	-	-	-	-	-	-
chr9	82625001	82650000	TRUE	-	-	-	TRUE	TRUE	-	-
chr9	88325001	88350000	-	-	-	-	TRUE	TRUE	FALSE	-
chr9	88450001	88475000	-	-	-	-	TRUE	TRUE	FALSE	-
chr9	88475001	88500000	-	-	-	-	TRUE	TRUE	FALSE	-
chr9	91725001	91750000	TRUE	TRUE	-	-	-	-	-	-
chr9	92325001	92350000	TRUE	TRUE	-	-	-	-	-	-
chr9	106600001	106625000	-	-	TRUE	TRUE	-	-	-	-
chr9	106625001	106650000	-	-	TRUE	TRUE	-	-	-	-
chr9	106925001	106950000	-	-	TRUE	TRUE	-	-	-	-
chr9	108125001	108150000	TRUE	TRUE	-	-	TRUE	TRUE	FALSE	-
chr9	108150001	108175000	TRUE	FALSE	-	-	-	-	-	-
chr9	122625001	122650000	TRUE	FALSE	-	-	TRUE	TRUE	FALSE	-
chr9	125450001	125475000	-	-	-	-	TRUE	TRUE	FALSE	-
chr9	125525001	125550000	TRUE	TRUE	-	-	TRUE	TRUE	TRUE	-
chr9	132600001	132625000	-	-	TRUE	FALSE	-	-	-	-
chr9	132625001	132650000	-	-	TRUE	FALSE	-	-	-	-
chr9	137450001	137475000	TRUE	FALSE	-	-	TRUE	TRUE	FALSE	-
chr9	140700001	140725000	-	-	-	-	TRUE	TRUE	FALSE	-
chr10	900001	925000	-	-	TRUE	FALSE	-	-	-	-
chr10	3150001	3175000	-	-	TRUE	FALSE	-	-	-	-
chr10	3950001	3975000	-	-	-	-	TRUE	TRUE	FALSE	-
chr10	10350001	10375000	-	-	-	-	TRUE	TRUE	FALSE	-
chr10	15950001	15975000	TRUE	FALSE	-	-	TRUE	TRUE	FALSE	-
chr10	2275001	22750000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr10	25650001	25675000	-	-	-	-	-	-	TRUE	FALSE
chr10	26250001	26275000	-	-	-	-	TRUE	-	-	-
chr10	26325001	26350000	-	-	-	-	FALSE	-	TRUE	FALSE
chr10	38975001	39000000	-	-	-	-	-	-	TRUE	FALSE
chr10	39000001	39025000	-	-	-	-	-	-	TRUE	FALSE
chr10	39025001	39050000	-	-	-	-	-	-	TRUE	FALSE
chr10	49875001	49900000	-	-	-	-	-	-	TRUE	FALSE
chr10	52925001	52950000	-	-	-	-	-	-	TRUE	FALSE
chr10	52950001	52975000	-	-	-	-	TRUE	-	-	-
chr10	53000001	53025000	-	-	-	-	TRUE	-	TRUE	FALSE
chr10	56175001	56200000	-	-	-	-	TRUE	-	-	-
chr10	56475001	56500000	-	-	-	-	TRUE	-	-	-
chr10	59250001	59275000	-	-	-	-	TRUE	-	-	-
chr10	59775001	59800000	-	-	-	-	TRUE	-	-	-
chr10	60000001	60025000	-	-	-	-	FALSE	-	-	-
chr10	63250001	63275000	-	-	-	-	TRUE	-	-	-
chr10	67125001	67150000	-	-	-	-	FALSE	-	-	-
chr10	86875001	86900000	-	-	-	-	TRUE	-	-	-
chr10	93025001	93050000	-	-	-	-	TRUE	-	-	-
chr10	102525001	102550000	TRUE	TRUE	-	-	TRUE	-	TRUE	FALSE
chr10	107175001	107200000	TRUE	TRUE	-	-	-	-	-	-
chr10	107200001	107225000	-	-	-	-	TRUE	-	TRUE	FALSE
chr10	107225001	107250000	TRUE	TRUE	-	-	TRUE	-	TRUE	FALSE
chr10	130875001	130900000	-	-	-	-	TRUE	-	TRUE	FALSE
chr10	131300001	131325000	-	-	-	-	TRUE	-	-	-
chr10	132600001	132625000	-	-	-	-	TRUE	-	TRUE	FALSE
chr10	134450001	134475000	-	-	-	-	TRUE	-	TRUE	FALSE
chr11	1075001	1100000	-	-	-	-	-	-	TRUE	FALSE
chr11	2175001	2200000	-	-	-	-	-	-	TRUE	FALSE
chr11	4800001	4825000	TRUE	TRUE	-	-	TRUE	-	-	-
chr11	5200001	5225000	-	TRUE	-	FALSE	-	-	-	-
chr11	5375001	5400000	-	TRUE	-	FALSE	-	-	TRUE	FALSE

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr11	5575001	5600000	-	-	TRUE	-	FALSE	-	-	-	-
chr11	23725001	23750000	TRUE	-	TRUE	-	TRUE	TRUE	TRUE	FALSE	-
chr11	26100001	26125000	-	-	TRUE	-	FALSE	TRUE	TRUE	FALSE	-
chr11	26125001	26150000	-	-	TRUE	-	TRUE	TRUE	TRUE	FALSE	-
chr11	26175001	26200000	-	-	TRUE	-	FALSE	-	-	FALSE	-
chr11	37950001	37975000	-	-	-	-	-	TRUE	TRUE	FALSE	-
chr11	38250001	38275000	-	-	TRUE	-	TRUE	TRUE	TRUE	FALSE	-
chr11	39775001	39800000	-	-	TRUE	-	FALSE	-	-	FALSE	-
chr11	42100001	42125000	-	-	-	-	TRUE	TRUE	TRUE	FALSE	-
chr11	50225001	50250000	-	-	-	-	-	-	-	TRUE	TRUE
chr11	60900001	60925000	TRUE	-	TRUE	-	-	-	-	TRUE	TRUE
chr11	61325001	61350000	-	-	TRUE	-	TRUE	TRUE	TRUE	FALSE	-
chr11	78300001	78325000	TRUE	-	FALSE	-	-	-	-	TRUE	TRUE
chr11	78325001	78350000	TRUE	-	FALSE	-	-	-	-	TRUE	TRUE
chr11	79200001	79225000	-	-	-	-	-	-	-	TRUE	TRUE
chr11	86950001	86975000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	89375001	89400000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	91175001	91200000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	102225001	102250000	-	-	-	-	-	TRUE	TRUE	TRUE	TRUE
chr11	102250001	102275000	-	-	-	-	-	TRUE	TRUE	TRUE	TRUE
chr11	106100001	106125000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	112250001	112275000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	112650001	112675000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	117750001	117775000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr11	130575001	130600000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr12	7325001	7350000	TRUE	-	FALSE	-	-	-	-	TRUE	TRUE
chr12	8775001	8800000	TRUE	-	FALSE	-	-	-	-	TRUE	TRUE
chr12	9625001	9650000	-	-	TRUE	-	FALSE	-	-	TRUE	TRUE
chr12	10925001	10950000	-	-	TRUE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr12	14750001	14775000	TRUE	-	FALSE	-	-	TRUE	TRUE	TRUE	TRUE
chr12	24875001	24900000	-	-	-	-	-	-	-	TRUE	TRUE
chr12	29550001	29575000	-	-	-	-	-	-	-	TRUE	TRUE

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr12	40725001	40750000	-	-	-	-	-	-	TRUE	FALSE
chr12	44775001	44800000	TRUE	TRUE	-	-	-	-	-	-
chr12	44800001	44825000	TRUE	TRUE	-	-	-	-	-	-
chr12	44850001	44875000	TRUE	TRUE	-	-	-	-	-	-
chr12	54200001	54225000	-	-	-	-	TRUE	FALSE	-	-
chr12	58450001	58475000	-	-	-	-	-	-	TRUE	FALSE
chr12	83150001	83175000	TRUE	FALSE	-	-	-	-	-	-
chr12	83200001	83225000	TRUE	FALSE	-	-	-	-	-	-
chr12	85250001	85275000	TRUE	FALSE	-	-	-	-	-	-
chr12	85275001	85300000	TRUE	TRUE	-	-	-	-	-	-
chr12	88050001	88075000	-	-	-	-	-	-	FALSE	-
chr12	124375001	124400000	-	-	-	-	-	-	TRUE	FALSE
chr12	127950001	127975000	-	-	-	-	-	-	TRUE	FALSE
chr12	131325001	131350000	-	-	TRUE	TRUE	-	-	-	-
chr12	132800001	132825000	-	-	-	-	-	-	TRUE	FALSE
chr12	132825001	132850000	-	-	-	-	-	-	TRUE	FALSE
chr12	133250001	133275000	-	-	-	-	-	-	TRUE	FALSE
chr13	19775001	19800000	-	-	TRUE	FALSE	-	-	-	-
chr13	19925001	19950000	-	-	TRUE	FALSE	-	-	-	-
chr13	19950001	19975000	-	-	TRUE	FALSE	-	-	-	-
chr13	33450001	33475000	TRUE	FALSE	-	-	TRUE	TRUE	-	-
chr13	38900001	38925000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr13	42200001	42225000	-	-	-	-	TRUE	TRUE	-	-
chr13	42225001	42250000	-	-	TRUE	FALSE	-	-	TRUE	FALSE
chr13	48000001	48025000	-	-	TRUE	FALSE	TRUE	TRUE	-	-
chr13	48025001	48050000	-	-	-	-	TRUE	TRUE	-	-
chr13	52900001	52925000	-	-	TRUE	FALSE	-	-	TRUE	FALSE
chr13	68350001	68375000	-	-	TRUE	FALSE	TRUE	TRUE	-	-
chr13	84350001	84375000	-	-	-	-	TRUE	FALSE	-	-
chr13	87775001	87800000	-	TRUE	FALSE	-	-	-	TRUE	FALSE
chr13	114525001	114550000	-	-	-	-	-	-	TRUE	FALSE
chr13	114775001	114800000	-	-	-	-	-	-	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr14	20700001	20725000	-	-	-	-	-	-	TRUE	FALSE
chr14	24400001	24425000	-	-	-	-	TRUE	TRUE	-	-
chr14	46000001	46025000	-	-	TRUE	FALSE	TRUE	TRUE	FALSE	-
chr14	48300001	48325000	-	-	-	-	-	-	-	-
chr14	50125001	50150000	TRUE	TRUE	-	-	-	-	-	-
chr14	68675001	68700000	TRUE	TRUE	-	-	-	-	-	-
chr14	90425001	90450000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-
chr14	97925001	97950000	-	-	-	-	TRUE	TRUE	TRUE	-
chr14	105775001	105800000	TRUE	TRUE	-	-	TRUE	TRUE	FALSE	-
chr14	105800001	105825000	TRUE	TRUE	-	-	TRUE	TRUE	FALSE	-
chr14	105825001	105850000	-	-	-	-	TRUE	TRUE	FALSE	-
chr15	31325001	31350000	TRUE	TRUE	-	-	-	-	-	-
chr15	42625001	42650000	-	-	TRUE	FALSE	-	TRUE	TRUE	FALSE
chr15	45725001	45750000	-	-	-	-	TRUE	TRUE	TRUE	FALSE
chr15	55225001	55250000	-	-	-	-	TRUE	TRUE	TRUE	FALSE
chr15	63700001	63725000	-	-	-	-	-	-	TRUE	FALSE
chr15	65400001	65425000	TRUE	TRUE	-	-	-	-	TRUE	FALSE
chr15	68575001	68600000	TRUE	FALSE	-	TRUE	TRUE	TRUE	TRUE	FALSE
chr15	78175001	78200000	-	TRUE	FALSE	-	-	-	-	-
chr15	89000001	89025000	TRUE	FALSE	-	TRUE	TRUE	TRUE	TRUE	TRUE
chr16	1500001	1525000	-	-	TRUE	FALSE	-	-	-	TRUE
chr16	5575001	5600000	-	-	TRUE	FALSE	-	-	-	TRUE
chr16	8750001	8775000	-	-	TRUE	FALSE	-	-	-	TRUE
chr16	16275001	16300000	-	-	-	-	-	-	TRUE	FALSE
chr16	17125001	17150000	-	-	-	-	-	-	TRUE	FALSE
chr16	18750001	18775000	-	-	-	-	-	-	TRUE	FALSE
chr16	31300001	31325000	TRUE	FALSE	-	TRUE	TRUE	TRUE	TRUE	-
chr16	67200001	67225000	-	-	-	-	-	-	TRUE	TRUE
chr16	67225001	67250000	TRUE	TRUE	-	-	-	-	TRUE	TRUE
chr16	67250001	67275000	TRUE	TRUE	-	-	-	-	TRUE	TRUE
chr16	67650001	67675000	-	-	-	-	TRUE	TRUE	-	TRUE
chr16	77550001	77575000	-	-	-	-	-	-	TRUE	FALSE

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's <i>D</i> Region	Tajima's <i>D</i> Present	Fay and Wu's <i>H</i> Region	Fay and Wu's <i>H</i> Present	Fu and Li's <i>F</i> Region	Fu and Li's <i>F</i> Present	iHS Region	iHS Present
chr16	78450001	78475000	-	-	TRUE	FALSE	-	-	-	-
chr17	3175001	3200000	-	-	-	FALSE	TRUE	TRUE	FALSE	-
chr17	3200001	3225000	-	-	TRUE	FALSE	TRUE	TRUE	FALSE	-
chr17	5425001	5450000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	29275001	2930000	TRUE	TRUE	-	TRUE	TRUE	TRUE	TRUE	-
chr17	39825001	39850000	-	-	TRUE	FALSE	-	-	-	-
chr17	4375001	43775000	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	-
chr17	43775001	43800000	-	-	TRUE	FALSE	TRUE	TRUE	FALSE	-
chr17	43800001	43825000	-	-	-	-	TRUE	TRUE	FALSE	-
chr17	43825001	43850000	-	-	-	-	TRUE	TRUE	FALSE	-
chr17	43875001	43900000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	43900001	43925000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	43925001	43950000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	43950001	43975000	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	-
chr17	43975001	44000000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44000001	44025000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44025001	44050000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44050001	44075000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44075001	44100000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44100001	44125000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44125001	44150000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44150001	44175000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44175001	44200000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44200001	44225000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44225001	44250000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	44250001	44275000	TRUE	FALSE	-	-	TRUE	TRUE	TRUE	-
chr17	44275001	44300000	-	-	-	-	TRUE	TRUE	TRUE	-
chr17	55875001	55900000	-	-	TRUE	FALSE	-	-	-	-
chr17	58825001	58850000	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	-
chr17	64000001	64025000	-	-	TRUE	FALSE	-	-	-	-
chr17	69900001	69925000	-	-	-	-	-	-	TRUE	FALSE
chr17	78325001	78350000	-	-	-	-	-	-	TRUE	FALSE

