THE UNIVERSITY
of EDINBURGH

The Genetics of Multiple Sclerosis in the Northern Isles of Scotland

Catriona Louise Kerr Barnes
BSc, MRes

Doctor of Philosophy in Population Health Sciences

University of Edinburgh

2019

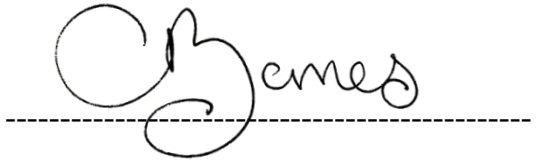
DECLARATION

I confirm that:

i) this thesis has been composed by me;

ii) the work in this thesis is my own, except as indicated below and;

iii) this work has not been submitted for any other degree or professional qualification.



Catriona Louise Kerr Barnes, 13th August 2019

Assistance

Drafting

My supervisors commented on drafts of all thesis chapters and suggested compositional changes and modifications to the figures and tables.

ORCADES and VIKING data

Genotype and phenotype data were provided for ORCADES and VIKING by Jim Wilson, for subsequent analysis in this thesis. This data, including collection procedures, is described in **Chapter 2: Study Data**.

Chapter 3: Heritability of Multiple Sclerosis in the Northern Isles of Scotland

The idea to determine the heritability of Multiple Sclerosis in the Northern Isles was suggested by Jim Wilson. I designed and wrote the analysis plan and computational scripts for analysis and carried out the analysis myself.

Chapter 4: Genome Wide Association Study of Multiple Sclerosis in the Northern Isles of Scotland

The idea to carry out a genome-wide association study was discussed prior to the start of this PhD by Jim Wilson and myself. All work to merge ORCADES and VIKING datasets was carried out by myself. Computational scripts for conducing the GWAS were provided by Peter Joshi, Paul Timmers, David Clark and Andrew Bretherick. These scripts ran all stages listed in the **Chapter 4.2** Methodology section, requiring input from myself on analysis parameters such as traits and covariates to include. The International Multiple Sclerosis Genetics Consortium data was provided by Steffan Daniël Bos-Haugen of the University of Oslo.

Chapter 5: Contribution of common risk variants to Multiple Sclerosis in the Northern Isles of Scotland

The idea to carry out a polygenic risk score was discussed prior to the start of this PhD by Jim Wilson and myself. I designed and wrote the analysis plan and computational scripts for analysis and carried out the analysis myself.

ACKNOWLEDGMENTS

Writing a short line of acknowledgment is not enough for the people and organisations who have helped me over the past four years, however I will try to keep it brief to avoid writing a second thesis of thanks.

Of course, my first and most grateful thanks go to my first supervisor, Professor Jim Wilson. I am so unbelievably lucky to have had you as both a supervisor and friend, and there is absolutely no doubt my life would have turned out very differently had I never met you. Every time I struggled with my PhD, I reminded myself how lucky I was to have this opportunity which you gave me, because you believed I was good enough to be here. Thank you will never be enough.

Secondly, I want to thank my second supervisor, Dr Peter Joshi. Peter, I'm pretty sure that if there was a dictionary entry for "PhD Supervisor" then it would have a photo of you beside it. You have been a fantastic supervisor, and again – I feel ridiculously lucky that I got both yourself and Jim as my supervisory team. Your explanations, feedback, advice and friendship have been immeasurable in helping me on my PhD journey.

This PhD was funded by the Shetland and Orkney Multiple Sclerosis Research Project and the QTL programme funding from the MRC HGU. I cannot thank everyone involved in these projects enough: the charity trustees, Tom and Alma Stove, Stephen Hagan, and everyone who participated to raise money for the funds for this research and those who donated. I am so grateful to the work that went into this funding. Thank you.

I would like to thank the people of Orkney and Shetland: those who gave up their time to contribute to the two fantastic ORCADES and VIKING cohorts; those who helped raise money for Multiple Sclerosis research; those who raise awareness and importance for Multiple Sclerosis research; and finally, those who supported me personally on my PhD journey. My family left Orkney this year after over 12 years on the islands, and I will miss everyone greatly.

Thank you to the staff and students at Edinburgh University, particularly those in the Wilson Group. I could not imagine a more fun, weird and supportive group of people to

work with. You have all made my PhD experience better than I could've imagined it to be, and I have had some of the best days of my life with, and because of, you all.

Finally, I would like to thank my family, friends and those whose presence made my life a little bit brighter. My friends for giving me every emotion under the sun and reminding me that life is an adventure, one which would not exist without your presence. To Benjamin Sisko, William Adama, Buffy Anne Summers, Rod Serling, Fox and Dana. David and Steve. Sturgill Simpson, Joe Dukie, David and Roger, Tom Petty, Stevie Nicks, John Fogerty, the Knopflers, Glenn and Don, Eddie Vedder, Captain EO. To name a few.

My family, who have always supported me in everything I do, particularly my parents – for being people I aspire to be, whether they realise it or not. To Hovan Barnes, who I could not be prouder of. And Moki and Queequeg too, of course.

Lastly, I would like to dedicate this PhD to my Nanna. Who began the PhD journey with myself and Hovan, but left too soon. I know how proud you would be that we both made it.

"Be open to your dreams, people. Embrace that distant shore.
Because our mortal journey is over all too soon."

- Chris Stevens

CONTENTS

Figures

Tables

Appendix Figures

Appendix Tables

ABSTRACT

Multiple Sclerosis affects around 2 million people worldwide (Kantarci and Wingerchuk, 2006; Dutta and Trapp, 2011). The disease is typified by the destruction of the central nervous system neurons' myelin sheaths, caused by the individual's own immune system (Hauser and Oksenberg, 2006). This destruction results in the inflammation and chronic degeneration of the CNS, causing varying symptoms including pain, fatigue, cognitive impairment and paralysis (Costelloe *et al.*, 2008). Not only is the life expectancy of individuals with MS 10 years below that of the age-matched general population (Ragonese *et al.*, 2008), but life quality is often severely affected from the start of disease onset (typically around 30 years of age (Hauser and Oksenberg, 2006)). There are several treatments available to aid in the relief of specific symptoms; however, the treatment is lifelong which places a burden on healthcare services. Current research looks to expand the available knowledge on the causes of MS, to improve preventative measures (such as lifestyle changes to accommodate environmental conditions) and targeted treatments (for example, focusing on the product of an MS-associated gene variant).

Of particular interest to MS research are the two population isolates of Orkney and Shetland, which together make up the Northern Isles of Scotland. Shetland has 295 MS cases per 100,000 individuals, while Orkney has the highest global prevalence of MS at 402 cases per 100,000 individuals (Visser *et al.*, 2012). Orkney, at a lower latitude than Shetland, has a significantly higher prevalence than what would be expected. Multiple theories behind the excess of MS cases in Orkney have been investigated, including vitamin D deficiency and homozygosity: neither were found to cause the high prevalence of MS. It is possible that this excess prevalence may be explained through unique genetics. This thesis sought to better understand these high rates of MS, with the aim of passing this knowledge on to the island residents of Orkney and Shetland and to contribute the findings to the wider understanding of MS.

Analyses were conducted using the ORCADES and VIKING datasets. ORCADES contained 2215 individuals from the Orkney islands, including 97 MS cases (some recruited because they were cases); VIKING contained 2015 individuals from the Shetland islands, including 15 cases. First, a heritability study was conducted using

GCTA to determine the SNP heritability of MS in both Orkney and Shetland and how it compared to published estimates of heritability. The SNP heritability of MS in Orkney was estimated at 0.31 (95% CI 0.13, 0.49). An estimate of SNP heritability for MS in Shetland could not be determined due to low case numbers.
Second, a genome-wide association study was conducted using a combined ORCADES/VIKING dataset containing 112 cases and 4223 controls. The aim of this study was to determine if unique common MS risk variants existed in the Northern Isles. Here, 89 SNPs were identified to suggestive significance, mostly within six key regions of the genome. Within the literature, only one of these (chromosome 6 SNP rs9268154) was associated with Multiple Sclerosis. Four of the five other regions had possible functions within the immune or nervous system. However, as these did not reach genome-wide significance it is likely these results were due to chance; further investigation is needed to clarify this.

Finally, a polygenic risk score study looked at the contribution of common risk variants to MS. The 127 most strongly associated MS SNPs were used to calculate risk scores in mainland Scotland, Orkney and Shetland. These risk scores were compared between controls in all three populations to determine if the Northern Isles, by chance, had higher frequencies of common risk variants and if this contributed to the excess of cases. These common risk variants explained 3% of variance in MS risk, and had an AUC score of 0.69 (95% CI 0.65, 0.74). However, no difference existed between common risk variants in the three populations, aside from one variant: rs9271069, a tag SNP for *HLA-DRB1*1501*. This SNP was found to have a significantly higher frequency in Orkney (RAF = 0.23, p-value = 8×10^{-13}) and Shetland (RAF = 0.21, p-value = 2.3×10^{-6}) than mainland Scotland (RAF = 0.17). This SNP accounted for 6 cases (95% CI 3, 8) out of 150 observed excess cases per 100,000 individuals in Shetland and 9 cases (95% CI 8, 11) of the observed 257 excess cases per 100,000 individuals in Orkney.

The question of why the Northern Isles have such high rates of MS remains open. This thesis explains a small proportion of this excess. It is hoped that the findings and discussions found here will encourage dialogue within the Northern Isles and bring awareness to the genetic, environmental and lifestyle factors that contribute to MS within the islands.

LAY SUMMARY

Multiple Sclerosis is a disease that affects the brain and spinal cord, resulting in pain, muscle problems, difficulties with thinking and paralysis. In a healthy individual, the immune system works to protect the body from infections and toxins. In an individual with MS, the body's immune system attacks itself. It is not fully understood why this happens, although environmental and genetic factors are both thought to influence the disease.

Of particular interest to MS research are the Northern Isles of Scotland, Orkney and Shetland. Orkney has the highest rate of MS in the world, with 402 MS cases per 100,000 people. Shetland has a similarly high rate of MS, with 295 MS cases per 100,000 people. The number of people affected with MS in the Northern Isles is higher than expected.