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D	Tajima's D Region	Fay and Wu's H Present	Fay and Wu's H Region	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr17	80225001	80250000	TRUE	FALSE	-	-	-	-	-	FALSE
chr18	15125001	15150000	-	-	-	-	TRUE	FALSE	-	-
chr18	25075001	25100000	-	-	-	-	TRUE	TRUE	-	-
chr18	30475001	30500000	-	-	-	-	TRUE	TRUE	-	-
chr18	32500001	32525000	-	-	-	-	FALSE	-	-	-
chr18	50300001	50325000	-	-	TRUE	TRUE	-	-	-	-
chr18	51475001	51500000	-	-	TRUE	TRUE	-	-	-	-
chr18	53825001	53850000	-	-	-	-	TRUE	FALSE	-	-
chr18	58425001	58450000	-	-	TRUE	TRUE	FALSE	TRUE	-	-
chr18	58450001	58475000	-	-	-	-	TRUE	TRUE	-	-
chr18	63475001	63500000	TRUE	TRUE	-	-	-	-	-	-
chr18	73300001	73325000	-	-	TRUE	TRUE	FALSE	-	-	TRUE
chr18	77700001	77725000	-	-	-	-	-	-	-	FALSE
chr19	96500001	9675000	TRUE	FALSE	-	-	-	-	-	FALSE
chr19	12550001	12575000	-	-	-	-	TRUE	TRUE	-	FALSE
chr19	15475001	15500000	TRUE	FALSE	-	-	TRUE	TRUE	-	TRUE
chr19	15700001	15725000	-	-	-	-	-	-	-	TRUE
chr19	24275001	24300000	-	-	-	-	-	-	-	TRUE
chr19	28300001	28325000	-	-	-	-	-	-	-	TRUE
chr19	32075001	32100000	TRUE	FALSE	-	-	-	-	-	-
chr19	32375001	32400000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr19	35025001	35050000	TRUE	FALSE	-	-	TRUE	TRUE	-	-
chr19	39775001	39800000	TRUE	TRUE	-	-	-	-	-	-
chr19	42750001	42775000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr19	49525001	49550000	-	-	TRUE	TRUE	-	-	-	-
chr19	54775001	54800000	-	-	TRUE	TRUE	FALSE	-	-	TRUE
chr20	3500001	3525000	-	-	TRUE	TRUE	FALSE	TRUE	TRUE	-
chr20	5425001	5450000	-	-	TRUE	TRUE	FALSE	-	-	-
chr20	12725001	12750000	-	-	-	-	-	-	TRUE	FALSE
chr20	24800001	24825000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr20	24825001	24850000	TRUE	TRUE	-	-	TRUE	TRUE	-	-
chr20	24850001	24875000	TRUE	TRUE	-	-	TRUE	TRUE	-	-

Table S1: Regions from Hider et al 2013 that were replicated for selection.
(continued)

Chrom	Start	End	Tajima's D Region	Tajima's D Present	Fay and Wu's H Region	Fay and Wu's H Present	Fu and Li's F Region	Fu and Li's F Present	iHS Region	iHS Present
chr20	24875001	24900000	TRUE	TRUE	-	FALSE	TRUE	TRUE	-	-
chr20	24900001	24925000	-	-	TRUE	-	-	-	TRUE	FALSE
chr20	26175001	26200000	-	-	-	-	-	-	-	-
chr20	30225001	30250000	TRUE	TRUE	-	FALSE	-	-	-	-
chr20	44175001	44200000	-	-	TRUE	-	TRUE	-	-	-
chr20	53100001	53125000	TRUE	FALSE	-	FALSE	-	TRUE	-	-
chr20	53225001	53250000	TRUE	FALSE	-	FALSE	-	-	-	-
chr20	53425001	53450000	-	-	TRUE	FALSE	-	-	-	-
chr20	53900001	53925000	-	-	-	-	TRUE	TRUE	-	-
chr20	53925001	53950000	-	-	-	-	TRUE	TRUE	-	-
chr20	59225001	59250000	-	-	-	-	TRUE	TRUE	-	-
chr21	19250001	19275000	-	-	-	-	TRUE	TRUE	-	-
chr21	22425001	22450000	-	-	TRUE	FALSE	-	TRUE	TRUE	-
chr21	31500001	31525000	TRUE	FALSE	-	TRUE	-	TRUE	TRUE	-
chr21	37725001	37750000	-	-	TRUE	FALSE	-	TRUE	TRUE	-
chr21	41200001	41225000	-	-	-	TRUE	-	TRUE	TRUE	-
chr21	44275001	44300000	-	-	-	-	-	TRUE	TRUE	-
chr21	44300001	44325000	-	-	-	-	-	TRUE	TRUE	-
chr22	23300001	23325000	TRUE	TRUE	-	-	-	-	-	-
chr22	23325001	23350000	TRUE	TRUE	-	TRUE	-	TRUE	TRUE	-
chr22	24275001	24300000	-	-	TRUE	FALSE	-	-	-	-
chr22	46775001	46775000	TRUE	TRUE	-	TRUE	-	-	-	-
chr22	46775001	46800000	TRUE	TRUE	-	TRUE	-	-	-	-
chr22	49125001	49150000	-	-	-	-	-	TRUE	TRUE	FALSE

A1.2 Genes in 1st percentile for Polynesian populations for frequency based statistics

Table S2: All genes that had a window in the 1st percentile and value < 0 from a Polynesian population. Table can be found in the electronic supplement as file Appendices/03-intrasfsPol.csv

A1.3 Significant iHS and nSL markers in Polynesian populations

Table S3: Significant markers for iHS and nSL in Polynesian populations. Table can be found in the electronic supplement as file Appendices/03-ihsnsl-genes.csv

A1.4 Significant XP-EHH markers in Polynesian populations

Table S4: Genes that had a significant XP-EHH value in Polynesian populations. Table can be found in the electronic supplement as file Appendices/03-xpehh-Pol.csv

A1.5 Inflammatory and auto-immune genes with significant results in selection statistics for Polynesian populations

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations.

Population	Gene	XP-EHH						Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS		
CIM	<i>ABHD6</i>	7	1		1				
NZM	<i>ABHD6</i>		4						
SAM	<i>ABHD6</i>		5						
TON	<i>ABHD6</i>	7							
SAM	<i>AGER</i>				1				
SAM	<i>AGPAT1</i>				1				
NZM	<i>AHI1</i>				1			6	
TON	<i>AHI1</i>					1	1		
CIM	<i>AHR</i>						2		
NZM	<i>AKAP11</i>					1			
CIM	<i>AMT</i>	3						2	
NZM	<i>AMT</i>	3		1					
TON	<i>AMT</i>	7	1		4			2	
SAM	<i>ANXA3</i>					1			
CIM	<i>APEH</i>				2				
TON	<i>APEH</i>	1			3				

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH								Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL			
CIM	<i>APOBEC3G</i>				1							
CIM	<i>ARAP1</i>	7	2	5	5	2	5	1				
NZM	<i>ARAP1</i>	2	1	5	4	2	5					
SAM	<i>ARID5B</i>		1	3					4			
CIM	<i>ARID5B</i>								1			
SAM	<i>ATF6B</i>					1		4				
CIM	<i>ATF6B</i>								2			
NZM	<i>ATF6B</i>							3			2	
TON	<i>ATF6B</i>							4				
CIM	<i>ATG16L2</i>		2	5	5	2	5					
NZM	<i>ATG16L2</i>		1	3	2		3					
CIM	<i>ATM</i>	7	3	5	6	1	3	6	6			
NZM	<i>ATM</i>	3								8		6
SAM	<i>ATM</i>	7		5			6	1				
TON	<i>ATM</i>	7		1								
SAM	<i>ATP6V0A1</i>							1	1		2	
TON	<i>ATP6V0A1</i>							1				
CIM	<i>ATRIP</i>								1			
TON	<i>ATRIP</i>								1			
TON	<i>BACH2</i>								1		1	
SAM	<i>BANK1</i>	1		4				11	13			
TON	<i>BANK1</i>	2						9	10			
NZM	<i>BANK1</i>							2		4		
CIM	<i>BLK</i>	1			5		5				6	
NZM	<i>BLK</i>	7	2		6		5					
SAM	<i>BLK</i>	7	1		5		5				5	
TON	<i>BLK</i>	7	2		6		5				6	
SAM	<i>BRD2</i>					2		1				
NZM	<i>BRD2</i>							1				
SAM	<i>BRE</i>					1						
CIM	<i>BSN</i>	1			3						5	3
NZM	<i>BSN</i>	1			1					2	6	6
TON	<i>BSN</i>	5	1		5						5	3
SAM	<i>BTNL2</i>			2		3		1				
CIM	<i>BTNL2</i>								7			
NZM	<i>BTNL2</i>							5				
NZM	<i>C1QBP</i>							1				
SAM	<i>C1QBP</i>	5						2				
TON	<i>C1QBP</i>	7		1	1			2				
SAM	<i>C2</i>								1			
SAM	<i>C2orf74</i>				1							
CIM	<i>C2orf74</i>								1			
CIM	<i>C5orf30</i>	2										
SAM	<i>C6orf10</i>					1		4	2			
CIM	<i>C6orf10</i>								1			
TON	<i>C6orf10</i>							2	1			
SAM	<i>C6orf15</i>		2									
NZM	<i>C6orf48</i>							1				
NZM	<i>C7orf72</i>	1					1	1	1			
CIM	<i>CAMK2A</i>							1	1			
TON	<i>CAMK2G</i>	1		1								
SAM	<i>CAMK2G</i>							1	1			
SAM	<i>CAMTA1</i>					1		2	2			
CIM	<i>CAMTA1</i>							2	1			
NZM	<i>CAMTA1</i>							3	1			
NZM	<i>CARD11</i>	1		1			3					
SAM	<i>CARD11</i>							2				
TON	<i>CARD11</i>	1		2			3					
NZM	<i>CARD6</i>								1	1		
TON	<i>CARD9</i>					1						
SAM	<i>CASP7</i>							1				

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

XP-EHH											
Population	Gene		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H
TON	<i>CCDC122</i>								1		
TON	<i>CCDC36</i>	2								2	5
TON	<i>CCNY</i>								1	1	
CIM	<i>CCR2</i>		3	2	1			3			
NZM	<i>CCR2</i>		4	2	3			5			
SAM	<i>CCR2</i>				1						
TON	<i>CCR2</i>			1	2						
CIM	<i>CCR3</i>	1	2	1	6			5			2
NZM	<i>CCR3</i>	2	4	1	6			5		1	4
SAM	<i>CCR3</i>	3	3	1	6			5			4
TON	<i>CCR3</i>	3	3	1	6			5		7	5
SAM	<i>CCR9</i>				1						2
TON	<i>CCR9</i>		1	1							
NZM	<i>CCRL2</i>	2	2					1			
TON	<i>CD80</i>			1							
SAM	<i>CD86</i>					2	2				
TON	<i>CD86</i>					2					
NZM	<i>CDH1</i>				1				1		
SAM	<i>CDH13</i>	2			3	1		5	2		
TON	<i>CDH13</i>					1		2	1		
CIM	<i>CDH13</i>							9			
NZM	<i>CDH13</i>							6			
SAM	<i>CDH23</i>		2			3		12	16		
TON	<i>CDH23</i>					1		11	12		
CIM	<i>CDH23</i>							3	2		
CIM	<i>CDH3</i>							1			
NZM	<i>CDH3</i>							2			
NZM	<i>CDHR5</i>							1			
SAM	<i>CDHR5</i>							1			
NZM	<i>CDKAL1</i>								1		
CIM	<i>CEP250</i>	6				2		5	5		
NZM	<i>CEP250</i>	4						2	2		
SAM	<i>CEP250</i>							2	2		
TON	<i>CEP250</i>							4	3		
CIM	<i>CEP57</i>							1		2	
NZM	<i>CEP57</i>							1		5	
NZM	<i>CFB</i>							1			
SAM	<i>CFB</i>							2	1		
TON	<i>CFB</i>							1			
SAM	<i>CFLAR</i>	1									
NZM	<i>CLEC16A</i>							1			
CIM	<i>CNTNAP2</i>	5	2		1					1	4
NZM	<i>CNTNAP2</i>	2	1					2	2	12	3
SAM	<i>CNTNAP2</i>	3	3	4		1	2	3	3	4	3
TON	<i>CNTNAP2</i>	2	1	5					4	2	2
CIM	<i>COBL</i>							1			
NZM	<i>COBL</i>							2			
CIM	<i>COG6</i>	2		1	4						
NZM	<i>COG6</i>	4			5						
CIM	<i>CPAMD8</i>		3	1						3	
NZM	<i>CPAMD8</i>			1							
SAM	<i>CPAMD8</i>				3						
TON	<i>CPAMD8</i>		3	5				1			
TON	<i>CTDSP1</i>						1				
CIM	<i>DAG1</i>	2									
NZM	<i>DAG1</i>	2				1					
TON	<i>DAG1</i>	7				2					
NZM	<i>DAGLB</i>			1							
CIM	<i>DAP</i>				5						
NZM	<i>DAP</i>				5						
SAM	<i>DAP</i>				1						

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH										
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H	Fu & Li's F	Tajima's D
TON	<i>DAP</i>			1								
SAM	<i>DDX6</i>							1				
CIM	<i>DGKD</i>	4						1				
SAM	<i>DGKD</i>							1				
TON	<i>DGKD</i>							1				
CIM	<i>DIEXF</i>							1				
SAM	<i>DIEXF</i>							1				
NZM	<i>DLEU1</i>						4	4			2	
TON	<i>DLEU1</i>						1	1				
TON	<i>DNLZ</i>				1							
CIM	<i>EDEM3</i>					1						
SAM	<i>EGFL8</i>					1						
CIM	<i>ERAP1</i>	2	4		6		5	6	4			
NZM	<i>ERAP1</i>	5	4		6		5					
SAM	<i>ERAP1</i>			4	4		4	1	1			
TON	<i>ERAP1</i>			2	3		1					
CIM	<i>ERAP2</i>				1							
NZM	<i>ERAP2</i>				1							
SAM	<i>ERAP2</i>								1			
TON	<i>ETS1</i>							1	1			
CIM	<i>FAM171B</i>		1					1	2			
NZM	<i>FAM171B</i>								2			
SAM	<i>FAM175B</i>				1							
CIM	<i>FAM98B</i>		2									
CIM	<i>FCHSD2</i>	7	3	5	6	2	5	1		1		
NZM	<i>FCHSD2</i>	3	2	5	6		5			1		
TON	<i>FCHSD2</i>						1			1		
NZM	<i>FGFR1OP</i>					1						
NZM	<i>FIGNL1</i>	6	4	5	5		5	1	1			
SAM	<i>FIGNL1</i>								1			
SAM	<i>FKBPL</i>					1		1				
NZM	<i>FKBPL</i>							1				
TON	<i>FKBPL</i>							1				
SAM	<i>FLI1</i>							1				
TON	<i>FLI1</i>							1	1			
SAM	<i>FNIP1</i>							1				
CIM	<i>FOXP1</i>	7			4	1	1			6		
NZM	<i>FOXP1</i>	6				4				1		
SAM	<i>FOXP1</i>	7	3		6	1	5		9	1	6	
TON	<i>FOXP1</i>	7	3		6	1	5		7		3	
SAM	<i>FYN</i>					1						
SAM	<i>GABBR1</i>			2								
TON	<i>GABBR1</i>			1								
SAM	<i>GART</i>			1			2					
TON	<i>GART</i>	1		3			4					
CIM	<i>GIN1</i>	6		3								
NZM	<i>GIN1</i>	2		1								
SAM	<i>GIN1</i>	2		1								
TON	<i>GIN1</i>	6		3								
NZM	<i>GLIS3</i>					1						
TON	<i>GLIS3</i>					1		1				
CIM	<i>GLIS3</i>							1				
NZM	<i>GM2A</i>							1				
TON	<i>GPR18</i>					1						
NZM	<i>GPR183</i>					1						
TON	<i>GPR183</i>					1						
SAM	<i>GPSM3</i>					1						
CIM	<i>GPX3</i>							1				
SAM	<i>GPX3</i>							1	1			
CIM	<i>GRHL2</i>		5									
NZM	<i>GRHL2</i>		3									