So far, studies investigating the rates of MS in the islands have not found the reason for the high burden of disease. This thesis looked to better understand why the rates of MS here are so high, and to pass that knowledge on to the island residents of Orkney and Shetland and to contribute the findings to the wider understanding of MS.

First, I looked at how much variation in risk of developing MS was caused by genetics (as opposed to the environment). I found that around one third of differences in the risk of developing MS in Orkney was due to differences in genetics between individuals; this is similar to previous studies of MS. This means that although genes are important in determining if you are likely to get MS in the Northern Isles, environmental factors also play an important role. I was unable to find an answer to this question for Shetland, as there were too few MS cases to answer this question accurately.

Second, I looked to see if there were any unique genetic variants which exist commonly among the populations in the Northern Isles that contributed to MS risk, however it does not appear that such a variant is likely to exist.

Third, I looked at the combined effect of genetic risk variants commonly found in the population. The combination of risk variants each individual has is unique, but it can be summarised into a personalised genetic risk score. When I compared the average

genetic risk score for people in Orkney and Shetland to the average genetic risk score for people in mainland Scotland, I found no difference between the populations.

Finally, I found the genetic risk variant which has the strongest effect on MS risk is more commonly found in the Northern Isles then mainland Scotland. However, this one genetic variant only contributes around 4% to the surplus number of cases that are present in the Northern Isles. This suggests that other factors, either genetic, environmental, or both, are causing the high rates of MS in the Northern Isles.

ABBREVIATIONS

1,25(OH)2D	-	1,25-dihydroxyvitamin D
ABCA1	-	Adenosine triphosphate binding cassette transporter 1
AD	-	Allelic dosage
AI	-	Artificial intelligence
AUC	-	Area under the curve
BMI	-	Body mass index
CD	-	Cluster of differentiation
CDCV	-	Common disease common variant
CI	-	Confidence interval
CIS	-	Clinically isolated syndrome
CLEC	-	C-Type lectin
cM	-	Centimorgan
CNP	-	Copy number polymorphism
CNS	-	Central nervous system
CNV	-	Copy number variation
CSF	-	Cerebrospinal fluid
CYP	-	Cytochrome P450
DDX	-	DExD-Box Helicase
DEXA	-	Dual energy X-ray absorptiometry
DHCR7	-	7-Dehydrocholesterol reductase
DMT	-	Disease modifying therapies
DNA	-	Deoxyribonucleic acid
EAE	-	Experimental autoimmune encephalomyelitis
EBNA	-	Epstein-Barr virus nuclear antigen
EBV	-	Epstein-Barr virus
EDSS	-	Expanded disability status scale
FDA	-	Food and Drug Administration
FDR	-	False discovery rate
GCTA	-	Genome-wide Complex Trait Analysis
GluN2B	-	Glutamate ionotropic receptor NMDA type subunit 2B
GRAMMAR	-	Genome-wide Rapid Association Using Mixed Model and Regression
Q-Q	-	Quantile-quantile
GREML	-	Genome-based restricted maximum likelihood
GRIN2B	-	Glutamate Receptor Ionotropic NMDA subunit 2B
GRM	-	Genetic relationship matrix
GWAS	-	Genome wide association study
GxE	-	Gene-environment interactions
GxG	-	Gene-gene interactions
H_2	-	Broad sense heritability
h_2	-	Narrow sense heritability

h_g^2	-	SNP heritability
h_{gwas}^2	-	GWAS heritability
HLA	-	Human leukocyte antigen
HRC	-	Haplotype Reference Consortium
HWE	-	Hardy Weinberg Equilibrium
IBS	-	Identical by state
IFN	-	Interferon
IgG	-	Immunoglobulin G
IL	-	Interleukin
IMSGC	-	International Multiple Sclerosis Genetics Consortium
kb	-	kilobase
LD	-	Linkage disequilibrium
LDLR	-	low density lipoprotein receptor
LMM	-	Linear mixed model
MAF	-	Minor allele frequency
Mb	-	Megabase
MHC	-	Major histocompatibility complex
MIR924HG	-	Long Intergenic Non-Protein Coding RNA 669
MALT1	-	Mucosa associated lymphoid tissue lymphoma translocation gene 1
GS	-	Generation Scotland
MR	-	Mendelian randomisation
MRI	-	Magnetic resonance imaging
MS	-	Multiple Sclerosis
n	-	Number
NADSYN	-	Nicotinamide adenine dinucleotide Synthetase
NAT	-	N-acetyltransferase
NIMS	-	Northern Isles Multiple Sclerosis Study
NLRP	-	Nucleotide-binding oligomerization domain-like receptor family pyrin domain containing
NMDA	-	N-methyl-D-aspartate
CTAGE1	-	Cutaneous T Cell Lymphoma-Associated Antigen 1
NR	-	Nuclear Receptor
OPC	-	Oligodendrocyte precursor cells
OR	-	Odds ratio
ORCADES	-	Orkney Complex Disease Study
PC	-	Principal component
PCA	-	Principal component analysis
PGRS	-	Polygenic risk score
PPMS	-	Primary progressive Multiple Sclerosis
pT	-	p-value threshold
QC	-	Quality control
QTL	-	Quantitative trait loci

r_2	-	Squared coefficient of correlation
LINC	-	Long Intergenic Non-Protein Coding RNA
RAF	-	Risk allele frequency
RNA	-	Ribonucleic acid
ROC	-	Receiver operating characteristic
IBD	-	Identical by descent
RRMS	-	Relapsing-remitting Multiple Sclerosis
RSID	-	Reference SNP cluster identification
SE	-	Standard error
SNP	-	Single nucleotide polymorphism
SPMS	-	Secondary progressive Multiple Sclerosis
TNF	-	Tumour necrosis factor
TYK	-	Tyrosine Kinase
UVB	-	Ultraviolet B
UVR	-	Ultraviolet radiation
VDRE	-	Vitamin D response element
VIKING	-	Viking Health Study – Shetland

CHAPTER 1: INTRODUCTION

1.1 Multiple Sclerosis

1.1.1 Introduction and Epidemiology

Multiple Sclerosis (MS) is the most common neurological disability found in young adults in the Western world (Tremlett and Rieckmann, 2010). The disease is characterised by the inflammation and chronic degeneration of the central nervous system (CNS), a result of the destruction of the myelin sheath surrounding CNS neurons by the individual's immune system (Hauser and Oksenberg, 2006). There is no definitive explanation as to the reason for these immune attacks, with genetic susceptibility and environmental factors both contributing to MS risk (Sotgiu *et al.*, 2004).

The manifestations of MS's clinical symptoms are unpredictable (Costelloe *et al.*, 2008). As the immune system causes the destruction of the myelin sheath, lesions (or scar tissue) form in the damaged regions, interrupting nerve impulses travelling through the CNS (Hauser and Oksenberg, 2006). The location of these regions of damaged myelin can result in varying symptoms that are experienced differently in each patient (Miller *et al.*, 2008). The mean age of onset of the disease is 30; within young adults, it is the most common reason for diagnosis with a non-traumatic neurological disability (Hauser and Oksenberg, 2006). Although MS has no unique clinical symptoms, some symptoms are very distinctive of the disease (Miller *et al.*, 2008). Among the most common are fatigue, vision impairment, poor balance, pain, paralysis and cognitive impairment (Miller *et al.*, 2008). The life expectancy of MS sufferers is also shorter, and currently sits at 10 years below the age-matched general population life expectancy (Brønnum-Hansen, Koch-Henriksen and Stenager, 2004; Grytten Torkildsen *et al.*, 2008; Ragonese *et al.*, 2008).

Symptoms alone cannot be used to diagnose MS; it is essential that evidence of damage in at least two distinct regions of the CNS occurring at different time periods is found (Polman *et al.*, 2011). It is also important to rule out other conditions which show similar neurological symptoms to MS, for example: CNS infections (also found in Lyme disease), CNS inflammatory disorders (systemic lupus erythematosus), structural CNS

damage (herniated disc) genetic disorders (hereditary myelopathies), brain tumours (lymphoma), deficiencies (copper or vitamin B12) and other non-MS demyelinating disorders (neuromyelitis optica) (Magro Checa *et al.*, 2013). Ruling these conditions out requires additional tests above the assessment of a patient's medical history and sensory functions, and magnetic resonance imaging (MRI), cerebrospinal fluid (CSF) evaluation and evoked potentials can be used to make a formal diagnosis (Calabresi, 2004). MRI scans measure relative water content within tissues; regions in the brain where the water-repellent myelin has been degraded hold more water, and so can be visibly identified on scans (Polman *et al.*, 2011). Cerebrospinal fluid, which immerses the brain and spinal cord, can be collected via a lumbar puncture and checked for elevated IgG antibodies and oligoclonal bands, whose presence indicate an irregular immune response (Polman *et al.*, 2011). Evoked potentials measure brain electrical activity in response to stimulation and can identify slower electrical transmission caused by demyelination (Polman *et al.*, 2011). Investigating these criteria in potential MS patients often rules out or diagnoses MS (Polman *et al.*, 2011).

The burden of MS is considerable, with approximately 2 million people worldwide affected (Kantarci and Wingerchuk, 2006; Dutta and Trapp, 2011). In general, countries closer to the equator have a lower prevalence of MS, with prevalence increasing as you move further north or south away from the equator (Simpson *et al.*, 2011). There is a significant association between latitude and MS that persists after adjustment for the most strongly associated MS risk allele*, HLA-DRB1* (Simpson *et al.*, 2011), which suggests a strong role of environmental influences that change with latitude (Simpson *et al.*, 2011).

Globally, some of the lowest prevalence rates are found within Africa, where prevalence ranges from only occasional cases in sub-Saharan Africa (Poser, 1994) to 10 per 100,000 in Tunisia (Poser, 1994) and 13 per 100,000 in English-speaking white South Africans (Dean, 1967). A low prevalence of <10 per 100,000 is seen in India, China and Japan (Wadia and Bhatia, 1990).

Higher prevalence rates are found in North America, which has an overall prevalence of 149 per 100,000 (Dilokthornsakul *et al.*, 2016). Canada has some of the highest rates of MS globally, at 380 per 100,000 (Amankwah *et al.*, 2017).

Prevalence in Europe is considerably more varied, with the highest rates found in Scotland and Nordic countries (Kingwell *et al.*, 2013). Scotland has an overall prevalence <143 per 100,000 (Kingwell *et al.*, 2013), with island populations such as Orkney and Shetland having substantially higher prevalence rates (402 and 295 per 100,000 respectively) (Visser *et al.*, 2012). Within the Nordic countries, prevalence is typically over 75 per 100,000, with Sweden having the highest prevalence of 253 per 100,000 (Kingwell *et al.*, 2013).

In the Southern hemisphere, New Zealand has prevalence rates that range from 23.6 to 68.5 per 100,000 (Alla *et al.*, 2014), although the indigenous Maori population is lower at 24.2 per 100,000 (Taylor *et al.*, 2010). Australia has a higher prevalence of 95.5 per 100,000 (Palmer *et al.*, 2013), although this does not include Aboriginal Australians. A 2017 study focusing on MS prevalence in Australia failed to identify any Aboriginal Australian cases, however they only used data from two out of the 10 Australian territories so further research is needed here (McLeod, Hammond and Hallpike, 1994).

Australia and New Zealand are not the only countries that have groups of individuals with significantly different prevalence rates from the general population; there are several regions or ethnic groups within countries that have unusually high or low prevalence to that expected based on the latitude and prevalence of surrounding populations. For example, although Canada has one of the highest MS prevalence rates globally, Native Aboriginals in Manitoba have a prevalence of 40 per 100,000 (Rivera and Cabrera, 2001) and First Nations populations in Alberta have a prevalence of 99.9 per 100,000, lower than the population average (Marrie, Hall and Sadovnick, 2016). The Inuit are known to have a much lower prevalence of MS, although no official prevalence data has been recorded (Chan, 1977). The Native Norwegian Sami also have a lower than average prevalence of MS of 30 per 100,000 (Lincoln *et al.*, 2009). Conversely, populations such as the Italian island of Sardinia have a much higher prevalence of MS than expected at 247.6 per 100,000 (Pugliatti, Sotgiu and Rosati, 2002; Sotgiu *et al.*, 2004). Within the broadly low-prevalence India, some comparatively high pockets of MS prevalence exist: Parsis in Poona, India, who migrated to India in the seventeenth century from Persia have prevalence rates of 58 per 100,000 (Rosati, 2001). Thus, although MS varies by latitude, the population

disparities within regions show that other factors such as genetic differences or behavioural-cultural distinctions influence MS risk.

Woman are more likely than men to develop MS. The overall incidence rate (the number of new cases of MS per year) for MS in Europe, North America, Australia and New Zealand is 3.6 and 2.0 cases per 100,000 for women and men respectively (Alonso and Hernan, 2008), with the female to male ratio for developing Multiple Sclerosis standing at 2:1 (World Health Organization, 2008; Browne *et al.*, 2014). However, men and those with the time of onset occurring at a later age have a worse prognosis and faster progression than women and younger individuals diagnosed (Weinshenker *et al.*, 1991; Confavreux, Vukusic and Adeleine, 2003; Tremlett, Paty and Devonshire, 2006; Debouverie *et al.*, 2008). The reason for women having higher rates of MS is not well understood; however, one theory suggests that women have higher levels of a blood vessel receptor protein S1PR2. S1PR2 determines if immune cells successfully cross the blood-brain barrier, and they have been found at increased numbers in areas of the brain which have been damaged by MS (Cruz-Orengo *et al.*, 2014).

1.1.2 Clinical Course

MS often first presents itself though a clinically isolated syndrome (CIS) (Efendi, 2015). The CIS will show features of inflammatory demyelination, which include optic neuritis (when the optic nerve becomes inflamed), brainstem syndromes (a grouping term for multiple conditions which affect the brainstem which can cause symptoms including pain, paralysis and sensation impairment) or transverse myelitis (inflammation of the spinal cord) (Efendi, 2015). In children, a CIS can appear as symptoms of encephalopathy (for example, vomiting, headache or seizure) (Efendi, 2015). It is not until later in the disease course that the type of MS becomes apparent. Three main types of MS exist: relapsing-remitting MS (RRMS), primary-progressive MS (PPMS) and secondary progressive MS (SPMS) (Lublin *et al.*, 2014).

Relapsing-remitting MS

Between 80-90% of patients with MS are diagnosed as relapsing-remitting MS (RRMS), where multiple periods of relapse and remission are experienced (Koch *et al.*, 2008; Tremlett and Devonshire, 2008).

Relapses, a prominent clinical feature of MS, are defined as the emergence of novel symptoms, or the reappearance of earlier symptoms, for a minimum period of 24 hours (Scalfari *et al.*, 2010). Relapses can lead to a temporary or even permanent loss of anatomical function (Scalfari *et al.*, 2010). More than 80% of individuals who experience an initial relapse will experience secondary disability progression (Weinshenker *et al.*, 1989; Lublin *et al.*, 2014) and early relapse frequency can indicate a faster clinical course (Scalfari *et al.*, 2010). Evidence has suggested that relapses are the external expression of recurrent inflammation within the central nervous system (Youl *et al.*, 1991) with an average of 10 novel MRI lesions detected for each relapse (McDonald, Miller and Thompson, 1994).

In contrast, remission is a period following a relapse where there no new signs of disease activity occur (Tsang and Macdonell, 2011). The length of a period of remission can vary between individuals, but can last for months to years before another relapse is experienced (Tsang and Macdonell, 2011). A relapse before a remission may leave problems in around 40% of patients, with the probability of experiencing difficulties increasing in individuals who have had MS for a longer period (Tsang and Macdonell, 2011). A small subgroup of individuals with RRMS may fully resolve relapse symptoms between attacks; in these cases, they are referred to as having benign MS (Sayao, Devonshire and Tremlett, 2007; Costelloe *et al.*, 2008). Benign MS symptoms do not progress past moderate disability in one functional system (Expanded Disability Status Scale stage 3) after 10 years following disease onset (Sayao, Devonshire and Tremlett, 2007; Costelloe *et al.*, 2008). From the previous definition, it is possible to group approximately 30% of MS patients as benign (Pittock *et al.*, 2004). Conversely, someone is described as having malignant MS if significant disability appears early on in disease progression (Gholipour *et al.*, 2011). This form of disease is unresponsive to established treatment methods, although high-dose chemotherapy combined with anti-thymocyte globulin and an autologous stem cell transplantation have proven effective in instigating neurological improvement and improvement on the Expanded Disability Status Scale (EDSS) (Kimiskidis *et al.*, 2008).

The clinical processes of relapses/remissions and chronic worsening are due to specific biological mechanisms that can be targeted by drugs to relieve symptoms (Scalfari *et al.*, 2010). For example, the drug alemtuzumab is a monoclonal antibody that binds to an

antigen expressed on B and T lymphocytes called CD52. By binding to this antigen, circulating lymphocytes are depleted and symptoms of relapses are reduced (NICE, 2014). However, although this drug reduces the number of relapses by around 70%, it does not delay disease progression (Coles *et al.*, 1999). Drugs developed for MS tend to target specific points in the disease pathway, and so there is not one overall treatment that prevents MS in its entirety.

Primary progressive MS

Primary progressive MS (PPMS) is diagnosed in 10-20% of individuals with MS (Ebers, 2004). Unlike RRMS, symptoms worsen consistently over time, relapses are not present and there are few, if any, remissions or improvements in symptoms (Scalfari *et al.*, 2010). Primary-progressive MS is also the form of MS which most often affects late onset individuals (where diagnosis occurs after the age of 50) with 55-80% of these individuals being diagnosed with this form of MS (Tremlett and Rieckmann, 2010). While relapses are considered to be a symptom of CNS inflammation, consistent progressive MS is considered to be a symptom of early, chronic axonal loss (Evangelou *et al.*, 2000; Filippi *et al.*, 2003). A PPMS subtype, progressive-relapsing MS (Scalfari *et al.*, 2010) exists; this follows the initial pattern of primary-progressive MS, although the patient will go on to experience relapses (Tremlett and Rieckmann, 2010). This is the least common form of MS, with approximately 5% of cases being diagnosed with progressive-relapsing MS (Tremlett and Rieckmann, 2010).

Secondary progressive MS

Secondary progressive MS (SPMS) begins with relapses and remission of symptoms and follows on to a progressive phase with no definite period of remission, which may or may not have superimposed relapses (Scalfari *et al.*, 2010). Approximately 65% of those who are initially diagnosed with RRMS will develop SPMS (Scalfari *et al.*, 2010). Relapses tend to occur less as progression increases, with short remissions or plateaus occurring in some cases (Lublin and Reingold, 1996). However, as milestones of disease progression differ between patients it is often difficult to confirm a diagnosis due to the high levels of variation (Runmarker and Andersen, 1993; Koch *et al.*, 2008).

MS Treatment

Although no cure for MS currently exists, there are several treatment methods developed for specific types and stages of MS, with the earlier stages of MS targeted

with most success (Rizvi and Agius, 2004). The most current effective treatment for MS are disease-modifying therapies (DMT), which can slow the progression of MS by reducing the frequency and severity of relapses and novel lesions (Wingerchuk and Carter, 2014). Fifteen DMTs are currently approved by the US Food and Drug Administration (FDA) to treat RRMS, PPMS and SPMS (Wingerchuk and Carter, 2014). The most recent of these is ocrelizumab, which slows the progression of PPMS by decreasing the number of B cells which are CD20-positive (Corboy and Miravalle, 2010). Compared with interferon beta-1a, a previously developed MS drug, ocrelizumab reduced relapse rates by up to 47% and decreased brain lesions by 95% (Corboy and Miravalle, 2010).