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

		XP-EHH											
Population	Gene	AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
SAM	<i>GRHL2</i>			1									
TON	<i>GRHL2</i>			4									
NZM	<i>GRID2IP</i>		1										
SAM	<i>HLA-DOB</i>	2	5	3	3	1	1						
TON	<i>HLA-DOB</i>		2		1								
CIM	<i>HLA-DOB</i>							2					
SAM	<i>HLA-DQA1</i>			1		1							
SAM	<i>HLA-DQA2</i>	1	4	1	1	3		1	1				
CIM	<i>HLA-DQA2</i>							1					
NZM	<i>HLA-DQB1</i>			1			1						
SAM	<i>HLA-DQB1</i>			4	2	2	4	3					
TON	<i>HLA-DQB1</i>			1									
SAM	<i>HLA-DQB2</i>			3		1	3						
CIM	<i>HLA-DQB2</i>							4					
NZM	<i>HSPA1L</i>							1		1			
SAM	<i>IFNGR2</i>			1			2						
TON	<i>IFNGR2</i>			3			4						
CIM	<i>IFNGR2</i>							1					
CIM	<i>IKZF1</i>			2		1							
NZM	<i>IKZF1</i>	4	4	5	6	1	5	1	1				
SAM	<i>IKZF1</i>			1		1							
CIM	<i>IL12B</i>				1								
CIM	<i>IL12RB2</i>						1						
TON	<i>IL18R1</i>						1						
TON	<i>IL18RAP</i>							1					
TON	<i>IL1R1</i>							1					
TON	<i>IL1RL1</i>							1					
TON	<i>IL1RL2</i>			3			1						
NZM	<i>IL22RA2</i>						1						
CIM	<i>IL31RA</i>		1	5	5		3						
CIM	<i>IL6R</i>						1						
NZM	<i>IL6R</i>						1						
CIM	<i>IL6ST</i>				5								
NZM	<i>IL6ST</i>				2								
SAM	<i>INPP5D</i>							1					
TON	<i>INPP5D</i>							1					
TON	<i>INS</i>					1							
CIM	<i>IP6K1</i>			2				1		2			
TON	<i>IP6K1</i>			3									
SAM	<i>IP6K3</i>			2					1		3		
NZM	<i>IRF7</i>					1							
SAM	<i>IRF7</i>					1							
TON	<i>IRGM</i>					1							
CIM	<i>IRGM</i>						2						
CIM	<i>ITGA4</i>							1					
TON	<i>ITGAM</i>					1							
CIM	<i>ITGAV</i>						1		1				
SAM	<i>ITPR3</i>		1	4									
TON	<i>ITPR3</i>			4						5	3		
CIM	<i>JAK2</i>				1								
SAM	<i>JAK2</i>						1						
CIM	<i>JAZF1</i>			1									
SAM	<i>JAZF1</i>			4	3	2	1	4					
TON	<i>JAZF1</i>	5	4	5	6	1	5		1				
NZM	<i>KCNB2</i>	6	4	3	6	1	5	6	6				
SAM	<i>KCNB2</i>						1	2	2				
TON	<i>KCNB2</i>		2				4	5	5				
SAM	<i>KIAA1841</i>				1								
CIM	<i>KIAA1841</i>							1					
NZM	<i>KPNA7</i>					1							
TON	<i>KSR1</i>					1							

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH								Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL			
CIM	<i>LACC1</i>							1				
SAM	<i>LACC1</i>							1				
TON	<i>LACC1</i>							1	1			
CIM	<i>LEMD2</i>	1	1		4						2	
NZM	<i>LEMD2</i>	1	1		2							
TON	<i>LEMD2</i>	5	1		6						1	1
NZM	<i>LMO7</i>							1				
SAM	<i>LMO7</i>							2	1			
TON	<i>LMO7</i>							2	2			
CIM	<i>LPP</i>	7	2		6		5	3	1			
NZM	<i>LPP</i>	7	4		6	1	5	6	1			
SAM	<i>LPP</i>	7	2		5		4				1	
TON	<i>LPP</i>	7	3		5		4		1			
NZM	<i>LRRK18</i>					2	4					
CIM	<i>LRRK2</i>	1	3	4	5						1	
NZM	<i>LRRK2</i>	5	3	5	6		5					
SAM	<i>LRRK2</i>	6	3	5	6		1	1	2		3	
TON	<i>LRRK2</i>	1	3	4	5			3	3			
CIM	<i>LSP1</i>							1				
CIM	<i>LTF</i>				4							
NZM	<i>LTF</i>			4	5		1	3	1	1		
TON	<i>LY75</i>								3	3		
TON	<i>LYST</i>	2									5	3
NZM	<i>LYST</i>							1	1		2	5
CIM	<i>MAML2</i>								1			
CIM	<i>MANBA</i>	7	4	5	5			2				
NZM	<i>MANBA</i>	7	2	3	1							
SAM	<i>MANBA</i>	7	3	5	2							
TON	<i>MANBA</i>	7	2	3	1							
TON	<i>MARCH7</i>							1	1			
SAM	<i>MC1R</i>							5				
NZM	<i>MEG3</i>							1				
SAM	<i>METTL10</i>	6	1	3	1							
TON	<i>METTL10</i>		1									
CIM	<i>MIR146A</i>								1	1		
NZM	<i>MIR146A</i>								1	1		
SAM	<i>MIR210HG</i>						1					
TON	<i>MLH3</i>	3										
CIM	<i>MLN</i>			1		2						
NZM	<i>MLN</i>			1								
TON	<i>MLN</i>	5	1		6							
CIM	<i>MST1</i>					2						
NZM	<i>MST1</i>					1						
TON	<i>MST1</i>					3					2	2
SAM	<i>MSTO1</i>		5									
CIM	<i>MTMR3</i>							2	2		6	
NZM	<i>MTMR3</i>							2			1	
SAM	<i>MUC1</i>		1									
CIM	<i>NAB1</i>							1				
NZM	<i>NAB1</i>							1				
TON	<i>NFATC1</i>					1						
CIM	<i>NFKB1</i>	6		5	1							
NZM	<i>NFKB1</i>			1								
SAM	<i>NFKB1</i>	1		3								
TON	<i>NFKB1</i>			2								
NZM	<i>NFKBIL1</i>							1				
NZM	<i>NFKBIZ</i>			1								
SAM	<i>NOS2</i>			1				1				
TON	<i>NOS2</i>			1								
SAM	<i>NOTCH4</i>						1		2			
TON	<i>NUSAP1</i>								1			

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH						Fu & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	rSL		
SAM	<i>OSMR</i>							1			
TON	<i>OSMR</i>							1			
CIM	<i>PAM</i>	6	5			2		1	2		2
NZM	<i>PAM</i>	5	5							2	
SAM	<i>PAM</i>	4	5								
TON	<i>PAM</i>	7	5								
SAM	<i>PAPOLG</i>		2	5							
TON	<i>PAPOLG</i>			4							
TON	<i>PARK7</i>	1			1						
SAM	<i>PBX2</i>				1						
NZM	<i>PBX2</i>						1				
CIM	<i>PDE2A</i>	1	1		2	1					
NZM	<i>PDE2A</i>				2						
CIM	<i>PDGFB</i>				1						
SAM	<i>PDGFB</i>				1				1		
CIM	<i>PEX13</i>			2							
SAM	<i>PEX13</i>			3							
CIM	<i>PFKFB3</i>							1			
TON	<i>PFKFB4</i>							1		3	
NZM	<i>PHACTR2</i>						1	2			
NZM	<i>PHLDB1</i>						3			1	
SAM	<i>PHLDB1</i>						3			1	
NZM	<i>PHRF1</i>					1			3		
SAM	<i>PHRF1</i>					1					
CIM	<i>PHTF1</i>				1		1			4	
NZM	<i>PHTF1</i>						1			4	
SAM	<i>PLCL1</i>	2									
NZM	<i>PLCL2</i>	1									
TON	<i>PNKD</i>				1						
NZM	<i>PPAN-P2RY11</i>							1			
SAM	<i>PPCDC</i>						1				
CIM	<i>PPIP5K2</i>	6	3						6		
NZM	<i>PPIP5K2</i>	2	1								
SAM	<i>PPIP5K2</i>	2	1								
TON	<i>PPIP5K2</i>	6	3								
SAM	<i>PPT2</i>					1					
CIM	<i>PPT2</i>							1			
TON	<i>PPT2</i>							1			
SAM	<i>PPT2-EGFL8</i>				1						
CIM	<i>PPT2-EGFL8</i>							1			
TON	<i>PPT2-EGFL8</i>							1			
SAM	<i>PRDX6</i>			5							
TON	<i>PRDX6</i>			1							
NZM	<i>PRKCB</i>	3			3	4					
TON	<i>PRKCQ</i>							1			
CIM	<i>PROCR</i>					2					
SAM	<i>PRRT1</i>					1					
SAM	<i>PSMB9</i>	3	4	1	2	3					
TON	<i>PSMB9</i>	1	4		1	1					
NZM	<i>PSMB9</i>						4				
SAM	<i>PSMG1</i>				2	1	1				
TON	<i>PSMG1</i>				2	1	1				
TON	<i>PTPRC</i>							1			
CIM	<i>PUS10</i>	1		4							
NZM	<i>PUS10</i>				1						
SAM	<i>PUS10</i>		2		6				3		1
TON	<i>PUS10</i>		2		6				2	1	
SAM	<i>PVT1</i>			5				3	1		
TON	<i>PVT1</i>			2				2	1		
CIM	<i>PXK</i>	7	1		3						
NZM	<i>PXK</i>	6									

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH								Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL			
SAM	<i>PXK</i>	2										
TON	<i>PXK</i>	6										
SAM	<i>RAD51B</i>				1					10	31	31
NZM	<i>RAD51B</i>							1		4	2	20
SAM	<i>RAPGEF6</i>							1				
CIM	<i>RASGRP1</i>					1						
SAM	<i>RASSF7</i>				1							
TON	<i>RAVER1</i>							1				
CIM	<i>RBPJ</i>							1				
NZM	<i>RBPJ</i>							5				
CIM	<i>RDH10</i>			2								
NZM	<i>RDH10</i>	7	4	5	6	1	5					
SAM	<i>RDH10</i>			1	3			2	1			
TON	<i>RDH10</i>	1		3			2	1				
SAM	<i>REL</i>		2		6					2	4	
TON	<i>REL</i>		2		5					4	3	
SAM	<i>RERE</i>	7	1		1		4					1
TON	<i>RERE</i>	7	1		2	2	4					
NZM	<i>REV3L</i>						1					1
SAM	<i>REV3L</i>						1	1		2		3
CIM	<i>RHOA</i>	1										1
NZM	<i>RHOA</i>	2				1						3
TON	<i>RHOA</i>	7	1		3							1
TON	<i>RIPK2</i>								1			
NZM	<i>RMI2</i>						1					
SAM	<i>RNASEH2C</i>						1	1				
TON	<i>RNASEH2C</i>						1					
CIM	<i>RNF123</i>				2				1			1
NZM	<i>RNF123</i>				1					2		2
TON	<i>RNF123</i>				3							
SAM	<i>RNF5</i>					1						
CIM	<i>RPL7</i>		2									
NZM	<i>RPL7</i>	7	4	5	6	1	5					
SAM	<i>RPL7</i>		2	5	2			2				
TON	<i>RPL7</i>		1	5	2		2					
CIM	<i>RPS14</i>							1	1			
NZM	<i>RPS6KA2</i>					2						
TON	<i>RPS6KA2</i>					1						
NZM	<i>RTKN2</i>	6			1		4			1		
SAM	<i>RTKN2</i>	7	1		6	1	5			1		
TON	<i>RUNX1</i>					1						
CIM	<i>SBNO2</i>						1					
SAM	<i>SCAMP3</i>	2										
TON	<i>SDCCAG3</i>					1						
SAM	<i>SEC24C</i>			1						2	4	
TON	<i>SEC24C</i>			1								
CIM	<i>SELE</i>	5					4					
NZM	<i>SELE</i>	6		2			5					
CIM	<i>SELL</i>						4					
NZM	<i>SELL</i>	4			1		5					
TON	<i>SELP</i>				2							
NZM	<i>SEMA6D</i>							1	1			3
SAM	<i>SEMA6D</i>							1	2			
CIM	<i>SGIP1</i>					1						
TON	<i>SGIP1</i>	1										
CIM	<i>SHISA5</i>							1				
SAM	<i>SKTV2L</i>						4					
TON	<i>SLC11A1</i>						1					
SAM	<i>SLC15A2</i>						2					
TON	<i>SLC15A2</i>						1					
NZM	<i>SLC16A10</i>							1				

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH						Fu & Wu's H	Fu & Li's F	Tajima's D	Zeng's E
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL		
SAM	<i>SLC29A3</i>					1					
CIM	<i>SLC2A13</i>		3	4	5		2	7	4		3
NZM	<i>SLC2A13</i>	6	3	5	6		5				2
SAM	<i>SLC2A13</i>		3	4	5		3	3	3		
TON	<i>SLC2A13</i>		2	4				3	2		
CIM	<i>SLC36A2</i>							1			
CIM	<i>SLC36A3</i>							2	2		
TON	<i>SLC39A11</i>			1		1		1	1		
CIM	<i>SLC39A11</i>							1			
SAM	<i>SLC39A11</i>							1	1		
CIM	<i>SLC39A8</i>	5	1								
NZM	<i>SLC39A8</i>	4									
SAM	<i>SLC39A8</i>	6	1					1	4		
TON	<i>SLC39A8</i>								2		
SAM	<i>SLC44A4</i>								1		
TON	<i>SLC9A4</i>					1					
CIM	<i>SLCO6A1</i>	7	3	3	6		5	4	4		6
NZM	<i>SLCO6A1</i>	7	3	3	6		5	3	4		6
SAM	<i>SLCO6A1</i>	7	2	2	2		3				
TON	<i>SLCO6A1</i>	7	3	3	6		5				
CIM	<i>SMAD7</i>						2				
NZM	<i>SMAD7</i>		2	2	6	2	5		1		
TON	<i>SNAPC4</i>						1		1		
SAM	<i>SP140</i>							3			
NZM	<i>SPATA8</i>						1				
CIM	<i>STARD10</i>		1	5		2	1	1			
NZM	<i>STARD10</i>		3								
CIM	<i>STK11</i>							1	2		
NZM	<i>STK11</i>								1		
SAM	<i>STK19</i>							2	2		
CIM	<i>TCERG1L</i>			3		1			1		
SAM	<i>TCERG1L</i>								3		
TON	<i>TCERG1L</i>								1		
CIM	<i>TEF</i>	2									
NZM	<i>TEF</i>	2									
NZM	<i>TERF1</i>	7	2		6	1	5	1	2		
SAM	<i>TERF1</i>	1	1					2	2		
TON	<i>TERF1</i>	6	1		1		1	3	3		
NZM	<i>TET3</i>							1			
CIM	<i>THADA</i>	6	2		6		5			8	2
NZM	<i>THADA</i>	7	2		6		5			3	1
SAM	<i>THADA</i>	7	2		6		5			3	
TON	<i>THADA</i>	7	2		6		5			7	12
TON	<i>TMEM39A</i>			3			1				8
SAM	<i>TMEM50B</i>			1			2				
TON	<i>TMEM50B</i>			3			4				
CIM	<i>TNC</i>					1					
TON	<i>TNC</i>							1			
TON	<i>TNFAIP2</i>		1								
NZM	<i>TNFAIP3</i>							1	1		
TON	<i>TNFRSF9</i>					1					1
CIM	<i>TNFSF15</i>				2						
SAM	<i>TNFSF15</i>						1				
NZM	<i>TNFSF18</i>							1			1
CIM	<i>TNIP1</i>						3	2			
SAM	<i>TNIP1</i>						1				
SAM	<i>TNXB</i>					1		10			
CIM	<i>TNXB</i>						2				5
NZM	<i>TNXB</i>							3			8
TON	<i>TNXB</i>						3				
CIM	<i>TOB2</i>		2								

Table S5: Loci associated with various inflammatory and autoimmune diseases that showed signs of possible selection in Polynesian populations. (*continued*)

Population	Gene	XP-EHH										
		AFR	AMR	EAS	EUR	POL	SAS	iHS	nSL	Fay & Wu's H	Fu & Li's F	Tajima's D
NZM	<i>TOB2</i>	2										
NZM	<i>TRAF3IP2</i>				1							
CIM	<i>TREH</i>							1				
SAM	<i>TREH</i>							1				
TON	<i>TXND11</i>					2		1				
SAM	<i>UBD</i>		1									
CIM	<i>UBLCP1</i>							1				
SAM	<i>UNC5B</i>			1		1		1				
TON	<i>UNC5B</i>								1			
CIM	<i>USP34</i>			1								
SAM	<i>USP34</i>			2								
NZM	<i>USP34</i>						1					
CIM	<i>USP4</i>	2								6	6	
NZM	<i>USP4</i>	1								5	7	
TON	<i>USP4</i>	5	1		4				2		4	
CIM	<i>WDFY4</i>					1						
NZM	<i>WDFY4</i>	4	1	1		2	4					
CIM	<i>WDR78</i>					1				3		
CIM	<i>ZMIZ1</i>				5		5					
NZM	<i>ZMIZ1</i>				5		5					
SAM	<i>ZMIZ1</i>				5		4					
TON	<i>ZMIZ1</i>				5		1					
NZM	<i>ZNF965</i>					2			1		5	
SAM	<i>ZNF365</i>	3		1		1	4	1	1	2		
TON	<i>ZNF438</i>							1				
CIM	<i>ZNF831</i>							1				
NZM	<i>ZPBP</i>	1	1				2					
SAM	<i>ZPBP</i>						1					

XP-EHH is the number of populations from the super population that had at least one marker significant in the gene. Integrated haplotype homozygosity score and nSL are the number of significant markers. Fay and Wu's *H*, Fu and Li's *F*, Tajima's *D*, and Zeng's *E* are the number of windows intersecting the gene that met the lower threshold.