Drugs that have recently been developed or are in the final stages of clinical trials focus on regulating the activation of immune cells (for example, alemtuzumab) (Azzopardi, Coles and Sklerozda Alemtuzumab, 2011), preventing lymphocytes from departing from secondary lymphoid organs (fingolimod and natalizumab, respectively) (Polman *et al.*, 2006; Kappos *et al.*, 2010) or suppressing CNS inflammation (laquinimod) (Comi *et al.*, 2012). Laquinimod works to treat PPMS by reducing the numbers of particular cytokines and preventing immune cells from reaching the brain (Comi *et al.*, 2012). Another drug currently in development is MD1003, a highly concentrated form of biotin or vitamin B7 (10,000 times the recommended daily allowance) (Corboy and Miravalle, 2010). This helps to promote the repair of the myelin sheath surrounding CNS neurons (Corboy and Miravalle, 2010).

Autologous haematopoietic stem cell transplantation is currently being investigated as a potential solution for MS treatment: haematopoietic (or blood cell producing) stem cells are collected from the patient's bone marrow and stored (Muraro *et al.*, 2017). The patient undergoes aggressive chemotherapy to deplete their immune system, and the collected stem cells are used to rebuild the patient's immune system over 3-6 months (Muraro *et al.*, 2017). Although this treatment may halt disease progression for 5 years in 46% of MS patients, there is a significant risk involved in using chemotherapy (Muraro *et al.*, 2017).

1.1.3 Pathophysiology

Basic cell types of the immune system

There are multiple cell types involved in MS development and progression. Among these are T and B cells, lymphocytes that are major components in the adaptive immune response (Sawcer *et al.*, 2011). Many of the key cells which play a role in MS can be seen in Figure 1, which shows the process of haematopoiesis and the resulting lineages of blood cells.

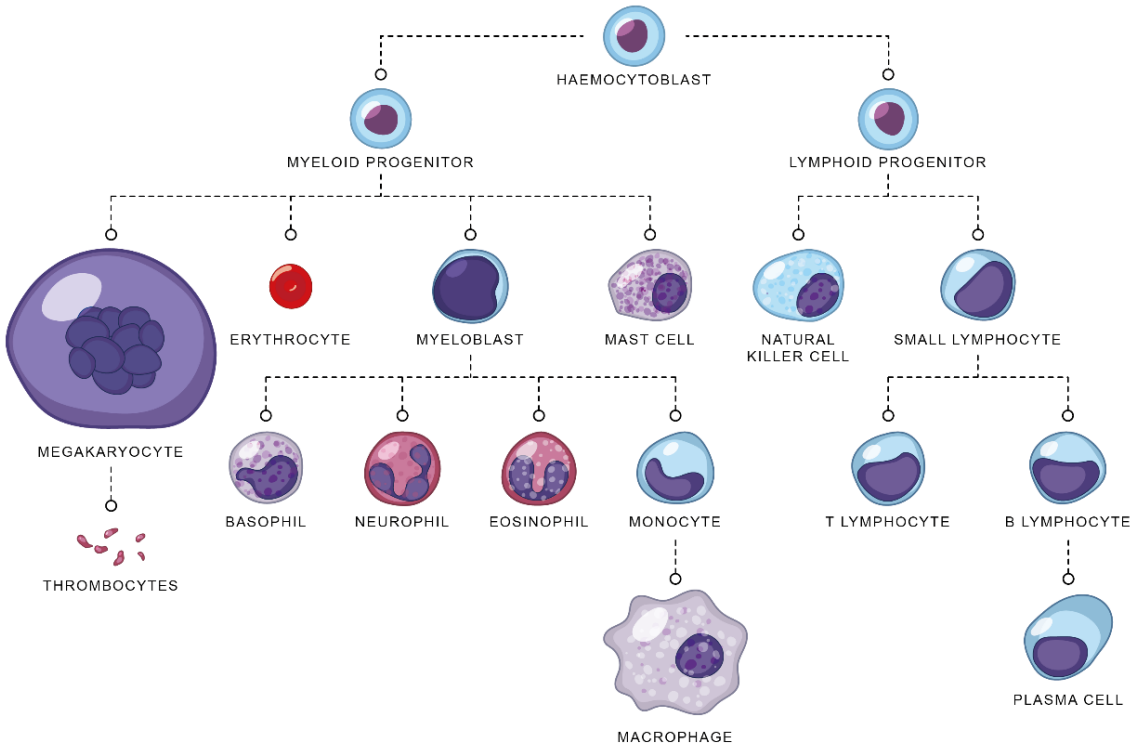


Figure 1: Haematopoiesis diagram

A diagram illustrating the various blood cell types produced during the process of haematopoiesis, beginning with haematopoietic stem cells (haemocytoblast) and ending in mature blood cell types.

T cells play a key role in cell-mediated immunity, a part of the immune system which also involves the activation of phagocytes (cells that ingest foreign material or dying cells) and release of cytokines (Mosmann and Sad, 1996). There are multiple types of T cells, including CD4+ T Cells (also known as T helper cells), CD8+ T cells (or cytotoxic T

cells) and regulatory T cells (Mosmann and Sad, 1996). CD4+ T cells become activated when presented with peptide antigens; small molecules from a toxin, foreign substance or even part of the host that can produce an immune response (Mosmann and Sad, 1996). Peptide antigens are displayed from MHC class II molecules on antigen presenting cells such as dendritic cells or B cells (Mangalam, Rodriguez and David, 1994). Once the CD4+ T cells detect an antigen and become activated, they divide and secrete various cytokines (Mangalam, Rodriguez and David, 1994). CD8+ T cells become activated when presented with antigens from MHC class I molecules, and proceed to destroy virus-infected cells, tumour cells or other damaged cells (Fletcher *et al.*, 2010). Regulatory T cells then secrete several molecules, including IL-10, to inactivate the CD8+ T cells (Fletcher *et al.*, 2010). Another important role of regulatory T cells is to suppress autoreactive T cells which have escaped the thymus; if left unsuppressed, autoreactive T cells can cause damage to other cell types, which is seen in MS (Fletcher *et al.*, 2010).

B cells, unlike T cells, are involved in humoral immunity, whereby they secrete antibodies and cytokines and present antigens (Høglund and Maghazachi, 2014). Similar to T cells, there are multiple types of B cells such as plasma B cells, which secrete antibodies such as immunoglobulin G1 (which is detected in the CSF of as many as 95% of diagnosed MS patients) (Høglund and Maghazachi, 2014).

Monocytes, another leukocyte, are part of the innate immune system but also influence the adaptive immune system (Mallucci *et al.*, 2015). This type of cell travels from the bloodstream to tissues around the body where they differentiate into macrophages (large phagocytes) or dendritic cells (which present antigens to T cells) (Mallucci *et al.*, 2015).

These cell types are involved in the two principal theories for the cellular processes underlying Multiple Sclerosis: the CNS-extrinsic model and the CNS-intrinsic model (Høglund and Maghazachi, 2014).

CNS-extrinsic model

Within the CNS-extrinsic model, also referred to as the peripheral model, MS is thought to be triggered in peripheral sites to the CNS rather than within the CNS itself. A diagram of this model can be seen in Figure 2.

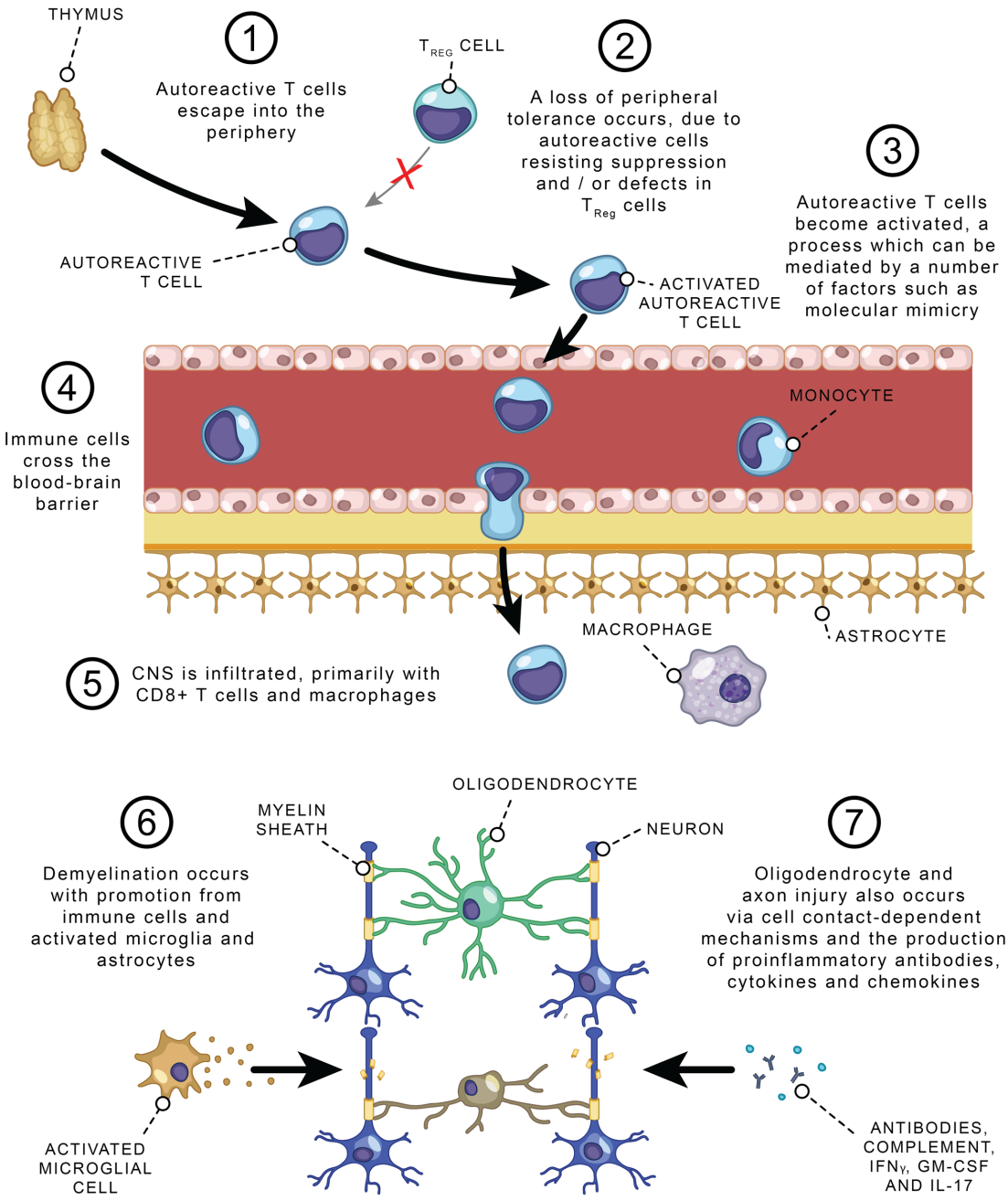


Figure 2: A diagram outline the CNS-extrinsic model for Multiple Sclerosis

This diagram outlines the CNS-extrinsic model for Multiple Sclerosis, with key stages highlighted from the escape of autoreactive T cells from the thymus to onset of attack within the CNS.