A2 Chapter 4 Tables

A2.1 Admixture cross-validation error

Table S6: Admixture cross-validation error for different values of K.

K	Cross-validation Error
1	0.4409
2	0.4155
3	0.3986
4	0.3929
5	0.3896
6	0.3870
7	0.3869
8	0.3860
9	0.3857
10	0.3858
11	0.3855
12	0.3856
13	0.3857
14	0.3857
15	0.3862

A2.2 GWAS catalog studies and references table

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog

Disease Trait	Reference
Auto-immunity and Auto-inflammatory	
Celiac disease	Dubois <i>et al.</i> (2010)
Celiac disease	Garner <i>et al.</i> (2014)
Celiac disease	Hunt <i>et al.</i> (2008)
Celiac disease	Östensson <i>et al.</i> (2013)
Celiac disease	van Heel <i>et al.</i> (2007)
Crohn's disease	Wellcome Trust Case Control Consortium (2007)
Crohn's disease	Barrett <i>et al.</i> (2008)
Crohn's disease	de Lange <i>et al.</i> (2017)
Crohn's disease	Franke <i>et al.</i> (2007)
Crohn's disease	Franke <i>et al.</i> (2010b)
Crohn's disease	Huang <i>et al.</i> (2012)
Crohn's disease	Jostins <i>et al.</i> (2012)
Crohn's disease	Julià <i>et al.</i> (2013)
Crohn's disease	Jung <i>et al.</i> (2016)
Crohn's disease	Kenny <i>et al.</i> (2012)
Crohn's disease	Libioulle <i>et al.</i> (2007)
Crohn's disease	Liu <i>et al.</i> (2015)
Crohn's disease	McGovern <i>et al.</i> (2010b)
Crohn's disease	Ostrowski <i>et al.</i> (2016)
Crohn's disease	Parkes <i>et al.</i> (2007)
Crohn's disease	Raelson <i>et al.</i> (2007)
Crohn's disease	Rioux <i>et al.</i> (2007)
Crohn's disease	Yamazaki <i>et al.</i> (2013)
Crohn's disease	Yang <i>et al.</i> (2014)
Inflammatory bowel disease	de Lange <i>et al.</i> (2017)
Inflammatory bowel disease	Duerr <i>et al.</i> (2006)
Inflammatory bowel disease	Jostins <i>et al.</i> (2012)
Inflammatory bowel disease	Kugathasan <i>et al.</i> (2008)

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog (*continued*)

Disease Trait	Reference
Inflammatory bowel disease	Liu <i>et al.</i> (2015)
Inflammatory bowel disease	Ostrowski <i>et al.</i> (2016)
Inflammatory bowel disease	Yang <i>et al.</i> (2016)
Multiple sclerosis	Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009)
Multiple sclerosis	Andlauer <i>et al.</i> (2016)
Multiple sclerosis	Aulchenko <i>et al.</i> (2008)
Multiple sclerosis	Comabella <i>et al.</i> (2008)
Multiple sclerosis	De Jager <i>et al.</i> (2009)
Multiple sclerosis	Gourraud <i>et al.</i> (2013)
Multiple sclerosis	Hafler <i>et al.</i> (2007)
Multiple sclerosis	Jakkula <i>et al.</i> (2010)
Multiple sclerosis	Martinelli-Boneschi <i>et al.</i> (2012)
Multiple sclerosis	Matesanz <i>et al.</i> (2012)
Multiple sclerosis	Nischwitz <i>et al.</i> (2010)
Multiple sclerosis	Patsopoulos <i>et al.</i> (2011)
Multiple sclerosis	Sanna <i>et al.</i> (2010)
Multiple sclerosis	Sawcer <i>et al.</i> (2011)
Multiple sclerosis	Wang <i>et al.</i> (2011a)
Primary biliary cirrhosis	Cordell <i>et al.</i> (2015)
Primary biliary cirrhosis	Hirschfield <i>et al.</i> (2009)
Primary biliary cirrhosis	Kawashima <i>et al.</i> (2017)
Primary biliary cirrhosis	Liu <i>et al.</i> (2010)
Primary biliary cirrhosis	Mells <i>et al.</i> (2011)
Primary biliary cirrhosis	Nakamura <i>et al.</i> (2012)
Psoriasis	Baurecht <i>et al.</i> (2015)
Psoriasis	Capon <i>et al.</i> (2008)
Psoriasis	Ellinghaus <i>et al.</i> (2010)
Psoriasis	Liu <i>et al.</i> (2008)
Psoriasis	Nair <i>et al.</i> (2009)
Psoriasis	Strange <i>et al.</i> (2010)
Psoriasis	Stuart <i>et al.</i> (2010)
Psoriasis	Tsoi <i>et al.</i> (2015)
Psoriasis	Yin <i>et al.</i> (2015)
Psoriasis	Zhang <i>et al.</i> (2009)
Rheumatoid arthritis	Wellcome Trust Case Control Consortium (2007)
Rheumatoid arthritis	Freudenberg <i>et al.</i> (2011)
Rheumatoid arthritis	Gregersen <i>et al.</i> (2009)
Rheumatoid arthritis	Hu <i>et al.</i> (2011)
Rheumatoid arthritis	Jiang <i>et al.</i> (2014)
Rheumatoid arthritis	Julià <i>et al.</i> (2008)
Rheumatoid arthritis	Kochi <i>et al.</i> (2010)
Rheumatoid arthritis	Myouzen <i>et al.</i> (2012)
Rheumatoid arthritis	Negi <i>et al.</i> (2013)
Rheumatoid arthritis	Okada <i>et al.</i> (2012c)
Rheumatoid arthritis	Okada <i>et al.</i> (2014)
Rheumatoid arthritis	Orozco <i>et al.</i> (2014)
Rheumatoid arthritis	Padyukov <i>et al.</i> (2011)
Rheumatoid arthritis	Plenge <i>et al.</i> (2007a)
Rheumatoid arthritis	Plenge <i>et al.</i> (2007b)
Rheumatoid arthritis	Raychaudhuri <i>et al.</i> (2008)
Rheumatoid arthritis	Saxena <i>et al.</i> (2017)

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog (*continued*)

Disease Trait	Reference
Rheumatoid arthritis	Stahl <i>et al.</i> (2010)
Rheumatoid arthritis	Terao <i>et al.</i> (2011)
Systemic lupus erythematosus	Alarcón-Riquelme <i>et al.</i> (2016)
Systemic lupus erythematosus	Armstrong <i>et al.</i> (2014)
Systemic lupus erythematosus	Bentham <i>et al.</i> (2015)
Systemic lupus erythematosus	Chung <i>et al.</i> (2011)
Systemic lupus erythematosus	Demirci <i>et al.</i> (2016)
Systemic lupus erythematosus	Graham <i>et al.</i> (2008)
Systemic lupus erythematosus	Han <i>et al.</i> (2009)
Systemic lupus erythematosus	Harley <i>et al.</i> (2008)
Systemic lupus erythematosus	Hom <i>et al.</i> (2008)
Systemic lupus erythematosus	Kozyrev <i>et al.</i> (2008)
Systemic lupus erythematosus	Lee <i>et al.</i> (2012)
Systemic lupus erythematosus	Lessard <i>et al.</i> (2016)
Systemic lupus erythematosus	Morris <i>et al.</i> (2016)
Systemic lupus erythematosus	Okada <i>et al.</i> (2012a)
Systemic lupus erythematosus	Yang <i>et al.</i> (2010c)
Systemic lupus erythematosus	Yang <i>et al.</i> (2011c)
Systemic lupus erythematosus	Yang <i>et al.</i> (2013b)
Type 1 diabetes	Wellcome Trust Case Control Consortium (2007)
Type 1 diabetes	Barrett <i>et al.</i> (2009a)
Type 1 diabetes	Bradfield <i>et al.</i> (2011)
Type 1 diabetes	Cooper <i>et al.</i> (2008)
Type 1 diabetes	Grant <i>et al.</i> (2009)
Type 1 diabetes	Hakonarson <i>et al.</i> (2007)
Type 1 diabetes	Hakonarson <i>et al.</i> (2008)
Type 1 diabetes	Huang <i>et al.</i> (2012)
Type 1 diabetes	Todd <i>et al.</i> (2007)
Type 1 diabetes	Wallace <i>et al.</i> (2010)
Ulcerative colitis	Anderson <i>et al.</i> (2011)
Ulcerative colitis	Asano <i>et al.</i> (2009)
Ulcerative colitis	Barrett <i>et al.</i> (2009b)
Ulcerative colitis	de Lange <i>et al.</i> (2017)
Ulcerative colitis	Franke <i>et al.</i> (2008)
Ulcerative colitis	Franke <i>et al.</i> (2010a)
Ulcerative colitis	Haritunians <i>et al.</i> (2010)
Ulcerative colitis	Jostins <i>et al.</i> (2012)
Ulcerative colitis	Julià <i>et al.</i> (2014)
Ulcerative colitis	Juyal <i>et al.</i> (2015)
Ulcerative colitis	Liu <i>et al.</i> (2015)
Ulcerative colitis	McGovern <i>et al.</i> (2010a)
Ulcerative colitis	Ostrowski <i>et al.</i> (2016)
Ulcerative colitis	Silverberg <i>et al.</i> (2009)
Ulcerative colitis	Yang <i>et al.</i> (2013a)
Vitiligo	Jin <i>et al.</i> (2010)
Vitiligo	Jin <i>et al.</i> (2011)
Vitiligo	Jin <i>et al.</i> (2012)
Vitiligo	Quan <i>et al.</i> (2010)
Vitiligo	Tang <i>et al.</i> (2013)
Gout and Urate	
Gout	Köttgen <i>et al.</i> (2013)
Gout	Li <i>et al.</i> (2015)

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog (*continued*)

Disease Trait	Reference
Gout	Matsuo <i>et al.</i> (2016)
Gout	Nakayama <i>et al.</i> (2017)
Gout	Sulem <i>et al.</i> (2011)
Renal overload gout	Nakayama <i>et al.</i> (2017)
Renal underexcretion gout	Nakayama <i>et al.</i> (2017)
Urate levels	Dehghan <i>et al.</i> (2008)
Urate levels	Döring <i>et al.</i> (2008)
Urate levels	Kamatani <i>et al.</i> (2010)
Urate levels	Köttgen <i>et al.</i> (2013)
Urate levels	Li <i>et al.</i> (2007)
Urate levels	Tin <i>et al.</i> (2011)
Urate levels	Vitart <i>et al.</i> (2008)
Urate levels	Wallace <i>et al.</i> (2008)
Urate levels	Yang <i>et al.</i> (2010b)
Urate levels in obese individuals	Huffman <i>et al.</i> (2015)
Urate levels in overweight individuals	Huffman <i>et al.</i> (2015)
Kidney Disease	
Chronic kidney disease	Köttgen <i>et al.</i> (2010)
Chronic kidney disease	Nanayakkara <i>et al.</i> (2014)
Chronic kidney disease	Pattaro <i>et al.</i> (2012)
Chronic kidney disease	Pattaro <i>et al.</i> (2016)
Chronic kidney disease and serum creatinine levels	Gudbjartsson <i>et al.</i> (2010)
End-stage renal disease (non-diabetic)	Bostrom <i>et al.</i> (2010)
Kidney function decline traits	Gorski <i>et al.</i> (2015)
Renal function and chronic kidney disease	Köttgen <i>et al.</i> (2009)
Malaria	
Malaria	Band <i>et al.</i> (2013)
Malaria	Jallow <i>et al.</i> (2009)
Malaria	Timmann <i>et al.</i> (2012)
Metabolic Syndrome	
Metabolic syndrome	Kraja <i>et al.</i> (2011)
Metabolic syndrome	Kristiansson <i>et al.</i> (2012)
Metabolic syndrome	Zabaneh and Balding (2010)
Metabolic syndrome (bivariate traits)	Kraja <i>et al.</i> (2011)
Neurological	
Alzheimer's disease	Abraham <i>et al.</i> (2008)
Alzheimer's disease	Antúnez <i>et al.</i> (2011)
Alzheimer's disease	Feulner <i>et al.</i> (2010)
Alzheimer's disease	Harold <i>et al.</i> (2009)
Alzheimer's disease	Heinzen <i>et al.</i> (2010)
Alzheimer's disease	Hollingsworth <i>et al.</i> (2011)
Alzheimer's disease	Jonsson <i>et al.</i> (2013)
Alzheimer's disease	Kamboh <i>et al.</i> (2012)
Alzheimer's disease	Lambert <i>et al.</i> (2009)
Alzheimer's disease	Lambert <i>et al.</i> (2013)
Alzheimer's disease	Li <i>et al.</i> (2008b)
Alzheimer's disease	Meda <i>et al.</i> (2012)
Alzheimer's disease	Nelson <i>et al.</i> (2014)
Alzheimer's disease	Pérez-Palma <i>et al.</i> (2014)
Alzheimer's disease	Seshadri <i>et al.</i> (2010)
Alzheimer's disease	Webster <i>et al.</i> (2008)
Parkinson's disease	Do <i>et al.</i> (2011)

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog (*continued*)

Disease Trait	Reference
Parkinson's disease	Foo <i>et al.</i> (2017)
Parkinson's disease	Hamza <i>et al.</i> (2010)
Parkinson's disease	Hill-Burns <i>et al.</i> (2014)
Parkinson's disease	Lill <i>et al.</i> (2012)
Parkinson's disease	Nalls <i>et al.</i> (2011)
Parkinson's disease	Nalls <i>et al.</i> (2014)
Parkinson's disease	Pankratz <i>et al.</i> (2012)
Parkinson's disease	Pickrell <i>et al.</i> (2016)
Parkinson's disease	Saad <i>et al.</i> (2011)
Parkinson's disease	Satake <i>et al.</i> (2009)
Parkinson's disease	Simón-Sánchez <i>et al.</i> (2009)
Parkinson's disease	Spencer <i>et al.</i> (2011)
Parkinson's disease	Vacic <i>et al.</i> (2014)
Obesity	
Body mass index	Berndt <i>et al.</i> (2013)
Body mass index	Frayling <i>et al.</i> (2007)
Body mass index	Graff <i>et al.</i> (2013)
Body mass index	Locke <i>et al.</i> (2015b)
Body mass index	Loos <i>et al.</i> (2008)
Body mass index	Monda <i>et al.</i> (2013)
Body mass index	Namjou <i>et al.</i> (2013)
Body mass index	Pei <i>et al.</i> (2014)
Body mass index	Speliotis <i>et al.</i> (2010)
Body mass index	Thorleifsson <i>et al.</i> (2009)
Body mass index	Warrington <i>et al.</i> (2015)
Body mass index	Wen <i>et al.</i> (2012)
Body mass index	Wen <i>et al.</i> (2014)
Body mass index	Willer <i>et al.</i> (2009)
Body mass index	Yang <i>et al.</i> (2012)
Body mass index variance	Ahmad <i>et al.</i> (2016)
Fat body mass	Pei <i>et al.</i> (2014)
Obesity	Berndt <i>et al.</i> (2013)
Obesity	Bradfield <i>et al.</i> (2012)
Obesity	Jiao <i>et al.</i> (2011)
Obesity	Meyre <i>et al.</i> (2009)
Obesity	Wang <i>et al.</i> (2011b)
Obesity (early onset extreme)	Scherag <i>et al.</i> (2010)
Obesity (early onset extreme)	Wheeler <i>et al.</i> (2013)
Obesity (extreme)	Cotsapas <i>et al.</i> (2009)
Obesity (extreme)	Paterno <i>et al.</i> (2011)
Obesity-related traits	Comuzzie <i>et al.</i> (2012)
Obesity-related traits	Melka <i>et al.</i> (2012)
Obesity-related traits	Scuteri <i>et al.</i> (2007)
Waist circumference adjusted for body mass index	Shungin <i>et al.</i> (2015)
Waist circumference adjusted for body mass index	Wen <i>et al.</i> (2016)
Waist-to-hip ratio adjusted for body mass index	Shungin <i>et al.</i> (2015)
Waist-to-hip ratio adjusted for body mass index	Wen <i>et al.</i> (2016)
T2D	
Type 2 diabetes	Wellcome Trust Case Control Consortium (2007)
Type 2 diabetes	Cho <i>et al.</i> (2011)
Type 2 diabetes	Cook and Morris (2016)
Type 2 diabetes	Cui <i>et al.</i> (2011)