In the CNS-extrinsic model, autoreactive T cells escape from the thymus into the periphery (Vizier *et al.*, 1999) and fail to be suppressed (likely due to a defect in T_{reg} cell function or a resistance to suppression (Friese and Fugger, 2005)). These autoreactive T cells then become activated, possibly through bystander signals from cytokines or molecular mimicry, where self-antigens are confused for foreign antigens (Friese and Fugger, 2005). The activated autoreactive T cells, along with B cells and monocytes, then travel to the CNS. At this stage, they cross the blood-brain barrier and mediate damage against the central neurons, principally to the myelin sheaths and axons. The destruction of the myelin sheath involves both direct cell contact and the production of damaging antibodies, cytokines and chemokines (Dendrou, Fugger and Friese, 2015).

CNS-intrinsic model

Unlike the CNS-extrinsic model, in the intrinsic model the events that trigger MS take place within the CNS with resident CNS cells (Høglund and Maghazachi, 2014). The events that would lead to the development of neurodegeneration in this model are not clear, although theories suggest that the inflammatory response may occur in response to a viral infection (Høglund and Maghazachi, 2014). Autoreactive lymphocytes would then appear in the CNS as a secondary incident (Høglund and Maghazachi, 2014).

It is more likely the pathology of Multiple Sclerosis follows that of the CNS-extrinsic model. The CNS-extrinsic model is more consistent with the procedure of inducing experimental autoimmune encephalomyelitis (EAE) in animals, where pathogenic CD4+ T cells are artificially generated and proceed to cross the blood-brain barrier to cause an immune response against the CNS (Kipp *et al.*, 2017).

Remyelination

Remyelination can often occur following the destruction of the myelin sheath as a repair tactic, which can help to prevent exposed axons from degeneration (Huebner and Strittmatter, 2009). CNS axons do not spontaneously regenerate and so remyelination is an important process to slow MS progression through improving the functionality of the CNS neurons and potentially protecting axons (Huebner and Strittmatter, 2009). Remyelination in humans is a highly variable process, which can occur often in some individuals (20% of individuals in a study remyelinated 60% of their lesions) but not in others (Patrikios *et al.*, 2006). It can occur when oligodendrocyte precursor cells (OPCs), a type of non-neuronal cell within the CNS, undergo fast proliferation and

differentiation into oligodendrocytes (Patrikios *et al.*, 2006). Oligodendrocytes are able to form the myelin sheath surrounding the CNS neurons (Patrikios *et al.*, 2006). This process of differentiation of OPCs into oligodendrocytes is aided by the removal of myelin debris by macrophages, as this debris contains OPC differentiation inhibitors (Levine, Reynolds and Fawcett, 2001). However, remyelination can fail due to differentiation failure or OPC recruitment failure (Levine, Reynolds and Fawcett, 2001), leaving CNS axons vulnerable to axon degeneration.

1.1.4 Environmental factors in Multiple Sclerosis

There is no one single factor that fully contributes to developing MS, however there is strong evidence that both environmental and genetic factors influence the disease (Ramagopalan, Giovannoni, *et al.*, 2009), with MS arising most frequently in genetically susceptible individuals that may have been exposed to risk-associated environmental factors or stochastic events (Goodin, 2009).

To investigate the influence of genetic factors versus environmental factors, twin studies can be used. These are valuable studies which examine disease prevalence among monozygotic twins (identical, sharing 100% of genes) and dizygotic twins (fraternal, sharing 50% of genes) (Sadovnick *et al.*, 2004). For MS in monozygotic twins, there is a concordance rate (where both twins develop MS) of 20-40% (Sadovnick *et al.*, 2004); therefore, monozygotic twins are more likely to be discordant for MS. This indicates that environmental factors are influencing MS development, as a high genetic influence would result in a higher concordance rate between identical twins (Sadovnick *et al.*, 2004). The concordance rate between dizygotic twins is much lower, estimated at 3-5% (Mumford *et al.*, 1994). The higher concordance rate between monozygotic than dizygotic twins indicates a significant genetic influence for MS, as if the disease were highly influenced by environmental factors, the concordance rate between dizygotic and monozygotic twins would be very similar (Mumford *et al.*, 1994).

The influence of environmental factors in MS development has been showed through migration studies of MS (Compston and Coles, 2008). An individual who resides in a low-risk region but migrates to a high-risk region during adulthood will retain their low risk of developing MS (Gale and Martyn, 1995). However, an individual who migrates

from a low-risk region to a high-risk region during childhood (under 15 years of age) will acquire the high risk of their new host country (Gale and Martyn, 1995). This has been seen across multiple studies in several population groups, for example, the United Kingdom and Australia (Barnett *et al.*, 2016), France and the French West Indies (Cabre, 2007) and European and South Africa (Kurtzke, Dean and Botha, 1970; Kurtzke, Delasnerie-Lauprêtre and Wallin, 1998). This indicates the importance of environmental effects on developing MS, particularly before the age of 15.

The prevalence and incidence of MS varies by geographic region, in particular an individual's latitude. As mentioned in **Section 1.1.1**, the prevalence of MS is lowest around the equator, with prevalence rates increasing as you move further north or south away from the equator line (Simpson *et al.*, 2011). A meta-analysis of 321 peer-reviewed studies of MS prevalence confirmed this latitudinal gradient as statistically significant ($p<0.001$) (Simpson *et al.*, 2011). Furthermore, this gradient remained when adjusting for frequencies of the most strongly associated MS genetic variant, *HLA-DRB1* (Simpson *et al.*, 2011). This indicates that a prominent environmental factor that varies by geographic latitude affects the development of MS: likely candidates for this are ultraviolet radiation (UVR) and vitamin D (Acheson, Bachrach and Wright, 1960). UVB is the type of UVR thought to be influential to MS risk. Although UVB is not the only solar radiation emitted from the sun, it is considered biologically important for MS in particular, as UVB begins vitamin D synthesis within the skin (Lucas *et al.*, 2015). Individuals further north and south of the equator will have lower exposures of UVB, and will therefore produce less vitamin D (Lucas *et al.*, 2015). However, it is challenging to disentangle the effect of sun exposure from vitamin D on MS, given the close connection between them (Lucas *et al.*, 2015). Many studies consider vitamin D and sun exposure as proxies for one another. For example, past sun exposure was used as a proxy for past vitamin D status, and a low past sun exposure was shown to be associated with an increased risk of developing MS (Handel and Ramagopalan, 2012). Lower sun exposure has also been connected to more severe forms of MS, however this may be due to lower vitamin D status (Martinelli *et al.*, 2014).

There is evidence that vitamin D could affect MS independently. A Mendelian randomisation (MR) study showed that genetically lowered levels of 25-hydroxyvitamin D has been shown to be strongly associated with an increased risk of MS (Mokry *et al.*,

2015). In particular, increased vitamin D levels before the age of 20 were shown to decrease MS risk in later life (Soilu-Hänninen *et al.*, 2005). Additionally, *CYP27B1* (which codes for an enzyme that converts 25-hydroxyvitamin D to the active form of vitamin D) and *CYP24A1* (which breaks down vitamin D metabolites) are associated with MS risk (Sawcer *et al.*, 2011). A low frequency variant (carried at 5% frequency in the population) in *CYP2R1* had a large effect on 25-hydroxyvitamin D levels, with heterozygote carriers having an increased risk of vitamin D insufficiency with an odds ratio (OR) of 2.2 (Manousaki *et al.*, 2017). Individuals with one copy of this variant had increased odds of developing MS (OR = 1.4) (Manousaki *et al.*, 2017). A vitamin D response element has also been found within the promoter region of *HLA-DRB1*1501*, the most strongly associated MS risk variant (Lucas *et al.*, 2015).

Although there has also been evidence to show that higher dietary intake of vitamin D is protective of MS (Mirzaei *et al.*, 2011; Bäärnhielm, Olsson and Alfredsson, 2014), vitamin D supplementation trials have only shown improvements in some aspects of the patients' immune system and MRI scans, with little clinical benefit observed (Jagannath *et al.*, 2010; Hewer *et al.*, 2013).

Other environmental factors thought to influence MS susceptibility include viral exposure, particularly Epstein-Barr virus (although human herpes virus 6 and canine distemper virus have been investigated in the past (Holmøy and Hestvik, 2008)). Although there is not a particular risk for an individual to develop MS if they have at some point experienced an EBV infection (Holmøy and Hestvik, 2008), individuals who have been seronegative for EBV had almost no risk of developing MS when compared to seropositive individuals (Ascherio and Munger, 2007). However, there appears to be a critical time period for EBV, with infection during adolescence and onwards being important for MS susceptibility (Makhani *et al.*, 2016).