Table S7: GWAS catalog studies used. Disease trait as specified in GWAS catalog (*continued*)

Disease Trait	Reference
Type 2 diabetes	Ghassibe-Sabbagh <i>et al.</i> (2014)
Type 2 diabetes	Hanson <i>et al.</i> (2014)
Type 2 diabetes	Hara <i>et al.</i> (2014)
Type 2 diabetes	Huang <i>et al.</i> (2012)
Type 2 diabetes	Imamura <i>et al.</i> (2012)
Type 2 diabetes	Imamura <i>et al.</i> (2016)
Type 2 diabetes	Kooner <i>et al.</i> (2011)
Type 2 diabetes	Li <i>et al.</i> (2013)
Type 2 diabetes	Ma <i>et al.</i> (2013)
Type 2 diabetes	Mahajan <i>et al.</i> (2014)
Type 2 diabetes	Ng <i>et al.</i> (2014)
Type 2 diabetes	Palmer <i>et al.</i> (2012)
Type 2 diabetes	Parra <i>et al.</i> (2011)
Type 2 diabetes	Perry <i>et al.</i> (2012)
Type 2 diabetes	Qi <i>et al.</i> (2010)
Type 2 diabetes	Saxena <i>et al.</i> (2007)
Type 2 diabetes	Saxena <i>et al.</i> (2013)
Type 2 diabetes	Scott <i>et al.</i> (2007)
Type 2 diabetes	Shu <i>et al.</i> (2010)
Type 2 diabetes	Sim <i>et al.</i> (2011)
Type 2 diabetes	Sladek <i>et al.</i> (2007)
Type 2 diabetes	Steinthorsdottir <i>et al.</i> (2007)
Type 2 diabetes	Tabassum <i>et al.</i> (2013)
Type 2 diabetes	Takeuchi <i>et al.</i> (2009)
Type 2 diabetes	Timpson <i>et al.</i> (2009)
Type 2 diabetes	Tsai <i>et al.</i> (2010)
Type 2 diabetes	Unoki <i>et al.</i> (2008)
Type 2 diabetes	Voight <i>et al.</i> (2010)
Type 2 diabetes	Williams <i>et al.</i> (2014)
Type 2 diabetes	Yamauchi <i>et al.</i> (2010)
Type 2 diabetes	Yasuda <i>et al.</i> (2008)
Type 2 diabetes	Zeggini <i>et al.</i> (2007)
Type 2 diabetes	Zeggini <i>et al.</i> (2008)
Type 2 diabetes and other traits	Rung <i>et al.</i> (2009)

A2.3 GWAS catalog disease gene lists

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7.

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
A1CF	Yes				
ABCA1		Yes			Yes
ABCG2	Yes				
ABO		Yes			
ACAN		Yes			
ACSM5				Yes	
ACVR1B	Yes				
ACVR2A	Yes				
ACVRL1	Yes				

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
ADAM30			Yes		
ADAMTS10		Yes			
ADAMTS17		Yes			
ADAMTS9		Yes	Yes		
ADAMTSL3		Yes			
ADCY3		Yes			
ADCY5			Yes		
ADCY9		Yes			
AGBL4		Yes			
ALDH16A1	Yes				
ALDH2		Yes			
ALMS1				Yes	
ANAPC13		Yes			
ANK1			Yes		
ANKRD55		Yes	Yes		
ANKS1A		Yes			
ANXA9				Yes	
AP3S2			Yes		
APOA5		Yes			Yes
APOB					Yes
APOE			Yes		Yes
ARF5			Yes		
ARL15		Yes	Yes		
ASAH2	Yes				
ASB3			Yes		
ATP2A1		Yes			
ATP8B2			Yes		
ATXN2	Yes			Yes	
ATXN2L		Yes			
B3GNT4	Yes				
BAZ1B	Yes				
BCAS3	Yes			Yes	
BCDIN3D		Yes			
BCL11A			Yes		
BCL2		Yes			
BDNF		Yes			
BMP2		Yes			
BNIPL				Yes	
BTNL2		Yes			
BUD13					Yes
C17orf82	Yes	Yes	Yes		
C18orf8		Yes			
C2CD4A			Yes		
C2CD4B			Yes		
C6orf106		Yes			
CABLES1		Yes			
CADM1		Yes			
CADM2		Yes			
CALCRL		Yes			
CAMK1D			Yes		
CAMK2B					Yes
CBLN1		Yes			
CCDC158				Yes	

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
CCDC85A			Yes		
CCDC91		Yes			
CCDC92		Yes			
CCNLJL		Yes			
CCR2		Yes			
CCR3		Yes			
CDC123			Yes		
CDC42BPG	Yes				
CDKAL1		Yes	Yes		
CDKN2A			Yes		
CDKN2B			Yes		
CEBPA		Yes			
CEP63		Yes			
CETP				Yes	
CHST8		Yes			
CLIP1		Yes			
CMIP		Yes			
CNPY2		Yes			
COBLL1		Yes			
COL6A1		Yes			
COL6A5		Yes			
CPEB4		Yes			
CPS1			Yes		
CREB1		Yes			
CST3			Yes		
CST4			Yes		
CST9			Yes		
CTCFL		Yes			
CTSS		Yes			
CUX2	Yes				
DACH1			Yes		
DCST2		Yes			
DDX1			Yes		
DGKB				Yes	
DGKG		Yes			
DIS3L2		Yes			
DMRTA1			Yes		
DMXL2		Yes			
DNAH10		Yes			
DNAJC27		Yes			
DNM3		Yes			
DNMT3A		Yes			
EDC4				Yes	
EFEMP1		Yes			
EHBP1		Yes			
ELavl4		Yes			
EPB41L4B		Yes			
ERBB4		Yes			
ETS2		Yes			
ETV5		Yes			
EYA2		Yes			
EZH2		Yes			
F12			Yes		

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
FAF1			Yes		
FAIM2		Yes			
FAM122A				Yes	
FAM13A		Yes			
FAM35A	Yes				
FAM60A			Yes		
FAM63A				Yes	
FANCL		Yes			
FBXW11		Yes			
FCER1A		Yes			
FER		Yes			
FGF2		Yes			
FGFR4		Yes			
FHIT		Yes			
FILIP1		Yes			
FITM2			Yes		
FLJ33534		Yes			
FNDC3B		Yes			
FNDC4				Yes	
FOXE1		Yes			
FOXO3		Yes			
FPGT-TNNI3K		Yes			
FSCN3			Yes		
FTO		Yes	Yes		Yes
FUBP1		Yes			
G6PC2					Yes
GALNT10		Yes			
GALNT2					Yes
GATM				Yes	
GBE1		Yes			
GCC1			Yes		
GCK					Yes
GCKR	Yes			Yes	Yes
GDF5		Yes			
GIPR		Yes			
GLIS3			Yes		
GNA12		Yes			
GNAS		Yes			
GNAT2		Yes			
GNPDA2		Yes			
GORAB		Yes			
GP2		Yes		Yes	
GPRC5B		Yes			
GPSM1			Yes		
GRB14		Yes	Yes		
GRID1		Yes			
GRK5			Yes		
GRK6				Yes	
GRP		Yes			
GTF3A		Yes			
HCG26					Yes
HHEX					
HHIP		Yes			

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
HIF1AN		Yes			
HIP1		Yes			
HIST1H2BF	Yes				
HIST1H2BH		Yes			
HIST1H3G		Yes			
HIST1H4E	Yes				
HLA-B			Yes		
HLA-DRB5		Yes			
HLF	Yes				
HLX		Yes			
HMG20A			Yes		
HMGA1		Yes			
HMGA2		Yes	Yes		
HMGCR		Yes			
HNF1A			Yes		
HNF1B			Yes		
HNF4A			Yes		
HNF4G	Yes	Yes			
HOXA11		Yes			
HOXB5		Yes			
HOXC12		Yes			
HOXC13		Yes			
HOXC4		Yes			
HOXC5		Yes			
HOXC6		Yes			
HS6ST3		Yes			
HSD17B12		Yes			
HSD17B4		Yes			
IDE			Yes		
IFNGR1		Yes			
IFT172				Yes	
IGF1R	Yes				
IGF2BP2			Yes		
IGFBP2			Yes		
INHBB	Yes				
INHBC	Yes				
INHBE	Yes				
INS-IGF2			Yes		
IQCK		Yes			
IRS1			Yes		
ITGB6		Yes	Yes		
ITGB8		Yes			
ITIH4		Yes			
ITPR2		Yes			
JAZF1		Yes	Yes		
JUND		Yes			
KAT8		Yes			
KCNJ11			Yes		
KCNJ2		Yes			
KCNK16			Yes		
KCNK3		Yes			
KCNMA1		Yes			
KCNQ1	Yes	Yes	Yes	Yes	

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
KCTD15		Yes			
KCTD19		Yes			
KLF13		Yes			
KLF14			Yes		
KLF9		Yes			
KREMEN1		Yes			
LAMA1			Yes		
LCORL		Yes			
LEKR1		Yes			
LEP			Yes		
LEPR		Yes			
LGR5			Yes		
LIN28B		Yes			
LINGO2		Yes			
LIPC					Yes
LMAN2				Yes	
LMX1B		Yes			
LOC285762		Yes			
LOC646736		Yes			
LOXL1		Yes			
LPL					Yes
LPP			Yes		
LRP1B			Yes		
LRP2	Yes				
LTBP1			Yes		
LTBP3	Yes				
LY86		Yes			
LYPLAL1		Yes			
MACF1			Yes		
MACROD1		Yes			
MAEA			Yes		
MAF	Yes	Yes			
MAP2K5		Yes			
MAP3K1		Yes			
MAP4K2	Yes				
MATK		Yes			
MBOAT7		Yes			
MC4R			Yes	Yes	
MEIS1		Yes			
MEN1	Yes				
MFAP2		Yes			
MICB					Yes
MIR148A		Yes			
MIR4686				Yes	
MIR548A2		Yes			
MLXIPL	Yes				Yes
MPHOSPH9			Yes		
MPPED2				Yes	
MSC		Yes			
MSRA		Yes			
MTCH2		Yes			
MTIF3		Yes			
MTNR1B		Yes	Yes		Yes

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
MUSTN1	Yes				
MYH9				Yes	
MYL2	Yes	Yes			
NAT8				Yes	
NAV1		Yes			
NCAM2		Yes			
NDUFS3		Yes			
NEGR1		Yes			
NFAT5	Yes				
NFE2L3		Yes			
NID2		Yes			
NIPAL1	Yes				
NKX2-6		Yes			
NLRC3		Yes			
NLRP3		Yes			
NOTCH2			Yes		
NR1H3					Yes
NRG4	Yes				
NRXN2	Yes				
NRXN3		Yes			
NT5C2		Yes			
NT5DC2		Yes			
NUDT3		Yes			
OLFM4		Yes			
OR10J1		Yes			
OR10J5		Yes			
OR2W5		Yes			
OVOL1	Yes				
PACS1		Yes			
PARK2		Yes			
PAX4			Yes		
PAX5		Yes			
PBRM1		Yes			
PCSK1		Yes			
PCSK5		Yes			
PDILT				Yes	
PDZK1	Yes				
PEMT		Yes			
PEPD		Yes	Yes		
PFN3				Yes	
PGPEP1		Yes			
PHTF2				Yes	
PIGC		Yes			
PIP5K1B				Yes	
PKLR	Yes				
PLXND1		Yes			
PMAIP1		Yes			
POC5		Yes			
POMC		Yes			
POU5F1			Yes		
PPARG		Yes	Yes		
PPP2R2C			Yes		
PPP2R3A		Yes			

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
PRC1			Yes		
PRKAG2	Yes			Yes	
PRKCH		Yes			
PRKD1		Yes			
PRKG2		Yes			
PRR7				Yes	
PSMD6			Yes		
PTBP2			Yes		
PTCH1			Yes		
PTPN11	Yes				
PTPRD			Yes		
PYGM	Yes				
QPCTL			Yes		
R3HDM2	Yes				
R3HDM1				Yes	
RABEP1			Yes		
RABEP2			Yes		
RARB			Yes		
RASA2			Yes		
RASGRP1			Yes		
RASGRP2	Yes				
RBM43				Yes	
RBMS1				Yes	
RFX3	Yes				
RFX7			Yes		
RGS14				Yes	
RMST			Yes		
RND3				Yes	
RPL27A			Yes		
RPTOR			Yes		
RREB1	Yes			Yes	
RSBN1L					Yes
RSPO3			Yes		
RYBP			Yes		
SACS				Yes	
SBK1			Yes		
SCARB2			Yes		
SEC16B			Yes		
SETDB1					Yes
SF1	Yes				
SF3B4			Yes		
SFMBT1	Yes				
SFXN2			Yes		
SGCG				Yes	
SH2B1			Yes		
SHROOM3				Yes	
SLC13A3				Yes	
SLC16A11				Yes	
SLC16A13				Yes	
SLC16A9	Yes				
SLC17A1	Yes				
SLC17A3	Yes				
SLC17A4	Yes				

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
SLC22A11	Yes				
SLC22A12	Yes				
SLC22A2				Yes	
SLC22A4		Yes			
SLC2A9	Yes				
SLC30A8			Yes		
SLC34A1				Yes	
SLC39A8		Yes			
SLC6A12				Yes	
SLC6A13				Yes	
SLC7A9				Yes	
SMAD3		Yes			
SMAD6		Yes			
SMG6		Yes			
SND1			Yes		
SNX10		Yes			
SPAG17		Yes			
SPATA5		Yes			
SPATA5L1				Yes	
SPRY2			Yes		
SRPK2		Yes			
SRR				Yes	
SSPN		Yes			
SSR1			Yes		
ST6GAL1			Yes		
STC1	Yes			Yes	
STXBP6		Yes			
SUFU		Yes			
SULT1A2		Yes			
TAL1		Yes			
TBX15		Yes			
TBX2				Yes	
TCF19			Yes		
TCF7L2		Yes	Yes		
TDRG1		Yes			
TFAP2B		Yes			
TFDP2				Yes	
TGFB2		Yes			
THADA				Yes	
TLE4				Yes	
TLR4		Yes			
TMEM154				Yes	
TMEM160		Yes			
TMEM171	Yes				
TMEM18		Yes			
TMEM60				Yes	
TNFAIP8		Yes			
TNKS		Yes			
TNNI3K		Yes			
TOMM40		Yes	Yes		Yes
TP53INP1			Yes		
TRIM46	Yes				
TRIM66		Yes			

Table S8: Disease associated genes by category from the GWAS catalog. For references see Table S7. (*continued*)

Gene name	Urate and gout	Obesity	T2D	Kidney disease	Metabolic syndrome
TRIP11		Yes			
TSEN15		Yes			
TSEN34		Yes			
TSPAN8			Yes		
TUB		Yes			
TUFM		Yes			
UBE2E2			Yes		
UBE2E3		Yes			
UBE2Q2	Yes			Yes	
UMOD				Yes	
USP37		Yes			
VEGFA	Yes	Yes		Yes	
VEGFB		Yes			
VPS26A			Yes		
WARS2		Yes			
WDR1	Yes				
WDR37				Yes	
WDR72				Yes	
WFS1			Yes		
XPA		Yes			
ZBED3			Yes		
ZBTB10		Yes			
ZBTB38		Yes			
ZC3H4		Yes			
ZFAND3			Yes		
ZFAND6			Yes		
ZFP64		Yes			
ZMIZ1			Yes		
ZNF608		Yes			
ZNF664		Yes			
ZNRF3		Yes			
ZZZ3		Yes			

A2.4 Site-frequency spectrum based selection and neutrality statistics summary tables