Tobacco smoking has been shown to substantially increase the risk of MS, with smokers 1.8 times more likely to develop MS than non-smokers (Riise, Nortvedt and Ascherio, 2003). The risk of developing secondary progression is also 3.6-fold higher in smokers than non-smokers (Hernán *et al.*, 2005), emphasising the importance of individuals with MS to stop smoking as environmental factors affect disease progression as well as risk (Brey, 2003). Smoking triggers lung inflammation and supports proinflammatory pathways (Shan *et al.*, 2009). It is possible that if CNS autoantigenic cells (cells that

target self-antigens) exist in the lung, they may become activated to attack the CNS (Odoardi *et al.*, 2012).

A higher body mass index (BMI) has been linked to an increased risk of MS, where a Mendelian Randomisation study found that 1 standard deviation increase in BMI was associated with an OR of 1.41 for MS risk (Mokry *et al.*, 2016). This may be because obesity is broadly a low-grade inflammation which causes proinflammatory mediators to be produced within fat tissues (which could potentially trigger neuroinflammatory activity) (Lumeng, Bodzin and Saltiel, 2007). Alternatively, obesity leads to a decrease in the bioavailability of vitamin D due to its deposition in body fat compartments (Wortsman *et al.*, 2000).

Recent research has suggested a potential link between MS progression and gut microbiota. A study in 2017 identified specific gut bacteria associated with MS that increased the proinflammatory T cell response, thereby increasing the proinflammatory environment within MS patients (Cekanaviciute *et al.*, 2017). Following transplantation of the gut bacteria from MS cases into germ-free mice, the mice experienced more severe symptoms of experimental autoimmune encephalomyelitis (an artificial mouse model of MS) than the mice with bacteria transplants from healthy control individuals (Cekanaviciute *et al.*, 2017). This was supported by another study that carried out faecal transplants to germ-free mice from monozygotic twins discordant for MS (Ridaura *et al.*, 2013). The mice that received the MS microbiota transfer again showed an exacerbation of EAE symptoms, along with a decrease in Interleukin 10 (IL-10) production (Ridaura *et al.*, 2013). IL-10 is an anti-inflammatory cytokine, a protein secreted by certain immune system cells that are important in cell signalling and is thought to have therapeutic effects in MS patients (Ersoy *et al.*, 2005). This research suggests that gut microbiota in MS patients may produce a sustained proinflammatory environment, giving potential for therapeutic targeting of the gut microbiota for MS treatment. Although other chronic inflammatory diseases have been associated with a change in gut microbiota, there needs to be more exploration in determining if the relationship between disease status and gut microbiota is causal. A causal relationship between gut microbiota and obesity has been previously identified (Ridaura *et al.*, 2013), and so a similar relationship with MS should not be readily ruled out.

1.2 Genetics of Multiple Sclerosis

1.2.1 Introduction

Multiple Sclerosis is a complex, relatively common disease, with a unique genetic architecture. The evidence that has been accumulated over the years points to MS having one locus with moderate effect (*HLA-DRB1*15:01*) and multiple loci with small effects (O'Gorman *et al.*, 2013). Over the past 50 years, genetic studies have identified over 200 genetic associations with Multiple Sclerosis (Patsopoulos, 2018). Particularly recently, large studies conducted by consortia such as the International Multiple Sclerosis Genetics Consortium (IMSGC) have involved upwards of 45,000 Multiple Sclerosis cases (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). These large-scale studies have created detailed genetic maps of MS, along with integrated functional annotations (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). The majority of the associations identified in these genome wide association studies (GWAS) have implicated pathways within the innate and adaptive immune system, with the most strongly associated MS risk gene found in the HLA complex (International Multiple Sclerosis Genetics Consortium *et al.*, 2017).

The identification of genetic associations is not the final objective in studying the genetics of MS. After identifying the causal gene, the specific mechanisms and pathways the gene influences need to be clarified to determine what role the gene has in MS susceptibility and development. This is not always obvious, as the majority of MS-associated variants lie in intronic or intragenic regions of the genome (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). However, functional studies are improving with the advancement in high-throughput technologies to create MS-specific genomic maps to provide more information for the consequences of the genes implicated in MS (Patsopoulos, 2018).

1.2.2 Heritability

The initial suggestion that MS was partly hereditary came in 1896 from Eichhorst (Hermann Eichhorst in Ziirich, 1896), who discovered the disease clustered in families. Other observations suggested MS was caused at least somewhat by genetic factors,

including an excess occurrence of the disease in Northern Europeans in comparison to indigenous populations residing in the same latitude (Oksenberg and Barcellos, 2000).

Multiple Sclerosis disease risk has been measured in a variety of familial relationships, including parents and children, monozygotic twins, dizygotic twins, siblings and half-siblings (Sadovnick *et al.*, 1996; Ebers *et al.*, 2004; Hoppenbrouwers *et al.*, 2008; Hawkes and Macgregor, 2009). Within the family structure of an MS patient, it was found that the risk of developing MS in first-degree relatives is approximately 1-3%; siblings have a risk of 2.2% while parents have a risk of 1.4% (O'Gorman *et al.*, 2013). A risk of approximately 17.3% is expected for identical twins (O'Gorman *et al.*, 2013). In comparison, the general population risk of developing MS is 0.3% (O'Gorman *et al.*, 2013).

Women have been shown to be more at risk for developing MS; the male to female risk ratio was 1:2 throughout the twentieth century, however Orton et al. (2006) reported an increase to 1:3 over the past 50 years across Canada (Orton *et al.*, 2006). There is no consensus on parent-of-origin effects, as there are many differences in family analysis although there appears to be a trend for maternal transmission (Ebers *et al.*, 2004; Herrera *et al.*, 2008; Hoppenbrouwers *et al.*, 2008) as there are more mother-daughter pairs of MS than father-son pairs (Sadovnick, Bulman and Ebers, 1991). In a study of half-siblings, maternal half-siblings (where two individuals share the same mother but not the same father) had an MS recurrence rate of 2.35%, whereas paternal half-siblings had a lower rate of 1.31% (Ebers *et al.*, 2004). A study with avuncular relationships (aunts/uncles-nieces/nephews) examined MS-affected avuncular pairs with an unaffected mother in comparison to MS-affected avuncular pairs with an unaffected father (paternal pair) (Herrera *et al.*, 2008). A significantly higher ($p=0.038$) number of maternal avuncular pairs were affected with MS (Herrera *et al.*, 2008).

The heritability, or the proportion of variance in MS liability that can be explained by genetic factors, has been estimated as 0.64 (with a 95% confidence interval of 0.36 - 0.76), with twin, sibling and half-sibling data (Westerlind, Ramanujam, *et al.*, 2014). This is a similar finding to other autoimmune diseases, where median values of autoimmune disease heritability is 0.60 (Selmi, Lu and Humble, 2012). However, single nucleotide polymorphism (SNP) heritability estimates from the most recent IMSGC study are lower at 0.19 (95% CI 0.18, 0.20) (International Multiple Sclerosis

Genetics Consortium *et al.*, 2017). These heritability estimates, along with the concept of "missing heritability" will be discussed further in **Chapter 3**.

1.2.3 HLA genes

Among the genetic contributors to MS, the Human Leukocyte Antigen (HLA) genes associated with MS are considered to be the most influential (Chao *et al.*, 2011). The HLA complex, found on chromosome 6p21, contains a group of highly polymorphic genes involved in human immune function, including genes that encode the major histocompatibility complex (MHC) proteins (Lincoln *et al.*, 2005). These MHC proteins are necessary for the recognition of pathogens by the acquired immune system and are essential for mediating cell interactions with leukocytes (Lincoln *et al.*, 2005). There are three classes of MHC proteins, with MS risk associated with genes belonging to classes I and II (Lincoln *et al.*, 2005). MHC class I molecules (HLA-A,-B and -C) present non-self peptides from inside the cell (Martin, 2008). For example, destroyed virus fragments are presented to instruct the immune system to destroy the cell. This destruction is carried out by CD8+ T cells (Høglund and Maghazachi, 2014). Unlike MHC class I molecules, MHC class II molecules (HLA-DP,-DM,-DOA,-DOB,-DQ and -DR) present non-self antigens that are found extracellularly, to stimulate CD4+ T cells (Moutsianas *et al.*, 2015). These CD4+ T cells then go on to stimulate B cells to produce antibodies for that specific antigen (Mallucci *et al.*, 2015). Any cells that have self-antigens presented by MHC class II molecules are suppressed by regulatory T cells. The connection of HLA class II genes with MS risk corresponds with the biology of the disease: after naive T cells have matured within the thymus, they enter the secondary lymphoid tissues following circulation in the blood, and it is within the lymphoid tissues that these cells interact with antigen-presenting cells (Dendrou, Fugger and Friese, 2015). HLA-II molecules present antigens to the T cell receptors, and it is the genes that code for these molecules which have been associated to an increased risk of developing MS (Moutsianas *et al.*, 2015).

A total of 32 independent effects have been identified within the MHC region (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). Although MS-associated genes have been found in both classes of MHC molecules, HLA class II genes are the main genetic contributors to MS, explaining up to 10.5% of the genetic variance

underlying MS risk (Chao *et al.*, 2011; Sawcer *et al.*, 2011). The first association between MS and the HLA region was found in 1972 (Jersild C, Svejgaard A, 1972), and this signal was refined in 2002 to the class II allele *HLA-DRB1*15:01* (Barcellos *et al.*, 2002). The odds ratio for this allele was 3.08 (Barcellos *et al.*, 2002); i.e., the odds of an individual with that specific allele having MS are 3.08 times the odds of an individual without that allele having MS. Fine-mapping of the HLA region using HLA-specific reference panels confirmed this allele as having the strongest effect on MS risk within European ancestry populations (Patsopoulos *et al.*, 2013). Other independent *HLA-DRB1* associations include *03:01*, *04:04*, *04:01*, *13:03* and *14:01* (Patsopoulos *et al.*, 2013). The majority of the collective effect of these alleles can be explained by changes to four amino acids which exist in the peptide-binding groove of the HLA-DR molecule (Patsopoulos *et al.*, 2013). These changes would affect the recognition and binding of antigens to this molecule (Patsopoulos *et al.*, 2013).

In addition to HLA class II alleles, HLA class I alleles have been implicated in affecting MS risk (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). A role for MS protection has been identified within the class I region: *HLA-A*02:01* has been shown to protect independently from *HLA-DRB1*15* activity (Fogdell-Hahn *et al.*, 2000; Patsopoulos *et al.*, 2013), and has a reported odds ratio of 0.52 (Yeo *et al.*, 2007). *HLA-C*05* was also shown to have a protective effect against MS susceptibility, independent of *HLA-DRB1* effects (Yeo *et al.*, 2007). *HLA-B* appears to contain 6 independent effects for MS susceptibility (International Multiple Sclerosis Genetics Consortium *et al.*, 2017).

There has also been a suggestion that interactions between HLA alleles may further influence MS risk, with identified interactions between *HLA-DQA1*01:01-HLA-DRB1*15:01* and *HLA-DQB1*03:01-HLA-DQB1*03:02* (Moutsianas *et al.*, 2015). However, their role in MS risk has yet to be examined fully (Patsopoulos, 2018).

1.2.4 Non-HLA genes

Although genes within the HLA region exert the strongest effect on MS susceptibility, over 200 variants have been identified that lie outside of this region, including within the X chromosome (International Multiple Sclerosis Genetics Consortium *et al.*, 2017).

The largest and most recent study conducted by the IMSGC (with 47,351 MS cases and 68,284 controls) identified novel genes with OR from 1.05 to 1.20 and described many of the molecular events that underpin MS susceptibility (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). These variants explain approximately 20% of the genetics of MS. All major immune cell types were shown to have a high enrichment of MS-associated variants, with MS-associated molecular events dispersed widely across the innate and adaptive immune system.

Within the adaptive immune system, T cells and B cells are enriched with MS variants, and within the innate immune system both natural killer and dendritic cells are strongly enriched. In specific tissues, the thymus has an enrichment of MS susceptibility genes, which indicates a possible role for the thymus in selecting autoreactive T cells in MS (Pugliese *et al.*, 1997; International Multiple Sclerosis Genetics Consortium *et al.*, 2017). Within the brain, there is an enrichment of MS genes within the microglia, but not within neurons or astrocytes. This indicates that immune cells that are resident within the brain such as microglia, contribute to MS susceptibility. Although there does not appear to be an enrichment of MS susceptibility loci within CNS tissues, this does not exclude the idea that some variants may directly affect neuronal tissues or their supporting cell types (Baranzini and Oksenberg, 2017). The IMSGC study also predicted the functional consequences for non-HLA MS variants using gene expression levels within different tissue types. For example, MS variant CLECL1 had an expression level that was 20-fold greater in cortical microglia when compared to bulk cortical tissue, suggesting a role for microglia in MS susceptibility. Overall, the findings suggest that the origin of MS begins in the peripheral immune system, with functional failures across all parts of the immune system cumulating in the disease.

However, determining the precise functional effects exerted from identified MS variants is often complex. For instance, a pathway that has been implicated in MS susceptibility can elicit different responses depending on which cell type the pathway occurs in. A good example of this is the response to type I IFNs: MS risk has been associated with response to type I IFNs, however it is unknown if the disease risk comes from the modified function of one specific cell type, or if all the cell types contribute to MS risk equally. Innate immune cells respond to type I INFs by increasing the presentation of antigens and the production of cytokines and chemokines, while in the adaptive

immune system, B cells enhance antibody production and T cells amplify their effector function (Ivashkiv and Donlin, 2014). This emphasises the importance of the context of each genetic variant.

1.2.5 Rare variants

A rare variant is a genetic variant that appears at a low frequency (<0.5%) in a population. Rare variants have the potential to have a large effect on disease susceptibility and might be kept rare by the action of selection. It is possible that there are some forms of MS that are caused by rare variants: although some evidence has been presented which supports this, replication of these findings has been limited. For example, a rare variant in the *CYP27B1* gene was found when 43 families with at least four cases were exome sequenced (Ramagopalan *et al.*, 2011), but large-scale follow-ups did not replicate this finding. Additionally, a *NR1H3* mutation was suggested to cause a Mendelian form of MS after discovery in 7 patients across 2 families (Wang *et al.*, 2016), however no evidence of this mutation was found in the 2016 IMSGC study (Antel *et al.*, 2016) and there have been individuals reported to carry this mutation who do not have MS or a similar disease (Minikel and MacArthur, 2016).

Exome sequencing, where all protein-coding genes within a genome are sequenced, has provided some evidence for rare variants affecting MS susceptibility (such as a mutation in the *NLRP1* gene (Maver *et al.*, 2017) and *TYK2* gene (Dyment *et al.*, 2012)), however follow up replication studies, particularly using whole-genome sequencing, are needed to confirm these.

1.2.6 Gene-gene interactions

Epistasis is the interaction between alleles at different loci, and is another potential contributor to MS heritability (Cordell, 2002). Although common variants account for the bulk of MS heritability, they do not account for it entirely (International Multiple Sclerosis Genetics Consortium, 2018). It is possible that some of these common variants interact with each other, and therefore contribute more to MS risk in combination than they do alone (International Multiple Sclerosis Genetics Consortium, 2018). Although limited, there has been some evidence that suggests that some heritability may come from these gene-gene interactions.

Evidence of interactions between HLA class II alleles has been found, with *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* having novel epistatic interactions (Lincoln *et al.*, 2009). Evidence for interactions between HLA class II alleles *HLA-DQA1*01:01-HLA-DRB1*15:01* and *HLA-DQB1*03:01-HLA-DQB1*03:02* were later confirmed in a large-scale study (17,465 cases and 30,385 controls) (Moutsianas *et al.*, 2015). Additionally, evidence has been presented for epistasis of *HLA-DRB1*1501* with several alleles: the *IL-2-330T* allele (Shahbazi *et al.*, 2010) and the *TNF-α* – 308 G/A polymorphism (Shahbazi *et al.*, 2011) have both been shown to interact with *HLA-DRB1*1501* to increase susceptibility to MS. Another HLA gene, *DDX39B*, was shown to have epistatic interactions with alleles in *IL7R* exon 6, a non-HLA locus (Galarza-Muñoz *et al.*, 2017). Thus, the extent of epistatic interactions for MS is not confined to the MHC region, although the role of these interactions is somewhat elusive (Patsopoulos, 2018). While the current evidence shows that epistasis is involved in MS risk on some level, the limited amount of results discovered suggests that they are not major contributors towards the remaining proportion of heritability that is not accounted for by common variants. This is consistent with other common complex diseases, where evidence of epistasis has not been abundant (Altshuler, Daly and Lander, 2008).

1.2.7 Gene-environment interactions

A genetic or environmental risk factor will have an absolute value for the effect it contributes to MS risk within an individual. When an individual is exposed to both these causal genetic and environmental factors and the disease risk caused by those two factors is higher than the sum of their two absolute effect values, then gene-environment interaction is present between these two factors. In other words, individuals who have different genotypes respond to variation in the environment in different ways. The interactions between environmental factors and genetic variants has the potential to explain a part of the heritability of MS (Baranzini and Oksenberg, 2017).

Smoking is a known risk factor for MS (Biran and Steiner, 2004; Handel and Ramagopalan, 2011). It has also been shown to interact with HLA risk variants, including *HLA-DRB1*15:01*. A combination of *HLA-DRB1*15* presence, absence of protective allele *HLA-A*02* and smoking gave an OR of 13.5 (compared to an OR of 4.9

in non-smokers). This finding has been replicated in individuals exposed to passive smoke (Hedström *et al.*, 2014). A non-HLA MS gene, *NAT1*, which encodes an enzyme for smoke product metabolism, has also been shown to interact with smoking to influence MS susceptibility (Briggs *et al.*, 2014). The effect of smoking on an individual's risk of MS therefore depends on not only the HLA genotype but also other parts of an individual's genome. Smoking is known to post-translationally modify peptides in the lungs through inducing enzyme activity (Klareskog, Catrina and Paget, 2009). It is possible that these modified peptides may be recognised by T cells which have not been removed in the thymus, resulting in autoreactive T cells (Hedström *et al.*, 2011) (Odoardi *et al.*, 2012).

In addition to smoking, there is a possible interaction between Epstein–Barr virus (EBV) and MS (Sundqvist *et al.*, 2012). Epstein–Barr nuclear antigen 1 (EBNA1) is a protein associated with EBV; it is the only EBV protein observed in all EBV-related malignancies (Sundqvist *et al.*, 2012). An interaction between increased EBNA1 titres and HLA variants has been found to increase MS susceptibility (Sundqvist *et al.*, 2012). This may suggest common pathogenetic pathways between EBV and HLA alleles, however it does not confirm a causative role for EBV with MS (Olsson, Barcellos and Alfredsson, 2017). It is possible that increased antibody titres for EBNA1 may be due to poor virus elimination from insufficient cell-mediated immunity against EBV in people with MS (Olsson, Barcellos and Alfredsson, 2017). Therefore, further research is needed to clarify the relationship between EBV and MS.

Low plasma Vitamin D (1,25-dihydroxyvitamin D3) is a well-known risk factor for MS (Cantorna, 2006). Strong evidence has been presented for the interaction between vitamin D and MS risk genes. Vitamin D response element (VDRE), onto which the receptor for 1,25(OH)2D binds, is found in the promoter region for *HLA-DRB1*; thus it is very likely that the expression of *HLA-DRB1* is controlled by vitamin D (Ramagopalan, Maugeri, *et al.*, 2009). Several other genes that are associated with MS susceptibility have been shown to be regulated by vitamin D, such as *NADSYN1* and *DHCR7* (Ahn *et al.*, 2010). This suggests that vitamin D has a role in modulating MS risk.