Table S9: Tajima's D summary statistics by population

Population	Mean	Std. Dev	Min	Lower 1%	Median	Upper 99%	Max
AFR							
ACB	2.410	0.735	-1.785	0.293	2.499	3.752	4.416
ASW	2.295	0.694	-1.703	0.252	2.385	3.548	4.186
ESN	2.390	0.708	-1.723	0.427	2.462	3.734	4.476
GWD	2.416	0.734	-1.543	0.370	2.492	3.804	4.610
LWK	2.403	0.715	-1.717	0.380	2.479	3.749	4.438
MSL	2.293	0.699	-1.644	0.369	2.359	3.628	4.297
YRI	2.440	0.714	-1.573	0.450	2.515	3.791	4.598
AMR							
CLM	2.388	0.904	-1.935	-0.290	2.539	3.899	4.659
MXL	2.113	0.894	-2.252	-0.527	2.260	3.635	4.319
PEL	1.925	0.999	-2.182	-0.853	2.070	3.688	4.505
PUR	2.484	0.885	-1.813	-0.141	2.634	3.970	4.827
EAS							
CDX	2.304	0.978	-2.257	-0.652	2.473	3.912	4.726
CHB	2.342	0.998	-1.977	-0.693	2.519	3.975	4.863
CHS	2.378	0.987	-2.171	-0.596	2.551	3.997	4.863
JPT	2.405	0.974	-2.078	-0.559	2.578	3.997	4.821
KHV	2.311	0.984	-2.642	-0.632	2.483	3.935	4.822
EUR							
CEU	2.317	0.978	-2.115	-0.613	2.490	3.916	4.859
FIN	2.354	0.954	-2.066	-0.520	2.521	3.927	4.747
GBR	2.288	0.966	-2.084	-0.618	2.462	3.880	4.719
IBS	2.332	0.978	-2.069	-0.581	2.497	3.953	4.780
NZC	2.271	1.006	-2.174	-0.731	2.449	3.919	4.779
TSI	2.361	0.974	-2.200	-0.558	2.530	3.965	4.884
POL							
CIM	1.875	1.073	-2.399	-1.093	2.024	3.771	4.673
NZM	1.929	1.042	-2.559	-0.945	2.074	3.780	4.742
SAM	2.061	1.041	-2.088	-0.979	2.226	3.836	4.657
TON	2.151	1.016	-2.129	-0.838	2.314	3.884	4.837
SAS							
BEB	2.393	0.879	-1.996	-0.279	2.542	3.872	4.535
GIH	2.454	0.903	-2.274	-0.281	2.607	3.978	4.700
ITU	2.460	0.890	-1.693	-0.232	2.603	3.972	4.772
PJL	2.442	0.892	-2.084	-0.268	2.592	3.930	4.717
STU	2.461	0.890	-1.881	-0.234	2.607	3.970	4.678

Table S10: Fay and Wu's H summary statistics by population

Population	Mean	Std. Dev	Min	Lower 1%	Median	Upper 99%	Max
AFR							
ACB	0.079	0.878	-6.602	-2.598	0.226	1.415	1.763
ASW	0.090	0.832	-6.403	-2.470	0.230	1.350	1.707
ESN	-0.016	0.947	-7.802	-2.924	0.141	1.398	1.728
GWD	-0.015	0.956	-8.113	-2.958	0.147	1.413	1.809
LWK	0.040	0.908	-7.157	-2.733	0.190	1.414	1.780
MSL	0.023	0.910	-6.862	-2.766	0.178	1.380	1.812
YRI	-0.004	0.949	-7.285	-2.922	0.156	1.418	1.794
AMR							
CLM	-0.614	1.277	-10.627	-4.682	-0.382	1.273	1.817
MXL	-0.758	1.317	-10.693	-4.926	-0.517	1.183	1.604
PEL	-1.171	1.537	-11.167	-5.894	-0.909	1.134	1.650
PUR	-0.503	1.219	-9.980	-4.355	-0.285	1.311	1.753
EAS							
CDX	-1.046	1.545	-11.374	-5.945	-0.749	1.220	1.708
CHB	-1.043	1.557	-10.958	-6.009	-0.745	1.233	1.796
CHS	-1.042	1.560	-10.566	-6.014	-0.741	1.241	1.860
JPT	-1.036	1.553	-10.782	-5.976	-0.740	1.239	1.745
KHV	-1.034	1.542	-10.703	-5.933	-0.739	1.228	1.755
EUR							
CEU	-0.816	1.417	-11.159	-5.314	-0.549	1.258	1.758
FIN	-0.826	1.424	-11.183	-5.354	-0.559	1.257	1.752
GBR	-0.808	1.407	-11.106	-5.286	-0.539	1.245	1.742
IBS	-0.810	1.416	-13.060	-5.312	-0.545	1.265	1.729
NZC	-0.814	1.411	-11.343	-5.303	-0.546	1.254	1.727
TSI	-0.801	1.413	-11.402	-5.274	-0.538	1.272	1.789
POL							
CIM	-1.379	1.685	-10.868	-6.540	-1.076	1.137	1.695
NZM	-1.281	1.627	-11.692	-6.294	-0.993	1.146	1.670
SAM	-1.261	1.648	-11.413	-6.408	-0.953	1.174	1.732
TON	-1.211	1.627	-11.282	-6.335	-0.904	1.185	1.702
SAS							
BEB	-0.683	1.321	-10.122	-4.924	-0.437	1.253	1.764
GIH	-0.723	1.367	-11.022	-5.124	-0.467	1.276	1.827
ITU	-0.715	1.359	-10.352	-5.060	-0.461	1.274	1.825
PJL	-0.681	1.333	-10.737	-4.953	-0.432	1.277	1.820
STU	-0.717	1.360	-10.861	-5.079	-0.466	1.273	1.807

Table S11: Fu and Li's F summary statistics by population

Population	Mean	Std. Dev	Min	Lower 1%	Median	Upper 99%	Max
AFR							
ACB	1.994	0.589	-4.068	0.024	2.072	3.004	3.995
ASW	1.958	0.560	-3.083	0.044	2.037	2.894	3.732
ESN	2.003	0.518	-3.209	0.431	2.050	2.985	4.004
GWD	2.004	0.547	-3.040	0.275	2.058	3.020	3.915
LWK	2.006	0.537	-2.779	0.286	2.060	2.998	3.997
MSL	1.956	0.518	-2.709	0.319	2.005	2.923	3.882
YRI	2.023	0.525	-2.574	0.411	2.070	3.020	4.072
AMR							
CLM	1.980	0.667	-3.024	-0.192	2.081	3.101	3.953
MXL	1.827	0.686	-4.809	-0.390	1.945	2.924	3.530
PEL	1.724	0.720	-4.263	-0.575	1.841	2.915	3.642
PUR	2.033	0.656	-4.016	-0.079	2.128	3.147	3.871
EAS							
CDX	1.945	0.648	-4.296	-0.068	2.030	3.073	3.859
CHB	1.931	0.693	-4.043	-0.288	2.030	3.113	3.863
CHS	1.969	0.658	-4.099	-0.036	2.055	3.128	3.934
JPT	2.000	0.633	-2.898	0.139	2.080	3.131	3.867
KHV	1.928	0.679	-5.262	-0.218	2.026	3.093	3.723
EUR							
CEU	1.900	0.737	-4.328	-0.513	2.025	3.095	3.919
FIN	1.953	0.685	-3.468	-0.253	2.055	3.103	3.964
GBR	1.894	0.726	-3.960	-0.465	2.018	3.070	3.852
IBS	1.903	0.741	-4.195	-0.486	2.026	3.117	3.885
NZC	1.821	0.824	-4.082	-0.890	1.982	3.090	4.013
TSI	1.922	0.735	-5.048	-0.459	2.043	3.121	3.903
POL							
CIM	1.642	0.778	-4.082	-0.818	1.775	2.929	3.700
NZM	1.699	0.753	-4.082	-0.699	1.827	2.950	3.650
SAM	1.728	0.761	-4.498	-0.715	1.857	2.981	3.951
TON	1.809	0.710	-4.427	-0.464	1.918	3.011	4.040
SAS							
BEB	1.988	0.642	-4.029	-0.089	2.085	3.069	3.660
GIH	2.015	0.651	-3.599	-0.077	2.110	3.118	3.797
ITU	2.022	0.638	-3.047	0.002	2.110	3.119	3.765
PJL	2.007	0.650	-4.068	-0.097	2.102	3.103	3.736
STU	2.026	0.634	-3.057	0.005	2.113	3.121	3.921

Table S12: Zeng's E summary statistics by population

Population	Mean	Std. Dev	Min	Lower 1%	Median	Upper 99%	Max
AFR							
ACB	1.838	0.857	-1.581	-0.132	1.835	3.862	6.147
ASW	1.731	0.822	-1.398	-0.172	1.729	3.682	5.790
ESN	1.893	0.866	-1.472	-0.051	1.877	3.973	6.511
GWD	1.916	0.879	-1.307	-0.062	1.905	4.025	6.077
LWK	1.862	0.856	-1.365	-0.085	1.850	3.902	6.283
MSL	1.785	0.853	-1.404	-0.118	1.770	3.836	5.996
YRI	1.925	0.868	-1.404	-0.026	1.912	4.010	6.411
AMR							
CLM	2.356	0.978	-1.684	0.037	2.356	4.683	7.109
MXL	2.259	0.972	-1.676	-0.017	2.248	4.623	6.992
PEL	2.423	1.052	-1.654	-0.004	2.407	4.998	7.652
PUR	2.345	0.966	-1.481	0.033	2.351	4.633	6.988
EAS							
CDX	2.614	0.996	-1.577	0.338	2.590	5.088	7.602
CHB	2.640	1.005	-1.464	0.319	2.619	5.145	7.427
CHS	2.665	1.002	-1.403	0.386	2.639	5.173	7.578
JPT	2.681	0.995	-1.291	0.409	2.661	5.167	7.412
KHV	2.610	1.002	-1.513	0.317	2.589	5.084	7.441
EUR							
CEU	2.452	1.026	-1.583	0.015	2.453	4.916	7.427
FIN	2.489	1.010	-1.828	0.118	2.485	4.928	7.413
GBR	2.425	1.015	-1.627	0.028	2.425	4.859	7.520
IBS	2.459	1.026	-1.619	0.024	2.456	4.917	8.544
NZC	2.415	1.044	-1.866	-0.097	2.423	4.897	7.444
TSI	2.474	1.021	-1.805	0.062	2.472	4.907	7.222
POL							
CIM	2.534	1.085	-1.561	0.051	2.510	5.208	7.827
NZM	2.501	1.078	-1.613	0.020	2.486	5.146	7.716
SAM	2.586	1.052	-1.505	0.201	2.557	5.210	7.566
TON	2.617	1.038	-1.444	0.271	2.587	5.203	7.240
SAS							
BEB	2.412	0.959	-1.633	0.163	2.408	4.726	7.032
GIH	2.486	0.975	-1.362	0.201	2.480	4.858	7.354
ITU	2.485	0.968	-1.450	0.219	2.480	4.831	7.125
PJL	2.446	0.969	-1.763	0.162	2.441	4.795	7.126
STU	2.487	0.969	-1.382	0.222	2.481	4.843	7.199

A2.5 Polynesian windows for clustering of the extremes

Table S13: Genes intersecting the windows of the 1st percentile Tajima's D in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-td-pol-firstper-genes.csv

Table S14: Genes intersecting the windows of the 99th percentile Tajima's D in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-td-pol-ninetyninthper-genes.csv

Table S15: Genes intersecting the windows of the 1st percentile Fay and Wu's H in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-fwh-pol-firstper-genes.csv

Table S16: Genes intersecting the windows of the 99th percentile Fay and Wu's H in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-fwh-pol-ninetyninethper-genes.csv

Table S17: Genes intersecting the windows of the 1st percentile Fu and Li's F in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-flf-pol-firstper-genes.csv

Table S18: Genes intersecting the windows of the 99th percentile Fu and Li's F in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-flf-pol-ninetyninethper-genes.csv

Table S19: Genes intersecting the windows of the 1st percentile Zeng's E in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-ze-pol-firstper-genes.csv

Table S20: Genes intersecting the windows of the 99th percentile Zeng's E in Polynesian populations. Table can be found in the electronic supplement as file Appendices/04-ze-pol-ninetyninethper-genes.csv

A2.6 Clustered regions from significant markers for iHS and nSL

Table S21: Positions of regions created by clustering significant markers for iHS or nSL by population. Table can be found in the electronic supplement as file Appendices/04-ihs-nsl-significant-marker-clusters.csv

A2.7 Clustered median centered metabolic disease genes

A2.7.1 Urate and Gout

Table S22: Tajima's D for windows at gout-associated loci. Table can be found in the electronic supplement as file Appendices/04-gout-td-gene-clus.csv

Table S23: Fay and Wu's H for windows at urate and gout-associated loci. Table can be found in the electronic supplement as file Appendices/04-gout-fwh-gene-clus.csv

Table S24: Fu and Li's F for windows at urate and gout-associated loci. Table can be found in the electronic supplement as file Appendices/04-gout-flf-gene-clus.csv

Table S25: Zeng's E for windows at urate and gout-associated loci. Table can be found in the electronic supplement as file Appendices/04-gout-ze-gene-clus.csv

A2.7.2 Obesity

Table S26: Tajima's D for windows at obesity-associated loci. Table can be found in the electronic supplement as file Appendices/04-obesity-td-gene-clus.csv

Table S27: Fay and Wu's H for windows at obesity-associated loci. Table can be found in the electronic supplement as file Appendices/04-obesity-fwh-gene-clus.csv

Table S28: Fu and Li's F for windows at obesity-associated loci. Table can be found in the electronic supplement as file Appendices/04-obesity-flf-gene-clus.csv

Table S29: Zeng's E for windows at obesity-associated loci. Table can be found in the electronic supplement as file Appendices/04-obesity-ze-gene-clus.csv

A2.7.3 Type 2 diabetes

Table S30: Tajima's D for windows at type 2 diabetes-associated loci. Table can be found in the electronic supplement as file Appendices/04-t2d-td-gene-clus.csv

Table S31: Fay and Wu's H for windows at type 2 diabetes-associated loci. Table can be found in the electronic supplement as file Appendices/04-t2d-fwh-gene-clus.csv

Table S32: Fu and Li's F for windows at type 2 diabetes-associated loci. Table can be found in the electronic supplement as file Appendices/04-t2d-flf-gene-clus.csv

Table S33: Zeng's E for windows at type 2 diabetes-associated loci. Table can be found in the electronic supplement as file Appendices/04-t2d-ze-gene-clus.csv

A2.7.4 Kidney disease

Table S34: Tajima's D for windows at kidney disease-associated loci. Table can be found in the electronic supplement as file Appendices/04-kd-td-gene-clus.csv

Table S35: Fay and Wu's H for windows at kidney disease-associated loci. Table can be found in the electronic supplement as file Appendices/04-kd-fwh-gene-clus.csv

Table S36: Fu and Li's F for windows at kidney disease-associated loci. Table can be found in the electronic supplement as file Appendices/04-kd-flf-gene-clus.csv

Table S37: Zeng's E for windows at kidney disease-associated loci. Table can be found in the electronic supplement as file Appendices/04-kd-ze-gene-clus.csv

A2.7.5 Metabolic syndrome

Table S38: Tajima's D for windows at metabolic syndrome-associated loci. Table can be found in the electronic supplement as file Appendices/04-metsyn-td-gene-clus.csv

Table S39: Fay and Wu's H for windows at metabolic syndrome-associated loci. Table can be found in the electronic supplement as file Appendices/04-metsyn-fwh-gene-clus.csv

Table S40: Fu and Li's F for windows at metabolic syndrome-associated loci. Table can be found in the electronic supplement as file Appendices/04-metsyn-flf-gene-clus.csv

Table S41: Zeng's E for windows at metabolic syndrome-associated loci. Table can be found in the electronic supplement as file Appendices/04-metsyn-ze-gene-clus.csv

Appendix B

Additional scripts

This appendix contains the main scripts that were used to generate results.

B1 GWAS catalog gene list creation

This was the R code used to generate the gene lists for the traits of interest in the thesis.

```
# libraries that are needed
library(tidyverse)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)

# read in the gwas catalog entries
gwas_cat <- read.delim(
  'gwas_catalog_v1.0.1-associations_e89_r2017-06-19.tsv',
  header=TRUE, stringsAsFactors = FALSE, sep='\t')

#filter gwas catalogue for results of genome wide significance
gwas_cat <- gwas_cat[ gwas_cat$P.VALUE < 5e-8,]

gwas_interested <- gwas_cat[grep(
  'metabolic syndrome|obesity|diabetes|urate|gout|body mass|lipid traits',
  gwas_cat$DISEASE.TRAIT, ignore.case = TRUE) ,]

# create kidney disease associated gene list
kd <- gwas_cat %>%
  filter(grep(
    DISEASE.TRAIT, pattern = 'kidney|renal', ignore.case=TRUE) &
```

```

!grepl(DISEASE.TRAIT,
       pattern = 'transplant|carcinoma>Type|stones|gout|related')) %>%
filter(DISEASE.TRAIT != "Diabetic kidney disease")

# filter the gene list down to diseases of interest
gwas_interested <- rbind(
  gwas_interested[grep(paste0("child|erectile|lean|autoantibodies|gestational|",
                               "cancer|psychopharmacol|metaformin|metformin|",
                               "obstructive|interaction|asthmatics|",
                               "omega|pain|cataracts|time|",
                               "bilirubin|chain|thyroid|zhi>Type 1|cystic"),
                        gwas_interested$DISEASE.TRAIT,
                        invert = TRUE, ignore.case = TRUE),],
  kd)

# pull out the genomic regions for all the transcripts for all the genes
# that we are interested in
gwas_genes <- sort(unique(unlist(
  strsplit(gwas_interested$REPORTED.GENE.S., split = ', ')))

gwas_genes_entrez <- na.omit(select(
  org.Hs.eg.db, keys = gwas_genes,
  columns = c('SYMBOL','ENTREZID'), keytype="SYMBOL"))

gwas_genes_ucsc <- merge(select(
  TxDb.Hsapiens.UCSC.hg19.knownGene,
  columns = c("TXID", "TXCHROM", "TXSTART", "TXEND", "TXSTRAND"),
  keys = gwas_genes_entrez$ENTREZID, keytype="GENEID"), gwas_genes_entrez,
  by.x ='GENEID', by.y = "ENTREZID")

gwas_genes_ucscGR <- GRanges(
  gwas_genes_ucsc[!is.na(gwas_genes_ucsc$TXID) ,])
gwas_genes_ucscGR <- gwas_genes_ucscGR[
  which(gwas_genes_ucscGR@seqnames %in% paste0('chr',1:22))]

# create the regions for the obesity associated genes
obesity_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
  unique(unlist(strsplit(
    gwas_interested[
      grep("obesity|body mass", gwas_interested$DISEASE.TRAIT,

```

```

ignore.case = TRUE),]$REPORTED.GENE.S., ', '))),]

# create the regions for the t2d associated genes
t2d_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
    unique(unlist(strsplit(
      gwas_interested[grep("diabetes",gwas_interested$DISEASE.TRAIT,
        ignore.case = TRUE),]$REPORTED.GENE.S., ', '))),]

# create the regions for metabolic syndrome associated genes
metsyn_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
    unique(unlist(strsplit(
      gwas_interested[grep("Syndrome",gwas_interested$DISEASE.TRAIT,
        ignore.case = TRUE),]$REPORTED.GENE.S., ', '))),]

# create the regions for urate and gout genes
gc_urate_gout_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
    unique(unlist(strsplit(
      gwas_interested[grep("urate|gout",gwas_interested$DISEASE.TRAIT,
        ignore.case = TRUE),]$REPORTED.GENE.S., ', '))),]

# the entries that match the diseases reported in Zhang et al 2013 Table 2
zhang_immune <- gwas_cat %>%
  filter(DISEASE.TRAIT %in% c("Crohn's disease", "Celiac disease",
    "Ulcerative colitis", "Inflammatory bowel disease",
    "Type 1 diabetes", "Rheumatoid arthritis",
    "Multiple sclerosis", "Psoriasis",
    "Systemic lupus erythematosus",
    "Primary biliary cirrhosis", "Vitiligo")) %>%
  dplyr::select(DISEASE.TRAIT, contains('gene'))

# entries that have Parkinson's or Alzheimers disease
neurological <- gwas_cat %>%
  filter(DISEASE.TRAIT %in% c("Parkinson's disease", "Alzheimer's disease")) %>%
  dplyr::select(DISEASE.TRAIT, contains('gene'))

#entries that have an association with malaria
malaria <- gwas_cat %>% filter(DISEASE.TRAIT %in% c( "Malaria")) %>%
  plyr::select(DISEASE.TRAIT, contains('gene'))

```

```

# find the coordinates for the transcripts for the genes
gwas_genes <- sort(unique(unlist(strsplit(c(zhang_immune$REPORTED.GENE.S.,
                                         neurological$REPORTED.GENE.S.,
                                         malaria$REPORTED.GENE.S.), split = ' ','))))

gwas_genes_entrez <- na.omit(AnnotationDbi::select(org.Hs.eg.db,
                                                       keys = gwas_genes,
                                                       columns = c('SYMBOL','ENTREZID'),
                                                       keytype="SYMBOL"))

gwas_genes_ucsc<- merge(AnnotationDbi::select(TxDb.Hsapiens.UCSC.hg19.knownGene ,
                                                 columns = c("TXID","TXCHROM",
                                                             "TXSTART","TXEND",
                                                             "TXSTRAND"),
                                                 keys = gwas_genes_entrez$ENTREZID,
                                                 keytype="GENEID"),
                           gwas_genes_entrez, by.x ='GENEID', by.y = "ENTREZID")

gwas_genes_ucscGR <- GRanges(gwas_genes_ucsc[!is.na(gwas_genes_ucsc$TXID) ,])
gwas_genes_ucscGR <- gwas_genes_ucscGR[which(gwas_genes_ucscGR@seqnames %in%
                                              paste0('chr',1:22))]

# get the coordinates for the transcripts for the malaria associated genes
malaria_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
  unique(unlist(strsplit(malaria$REPORTED.GENE.S., ' ', ))),]

# get the coordinates for the transcripts for the auto immune associated genes
zhang_immune_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
  unique(unlist(strsplit(zhang_immune$REPORTED.GENE.S., ' ', ))),]

# get the coordinates for the transcripts for the neuro disease associated genes
neurological_GR <- gwas_genes_ucscGR[
  gwas_genes_ucscGR$SYMBOL %in%
  unique(unlist(strsplit(neurological$REPORTED.GENE.S., ' ', ))),]

```

B2 SelectionTools Pipeline NeSI Scripts

The following are a series of script that were run in order to generate the results from the selectionTools 1.1 pipeline. They consist of

B2.1 unimputed_selection_pipeline.sl

This is the slurm workload manager script that was used on the NeSI PAN cluster to generate the selection results from selectionTools 1.1. It specifies the window and slide sizes for different statistics, as well as the gap size and penalties used in the iHS and nSL calculations.

NZ_1KGP_unimputed/unimputed_selection_pipeline.sl

```
#!/bin/bash

#SBATCH -J selection
#SBATCH -A uoo00008          # Project Account
#SBATCH --time=00:30:00        # Walltime
#SBATCH --mem-per-cpu=2048   # memory/cpu (in MB)
#SBATCH --cpus-per-task=1    # 12 OpenMP Threads
#SBATCH --array=1-22
#SBATCH --mail-user=murray.cadzow@otago.ac.nz
#SBATCH --mail-type=FAIL

POP=$1
i=$SLURM_ARRAY_TASK_ID
DIR=$SLURM_SUBMIT_DIR
module load Python/3.5.0-intel-2015a
module load R/3.2.1-intel-2015a
mkdir $TMP_DIR/${POP}
#srun tar -C $TMP_DIR/${POP}/ \
-xzf $DIR/${POP}.tar.gz ${POP}.chr${i}_biallelic_coreExome_markers.vcf
srun gzip -dc $DIR/${POP}.chr${i}.phased.vcf.gz > \
$TMP_DIR/${POP}/${POP}.chr${i}.phased.vcf
cd $TMP_DIR/${POP}/
#srun cat ${POP}.chr${i}_biallelic_coreExome_markers.vcf | \
grep -v '^##contig' > ${POP}_chr${i}_coreExome.vcf
srun python ~/.local/bin/selection_pipeline \
-i ${POP}.chr${i}.phased.vcf \
--phased-vcf \
-c $i \
--config-file NZ_1KGP_unimputed/unimputed_defaults_nesi_18-3-16.cfg \
--maf 0.01 \
```

```

--hwe 0.000001 \
--TajimaD 30 \
--fay-Window-Width 30 \
--fay-Window-Jump 30 \
--ehh-window-size 10 \
--ehh-overlap 2 \
--big-gap 200 \
--small-gap 20 \
--small-gap-penalty 20 \
--population $POP \
--cores 1 \
--no-clean-up \
--no-ihc

module unload Python/3.5.0-intel-2015a
module load Python/2.7.9-intel-2015a

PIPELINE_DIR=MerrimanSelectionPipeline/selection_pipeline
RESOURCE_DIR=MerrimanSelectionPipeline/referencefiles
srun python $PIPELINE_DIR/haps_interpolate.py \
--haps results/${POP}_aachanged.haps \
--output ${POP}_genetic_dist.haps \
--genetic-map $RESOURCE_DIR/genetic_maps/genetic_map_chr${i}_combined_b37.txt \
--physical-position-output ${POP}_genetic_dist.pos

srun python $PIPELINE_DIR/haps_to_selscan.py \
--haps ${POP}_genetic_dist.haps \
--pos ${POP}_genetic_dist.pos \
--output ${POP}_${i}_selscan \
--chr ${i}

srun gzip $TMP_DIR/${POP}/results/*vcforig
cd $TMP_DIR
srun tar -czf ${POP}.chr${i}.tar.gz *
mkdir -p /home/murray.cadzow/uoo00008/NZ_1KGP_unimputed/Indiv_pops_results/$POP
srun cp $TMP_DIR/*.tar.gz NZ_1KGP_unimputed/Indiv_pops_results/${POP}/

```

B2.2 unimputed_defaults_nesi-18-3-16.cfg

This the config file that is used as part of selectionTools 1.1.

unimputed_defaults_nesi_18-3-16.cfg

```

#
# Defaults config file for VCF process
#
# If the executables are on your path
# just the executable name is required.
#
# ? is the wildcard flag for the prefix options

[system]
cores_available = 1
# Library settings do not change, the library folder are
# appended to the path when running the program#
[environment]
LD_LIBRARY_PATH=MerrimanSelectionPipeline/lib
PERL5LIB=MerrimanSelectionPipeline/lib/perl5
[selection_pipeline]
selection_pipeline_executable = ~/.local/bin/selection_pipeline
[vcftools]
vcf_tools_executable = MerrimanSelectionPipeline/bin/vcftools
vcf_subset_executable = MerrimanSelectionPipeline/bin/vcf-subset
vcf_merge_executable = MerrimanSelectionPipeline/bin/vcf-merge
vcf_concat_executable = MerrimanSelectionPipeline/bin/vcf-concat
extra_args=
[genetic_map]
genetic_map_dir= MerrimanSelectionPipeline/referencefiles/genetic_maps
genetic_map_prefix=genetic_map_chr?_combined_b37.txt
[shapeit]
shapeit_executable= MerrimanSelectionPipeline/bin/shapeit
extra_args =
[impute2]
impute_executable = MerrimanSelectionPipeline/bin/impute2
impute_map_dir= MerrimanSelectionPipeline/referencefiles/impute_ref
impute_reference_dir= MerrimanSelectionPipeline/referencefiles/impute_ref
impute_map_prefix=genetic_map_chr?_combined_b37.txt
impute_reference_prefix=ALL.chr?.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nomono
extra_args =
[plink]
plink_executable =MerrimanSelectionPipeline/bin/plink
extra_args =

```

```

[Rscript]
rscript_executable = Rscript
indel_filter = MerrimanSelectionPipeline/corescripts/haps_indel_and_maf_filter.R
generate_rsb = MerrimanSelectionPipeline/corescripts/generate_rsb.R
extra_args=
[haps_scripts]
haps_to_hapmap_script= /home/murray.cadzow/.local/bin/haps_to_hapmap
haps_filter_script = /home/murray.cadzow/.local/bin/haps_filters
haps_interpolate_script = /home/murray.cadzow/.local/bin/haps_interpolate
haps_to_selscan_script = /home/murray.cadzow/.local/bin/haps_to_selscan
[ancestral_allele]
split_by_chromosome = True
# not used unless split_by_chromosome is set to False
ancestral_fasta_header_regex =
# not used unless split_by_chromosome is set to False
ancestral_fasta_file =
ancestral_allele_script= /home/murray.cadzow/.local/bin/ancestral_annotation

ancestral_fasta_dir=MerrimanSelectionPipeline/referencefiles/ancestral_ref/\homo_sapiens_ancestor_GRCh37_e65
ancestral_prefix=homo_sapiens_ancestor_?.fa
[qctool]
qctool_executable=MerrimanSelectionPipeline/bin/qctool
[selscan]
selscan_executable=MerrimanSelectionPipeline/bin/selscan
[multicore_ihh]
multicore_ihh = MerrimanSelectionPipeline/corescripts/multicore_iHH.R
[variscan]
variscan_executable = MerrimanSelectionPipeline/bin/variscan
[java]
java_executable = /usr/bin/java
[beagle]
beagle_jar = MerrimanSelectionPipeline/bin/beagle.jar
vm_size = 4g

```

B2.3 run_selscan.sl

This script was used on the results generated by NZ_1KGP_unimputed/unimputed_selection_pipeline.sl. It ran selscan on the haplotype files to calculate iHS.

run_selscan.sl

```

#!/bin/bash

#SBATCH -J selscan_array
#SBATCH -A uoo00008          # Project Account
#SBATCH --time=05:59:00        # Walltime
#SBATCH --mem=4096  # memory/node (in MB)
#SBATCH --cpus-per-task=4    # 10 OpenMP Threads
#SBATCH --array=1-22

POP=$1
i=$SLURM_ARRAY_TASK_ID
echo $POP chr $i

DIR=$SLURM_SUBMIT_DIR

srun tar -C $TMP_DIR -xzf ${POP}.chr${i}.tar.gz ${POP}/${POP}_${i}_selscan*
cd ${TMP_DIR}/${POP} && \
srun ~/uoo00008/selscan/src/selscan \
--ihs \
--hap ${POP}_${i}_selscan.selscanhaps \
--map ${POP}_${i}_selscan.selscanmap \
--ihs-detail \
--threads 4 \
--out ${TMP_DIR}/${POP}/${POP}_${i}

mkdir ${TMP_DIR}/${POP}/gap1mb
srun ~/uoo00008/selscan/src/selscan \
--ihs \
--hap ${POP}_${i}_selscan.selscanhaps \
--map ${POP}_${i}_selscan.selscanmap \
--ihs-detail \
--threads 4 \
--max-gap 1000000 \
--out ${TMP_DIR}/${POP}/gap1mb/${POP}_${i}

srun tar -czf ${POP}.chr${i}.ihs.tar.gz *log *.out gap1mb
srun cp ${POP}.chr${i}.ihs.tar.gz $DIR/

```

B2.4 run_nsl_selscan.sl

This script was used on the results generated by NZ_1KGP_unimputed/unimputed_selection_pipeline.sl. It ran selscan on the haplotype files to calculate nSL.

run_nsl_selscan.sl

```
#!/bin/bash

#SBATCH -J selscan_array
#SBATCH -A uoo00008          # Project Account
#SBATCH --time=05:59:00        # Walltime
#SBATCH --mem=2048   # memory/node (in MB)
#SBATCH --cpus-per-task=4    # 10 OpenMP Threads
#SBATCH --array=1-22

POP=$1
i=${SLURM_ARRAY_TASK_ID}
echo $POP chr $i

DIR=${SLURM_SUBMIT_DIR}

srun tar -C ${TMP_DIR} -xzf ${POP}.chr${i}.tar.gz ${POP}/${POP}_${i}_selscan*
cd ${TMP_DIR}/${POP} && \
srun selscan/src/selscan \
--nsl \
--hap ${POP}_${i}_selscan.selscanhaps \
--map ${POP}_${i}_selscan.selscanmap \
--threads 4 \
--out ${TMP_DIR}/${POP}/${POP}_${i}

mkdir ${TMP_DIR}/${POP}/gap1mb
srun selscan/src/selscan \
--nsl \
--hap ${POP}_${i}_selscan.selscanhaps \
--map ${POP}_${i}_selscan.selscanmap \
--threads 4 \
--max-gap 1000000 \
--out ${TMP_DIR}/${POP}/gap1mb/${POP}_${i}

srun tar -czf ${POP}.chr${i}.nsl.tar.gz *log *.out gap1mb
srun cp ${POP}.chr${i}.nsl.tar.gz $DIR
```

B2.5 run_xpehh.sl

This script was used on the results generated by NZ_1KGP_unimputed/unimputed_selection_pipeline.sl. It ran selscan on the haplotype files to calculate cross-population extended haplotype homozygosity (XP-EHH).

run_xpehh.sl

```
#!/bin/bash

#SBATCH -J selscan_array
#SBATCH -A uoo00008          # Project Account
#SBATCH --time=12:00:00        # Walltime
#SBATCH --mem=2048  # memory/node (in MB)
#SBATCH --cpus-per-task=4    # 10 OpenMP Threads
#SBATCH --array=1-22
#SBATCH -C sb

POP1=$1
POP2=$2
i=${SLURM_ARRAY_TASK_ID}
#echo $POP chr $i

DIR=${SLURM_SUBMIT_DIR}

module load Python/2.7.8-goolf-1.5.14

PIPELINE_DIR=MerrimanSelectionPipeline/selection_pipeline

srun tar -C ${TMP_DIR} -xzf \
${DIR}/${POP1}/${POP1}.chr${i}.tar.gz \
${POP1}/${POP1}_${i}_selscan.selscanhaps \
${POP1}/${POP1}_${i}_selscan.selscanmap

#need selscan file for POP
srun tar -C ${TMP_DIR} -xzf \
${DIR}/${POP2}/${POP2}.chr${i}.tar.gz \
${POP2}/${POP2}_${i}_selscan.selscanhaps \
${POP2}/${POP2}_${i}_selscan.selscanmap
#merge selscan files
mkdir ${TMP_DIR}/${POP1}_${POP2}

cd ${TMP_DIR}/${POP1}_${POP2} && \
srun python ${PIPELINE_DIR}/selscan_to_selscan_xpehh.py \
```

```

--pop1-prefix $TMP_DIR/$POP1/${POP1}_${i}_selscan \
--pop1-name ${POP1} \
--pop2-prefix $TMP_DIR/$POP2/${POP2}_${i}_selscan \
--pop2-name ${POP2} \
-c ${i} \
--out ./

ls $TMP_DIR/*

#srun tar -C $TMP_DIR -xzf ${POP1}_${POP2}.tar.gz \
${POP1}_${POP2}/*${i}.xpehh* ${POP1}_${POP2}/*${i}.*.xp*
cd $TMP_DIR && srun selscan/src/selscan \
--xpehh \
--hap ${POP1}_${POP2}/${POP1}_${i}.matches_${POP2}.xpehh_selscanhaps \
--map ${POP1}_${POP2}/${POP1}_${POP2}_${i}.xpehh_selscanmap \
--ref ${POP1}_${POP2}/${POP2}_${i}.matches_${POP1}.xp_ehh_selscanhaps \
--threads 4 \
--out ${TMP_DIR}/${POP1}_${POP2}_${i}

mkdir $TMP_DIR/gap1mb

srun ~/uoo00008/selscan/src/selscan \
--xpehh \
--hap ${POP1}_${POP2}/${POP1}_${i}.matches_${POP2}.xpehh_selscanhaps \
--map ${POP1}_${POP2}/${POP1}_${POP2}_${i}.xpehh_selscanmap \
--ref ${POP1}_${POP2}/${POP2}_${i}.matches_${POP1}.xp_ehh_selscanhaps \
--threads 4 \
--max-gap 1000000 \
--out ${TMP_DIR}/gap1mb/${POP1}_${POP2}_${i}

cd $TMP_DIR
srun tar -czf ${POP1}_${POP2}_chr${i}_xpehh.tar.gz ${POP1}_${POP2}* gap1mb
mkdir -p $DIR/xpehh/${POP1}_${POP2}
srun cp ${POP1}_${POP2}_chr${i}_xpehh.tar.gz $DIR/xpehh/${POP1}_${POP2}/

```

B2.6 Extract results

extract_results.sh

This bash script was used to extract the results from selectionTools 1.1 into a tidy directory structure, and also to normalise the iHSm nSL, and XP-EHH results.

```

#!/bin/bash

results_dir=NZ_coreExome_1kgp/data/dbload
input_dir=NZ_coreExome_1kgp/data/nesi_results

# extract all the results files generated from the selection pipeline
for i in $(seq 1 22)
do
    mkdir -p $results_dir/{daf,fawh,fixed_vcf,fst,ihc,kaks,nsl,tajd,xpehh}/chr${i}
    echo chr${i}
    echo ihc
    parallel 'tar -C {3}/ihc/chr{1} -xzf {2} *out *log --wildcards' :::: ${i} :::: \
        $(ls $input_dir/*/*chr${i}.ihc.tar.gz) :::: $results_dir
    echo nsl
    parallel 'tar -C {3}/nsl/chr{1} -xzf {2} *out *log --wildcards' :::: ${i} :::: \
        $(ls $input_dir/*/*chr${i}.nsl.tar.gz) :::: $results_dir
    echo tajd
    parallel 'tar -C {3}/tajd/chr{1} -xzf {2} *taj_d --wildcards' :::: ${i} :::: \
        $(ls $input_dir/*/*chr${i}.tar.gz) :::: $results_dir
    mv $results_dir/tajd/chr${i}/*/*results/* $results_dir/tajd/chr${i}/*
    rmdir $results_dir/tajd/chr${i}/*/*results
    echo fawh
    parallel 'tar -C {3}/fawh/chr{1} -xzf {2} *faw --wildcards' :::: ${i} :::: \
        $(ls $input_dir/*/*chr${i}.tar.gz) :::: $results_dir
    mv $results_dir/fawh/chr${i}/*/*results/* $results_dir/fawh/chr${i}/*
    rmdir $results_dir/fawh/chr${i}/*/*results

    echo daf
    parallel 'tar -C {3}/daf/chr{1} -xzf {2} *aachanged.af --wildcards' :::: ${i} \
        :::: $(ls $input_dir/*/*chr${i}.tar.gz) :::: $results_dir
    mv $results_dir/daf/chr${i}/*/*.af $results_dir/daf/chr${i}/*
    rename _aachanged.af ${i}_aachanged.af $results_dir/daf/chr${i}/*
    echo fixed_vcf
    parallel 'tar -C {3}/fixed_vcf/chr{1} -xzf {2} *fixed.vcf --wildcards' :::: ${i} \
        :::: $(ls $input_dir/*/*chr${i}.tar.gz) :::: $results_dir
    mv $results_dir/fixed_vcf/chr${i}/*/*vcf $results_dir/fixed_vcf/chr${i}/*
    echo xpehh
    parallel 'tar -C {3}/xpehh/chr{1} -xzf {2} *out *log --wildcards' :::: ${i} \
        :::: $(ls $input_dir/xpehh/*/*chr${i}_xpehh.tar.gz) :::: $results_dir

done

```

```

# calc the number of files for each population
for pop in $(basename -a $(find $input_dir -type d | \
    grep -v "xpehh" | grep [a-zA-Z] ) | grep -v "results")
do
    echo ----- $pop -----
    echo ihs $(ls $results_dir/ihc/chr*/${pop}*out | wc -l)

    echo ns1 $(ls $results_dir/ns1/chr*/${pop}*out | wc -l)

    echo fawh $(ls $results_dir/fawh/chr*/${pop}*faw | wc -l)

    echo tajd $(ls $results_dir/tajd/chr*/${pop}*taj_d | wc -l)

    echo daf $(ls $results_dir/daf/chr*/${pop}*af | wc -l)

    echo fixed_vcf $(ls $results_dir/fixed_vcf/chr*/${pop}*vcf | wc -l)

    echo xpehh $(ls $results_dir/xpehh/chr*/${pop}*out | wc -l)
    echo -----
    echo
    echo
done

# clean up some of the extraction
cd $results_dir/fixed_vcf/
for pop in $(basename -a $(find $input_dir -type d | \
    grep -v "xpehh" | grep [a-zA-Z] ) | grep -v "results")
do
    rmdir chr*/$pop
done

# combine vcfs, run snpEff and create panel files
parallel 'bgzip {} && tabix -f -p vcf {}.gz' :::: $(ls */*vcf)
mkdir combined
parallel -j 6 'bcftools merge $(ls chr{}/gz| \
    grep "AMR\|AFR\|EUR\|EAS\|SAS\|POL") | \
    bgzip -c > combined/combined_super_chr{}.vcf.gz' :::: $(seq 1 22)
parallel -j 6 'bcftools merge $(ls chr{}/gz | \
    grep -v "AMR\|AFR\|EUR\|EAS\|SAS\|POL") | \
    bgzip -c > combined/combined_chr{}.vcf.gz' :::: $(seq 1 22)
parallel -j 16 'java -jar snpEff4.2/snpEff/snpEff.jar \

```

```

-c.snpEff4.2/snpEff/snpEff.config \
-v.GRCh37.75 \
-strict \
-stats chr{2}/{1}{2}.html chr{2}/{1}{2}_fixed.vcf.gz \
bgzip -c > chr{2}/{1}{2}_fixed_ann.vcf.gz' :: $(ls chr22/*gz | \
sed 's/chr22///g' | cut -d'2' -f1) :: $(seq 1 22)
for pop in $(basename -a -s 1_fixed.vcf.gz chr1/*fixed.vcf.gz)
do
zcat chr1/${pop}1_fixed.vcf.gz | \
head -100 | grep '^#CHR' | \
cut -f10- | tr '\t' '\n' | \
awk '{print $1"\t"$1}' > ${pop}.panel
done

# calculate Fst
parallel -j 10 \
for p1 in POL \
do
for p2 in AFR AMR EUR EAS SAS
do
vcftools --gzvcf combined/combined_super_chr{}.vcf.gz \
--weir-fst-pop ${p1}.panel --weir-fst-pop ${p2}.panel \
--fst-window-size 1000000 --fst-window-step 1000 \
--out ../fst/chr{}/${p1}_${p2}_{}
done
done
' :: $(seq 1 22)

parallel -j 10 \
for p1 in CIM NZC NZM TON SAM \
do
for p2 in $(basename -a -s .panel *.panel | \
grep -v "AMR\|AFR\|EUR\|EAS\|SAS\|POL")
do
if [[ $p1 != $p2 ]]
then
vcftools --gzvcf combined/combined_chr{}.vcf.gz \
--weir-fst-pop ${p1}.panel --weir-fst-pop ${p2}.panel \
--fst-window-size 1000000 --fst-window-step 1000 \
--out ../fst/chr{}/${p1}_${p2}_{}
fi

```

```

done
done
' :::: $(seq 1 22)

# normalise the ihs files
cd $results_dir/ihc
for pop in $(ls chr1/*out | cut -d'/' -f2 | cut -d'_' -f1 | sort -u)
do
    selscan/bin/linux/norm --ihs \
        --files chr*/${pop}_*.out --crit-percent 0.99 --log ${pop}.log
    selscan/bin/linux/norm \
        --ihs --files chr*/gap1mb/${pop}_*.out --crit-percent 0.99 \
        --log ${pop}_gap1mb.log
done

# rename the ihs files to remove the underscore
for i in $(seq 1 22)
do
    rename _${i} ${i} chr${i}/*
    rename _${i} ${i} chr${i}/gap1mb/*
    mkdir -p norm/chr${i}/gap1mb
    mv chr${i}/*norm norm/chr${i}/
    mv chr${i}/gap1mb/*norm norm/chr${i}/gap1mb/
done

# normalise the ns1 files
cd $results_dir/ns1
for pop in $(ls chr1/*out | cut -d'/' -f2 | cut -d'_' -f1 | sort -u)
do
    ~/Murray/src/selscan/bin/linux/norm --ihs \
        --files chr*/${pop}_*.out --crit-percent 0.99 \
        --log ${pop}.log
    ~/Murray/src/selscan/bin/linux/norm --ihs \
        --files chr*/gap1mb/${pop}_*.out --crit-percent 0.99 \
        --log ${pop}_gap1mb.log
done

# rename and move the ns1 files
for i in $(seq 1 22)
do

```

```

rename _${i} ${i} chr${i}/*
mkdir -p norm/chr${i}
mv chr${i}/*norm norm/chr${i}/
mv chr${i}/gap1mb/*norm norm/chr${i}/gap1mb/

done

# normalise xpehh files
cd $results_dir/xpehh
for p1 in POL CIM NZC NZM NAD TON SAM EPN WPN;
do
  for p2 in $(ls chr1/${p1}*.*.out | cut -d'_' -f2)
  do
    if [[ ${p1} != ${p2} ]]
    then
      selscan/bin/linux/norm --xpehh \
        --files chr*/${p1}_${p2}.*.out --crit-percent 0.99 \
        --log ${p1}_${p2}.log
      selscan/bin/linux/norm --xpehh \
        --files chr*/gap1mb/${p1}_${p2}.*.out --crit-percent 0.99 \
        --log ${p1}_${p2}_gap1mb.log
    fi
  done
done

# move normalised xpehh files
for i in $(seq 1 22)
do
  mkdir -p norm/chr${i}/gap1mb
  mv chr${i}/*norm norm/chr${i}/
  mv chr${i}/gap1mb/*norm norm/chr${i}/gap1mb/
done

```


Appendix C

Papers published during the course of this thesis

C1 Papers relating to this thesis

Cadzow, M., Merriman, T.R., Dalbeth (2017) N. Performance of gout definitions for genetic epidemiological studies: Analysis of UK Biobank. *Arthritis Research and Therapy*. **19**, 181

Cadzow, M., Merriman, T.R., Boocock, J., Dalbeth, N., Stamp, L.K., Black, M.A., Visscher, P.M., Wilcox, P.L. (2016) Lack of direct evidence for natural selection at the candidate thrifty gene locus, *PPARGC1A*. *BMC Medical Genetics*. **17**, 80

C2 Other papers

Lacey, C.J., Doudney, K., Bridgman, P.G., George, P.M., Mulder, R.T., Zarifeh, J.J., Kimber, B., **Cadzow, M.J.**, Black, M.A., Merriman, T.R., Lehnert, K., Bickley, V.M., Pearson, J.F., Cameron, V.A., Kennedy, M.A. (2018) Copy number variants implicate cardiac function and development pathways in earthquake-induced stress cardiomyopathy. *Scientific Reports*. **8**, 7548

Tanner, C., Boocock, J., Stahl, E.A., Dobbyn, A., Mandal, A.K., **Cadzow, M.**, Phipps-Green, A.J., Topless, R.K., Hindmarsh, J.H., Stamp, L.K., Dalbeth, N., Choi, H.K., Mount, D.B., Merriman, T.R. (2017) Population-Specific Resequencing Associates the ATP-Binding Cassette Subfamily C Member 4 Gene With Gout in New Zealand Māori and Pacific Men. *Arthritis and Rheumatology*. **69**, 1461-1469.

Peters, B., Aidley, J., **Cadzow, M.**, Twell, D., Brownfield, L. (2017) Identification of Cis-regulatory modules that function in the male germline of flowering plants. *Methods in Molecular Biology*. **1669**, 275-293.

Flynn, T.J., **Cadzow, M.**, Dalbeth, N., Jones, P.B., Stamp, L.K., Hindmarsh, J.H., Todd, A.S., Walker,

R.J., Topless, R., Merriman, T.R. (2015) Positive association of tomato consumption with serum urate: Support for tomato consumption as an anecdotal trigger of gout flares. *BMC Musculoskeletal Disorders*, **16**, 196

Dalbeth, N., Topless, R., Flynn, T., **Cadzow, M.**, Bolland, M.J., Merriman, T.R. (2015) Mendelian randomization analysis to examine for a causal effect of urate on bone mineral density. *Journal of Bone and Mineral Research*. **30**, 985-991

Topless, R.K., Flynn, T.J., **Cadzow, M.**, Stamp, L.K., Dalbeth, N., Black, M.A., Merriman, T.R. (2015) Association of *SLC2A9* genotype with phenotypic variability of serum urate in pre-menopausal women. *Frontiers in Genetics*. **6**, 313