In the future, it is possible that the relationship between environmental and genetic factors will be understood further by using epigenetic studies (Olsson, Barcellos and

Alfredsson, 2017). Epigenetics is the study of changes made in an organism that are caused by the modification of gene expression (Petronis, 2010). These changes can be caused by methods including DNA methylation, where methyl groups are attached to the DNA molecule, and post-translational histone modifications, which can alter the structure of chromatin (Petronis, 2010). Many cell types within the immune system and CNS display different patterns of modification (Maltby *et al.*, 2015). Not only can these modifications change the activity of the genome in response to environmental factors (Gao *et al.*, 2015), but they could also mediate the effect of exposure of environmental factors to genetic variation (Olsson, Barcellos and Alfredsson, 2017). Studies are beginning to show differences in epigenetic modifications between MS cases and controls (Baranzini *et al.*, 2010; Maltby *et al.*, 2015), however more research will help improve the understanding of the interactions between environment and genes.

Although it is difficult to quantify gene-environment interactions due to problems such as confounding and reverse causation (Patsopoulos, 2018), assessing the additive interactions between environmental exposures and susceptibility variants could offer further insight into the genetic architecture of MS. Additionally, methods such as Mendelian Randomisation are able to help determine causative factors (Davey Smith and Ebrahim, 2005). Regardless, the presence of gene-environment interactions allows lifestyle changes to positively impact MS susceptibility.

1.2.8 Genetic link to other autoimmune diseases

Comorbidities are defined as the presence of one (or more) diseases that are in addition to an initial disease, such as Multiple Sclerosis. Many individuals who have developed MS often develop other immune-mediated inflammatory diseases, or have close family members that do (Nielsen *et al.*, 2008). The inflammatory diseases that are found to be in comorbidity with each other may share similar pathways and have an overlapping genetic structure (Barrett *et al.*, 2008; Hunt *et al.*, 2008; Stahl *et al.*, 2010; Sawcer *et al.*, 2011). For example, many genes that are associated in autoimmune and inflammatory diseases are pleiotropic, or exhibit effects on multiple traits (Wagner and Zhang, 2011). Many of the variants that have been associated with MS have been associated with diseases such as lupus, psoriasis, rheumatoid arthritis, ulcerative colitis, Crohn's disease and type 1 and 2 diabetes (Sawcer *et al.*, 2011). For example, type 1

diabetes susceptibility genes *CLEC16A* and *CD226* also affect susceptibility to MS (Booth *et al.*, 2009). However, being susceptible to one disease does not always result in a negative effect on another, as many autoimmune disease susceptibility loci can act protectively against other diseases (Sirota *et al.*, 2009). For example, key MS risk variant *HLA-DRB1*1501* offers a protective effect for type 1 diabetes.

A limited number of studies have examined the link between MS and neurological diseases (Patsopoulos, 2018). One study that focused on 25 brain disorders (such as Alzheimer's disease and MS) found no evidence that linked MS to any of the other neurological diseases (Anttila *et al.*, 2017). Additionally, no shared association has been found with MS and amyotrophic lateral sclerosis (Goris *et al.*, 2014) or schizophrenia (Goris *et al.*, 2014). This suggests that the pathogenic processes for MS differs from those of neurological disorders and is more in line with that of inflammatory, autoimmune disorders.

1.2.9 Population heterogeneity

The majority of large-scale genetic MS studies have been carried out in European populations (International Multiple Sclerosis Genetics Consortium, 2010; Patsopoulos and (IMSGS), 2016). However, studies from other subpopulations have suggested that different ethnic groups or subpopulations have different MS susceptibility variants.

For example, the Italian island Sardinia has one of the highest risks of MS in the world, with a prevalence of 247.6 per 100,000 individuals (Pugliatti, Sotgiu and Rosati, 2002; Sotgiu *et al.*, 2004). Sardinians have been found to have different HLA variants associated with MS than those found on mainland Italy, with haplotypes *DRB1*0301–DQA1*0501–DQB1*0201* and *DRB1*0405–DQA1*0501–DQB1*0301* having the strongest associations (Marrosu *et al.*, 1997). Ashkenazi Jews also have different HLA associations with MS, with *HLA-A68:02* and *HLA-B38:01-HLA-C12:03* implicated with MS risk (Khankhanian *et al.*, 2015).

Studies within African-American populations have been relatively small, and have therefore not revealed any novel variants associated with MS other than those found in European populations (Johnson *et al.*, 2010; Isobe *et al.*, 2013, 2015). However, in

Afro-Brazilian populations, *DQB1*0602* was found to have an association with MS (Caballero *et al.*, 1999).

These findings suggest that the broad immune pathways of MS are consistent among populations, however the variants implicated do vary. This has the potential to be significant for treatment and prevention strategies for smaller subpopulations.

1.3 Multiple Sclerosis in the Northern Isles

1.3.1 Population isolates

Population isolates are unique groups of individuals that are geographically, culturally or linguistically separated from nearby populations (Hatzikotoulas, Gilly and Zeggini, 2014). As a group, they have been of interest in of human genetics studies for a number of unique attributes (Jorde *et al.*, 2000). In general, they are less genetically diverse than larger, non-isolated populations as alleles are more likely to reach fixation or extinction than in larger populations (Kittles *et al.*, 1998). The founders, which are typically a small subset of individuals, may have particularly high or low frequencies of certain alleles by chance. This is particularly useful when a rare allele associated with a disease in the parent population drifts to higher frequency in the isolate population, making it easier to identify (genetic drift refers to the random fluctuations of allele frequencies within a population) (Hatzikotoulas, Gilly and Zeggini, 2014). Additionally, isolates tend to have a more uniform genetic, environmental and cultural background (Peltonen, Palotie and Lange, 2000), which makes them particularly useful for studying complex disease genetics. In general, isolates have lesser differences in diet, exercise, climate and infectious disease exposure than that found in a large, non-isolated population. Additionally, language and religious uniformity helps in fostering social unity. This helps to reduce environmental noise which can be confounding in complex diseases studies. However, population isolate growth is more susceptible to events which may cause bottlenecks, such as environmental change, infectious disease, war and famine (Jorde *et al.*, 2000). Recovery from a bottleneck is also influenced by higher rates of inbreeding and genetic drift.

There are several isolated populations which have higher frequencies of MS prevalence. Orkney, Sardinia and Iceland are excellent examples of geographically isolated populations with an MS prevalence higher than the surrounding populations, while the Finno-Ugric-speaking Saami population who are indigenous to northern Scandinavia and remain linguistically and culturally isolated, also have a high prevalence of MS (182.4 per 100,000) (Benjaminsen *et al.*, 2014).

Population isolates are likely to have unique factors, either genetic, environmental or both, that contribute to their high MS prevalence. The Italian island of Sardinia has one of the highest risks of MS in the world, with the adjusted total prevalence rate at 247.6 per 100,000 individuals (Pugliatti, Sotgiu and Rosati, 2002; Sotgiu *et al.*, 2004). Within this population, the incidence rate of MS has increased substantially over the past forty years (Sotgiu *et al.*, 2003). One possible reason for this is the "hygiene hypothesis". Post-World War II, a distinct lifestyle change (which included a malaria eradication campaign) left a population which had evolved to combat a multitude of parasites and pathogens with no need for immunogenetic mutations, leaving the Sardinian population prone to autoimmune diseases (Sotgiu *et al.*, 2003).

Population isolates can also be a key resource when studying a complex disease such as MS. A study of a high-risk population isolate within Finland revealed a novel *STAT3* gene variant that had a protective association with MS. The study noted how using a GWAS on an isolated population aided discovery of the gene (Jakkula *et al.*, 2010). Another study of Finland looked at the Seinäjoki-South region, which has a significantly high prevalence of MS (219 / 100,000 individuals). Due to the settlement history of the region, along with molecular genetic evidence, the study suggested that the high proportion of MS was caused by a founder effect (Tienari *et al.*, 2004). When using population isolates to study a complex genetic trait, there is always potential to discover unique genetics or environmental causes that can give more information on the aetiology of the disease.

1.3.2 Multiple Sclerosis in the Northern Isles of Scotland

It has previously been discussed that Multiple Sclerosis varies with geographical distance from the equator, with a positive association seen between MS prevalence and global latitude (Simpson *et al.*, 2011). This gradient is seen at a regional level within the

United Kingdom; Wales has a prevalence of 138 per 100,000 individuals, England with 164 per 100,000, Northern Ireland with 175 per 100,000 and Scotland with 209 per 100,000. However, the Northern Isles of Scotland have significantly higher rates: with 295 per 100,000 found in Shetland and 402 per 100,000 in Orkney, the highest prevalence of MS in the world (Visser *et al.*, 2012). Based on global trends, the prevalence found in Orkney is significantly higher than what would normally be expected; Orkney therefore has an unexplained excess of MS prevalence. A number of previous studies have investigated the potential cause of this excess of MS prevalence.

Vitamin D deficiency was investigated by Weiss et al. in 2016, as at 10 to 60 miles north of mainland Scotland, vitamin D levels were expected to be lower than those on the mainland (Weiss *et al.*, 2016). Since strong associations between MS and vitamin D deficiency have been identified, it was speculated that this may be the cause of the high MS prevalence in the Northern Isles. However, this cross-sectional study comparing MS control individuals in Orkney to those on mainland Scotland found this was not the case. Mean plasma 25-hydroxyvitamin D was found to be significantly higher in those on Orkney (mean 35.3 nmol/L compared to 31.7 nmol/L). Additionally, Orkney had a lower prevalence of severe plasma 25-hydroxyvitamin D deficiency (of 6.6% compared to 16.2% in mainland Scotland). A combination of a high number of farming and outdoor occupations, along with older age and foreign holidays were significantly associated with the higher levels of plasma 25-hydroxyvitamin D on the islands.

Another study investigated homozygosity on Orkney (McWhirter *et al.*, 2012). Many isolated communities often have higher degrees of parental relatedness, and it was thought this may contribute to the excess of MS prevalence (Roberts, Roberts and Poskanzer, 1983). Three measures of homozygosity were generated for 88 MS patients and 178 controls and assessed for association with MS. However, no association was detected, and so consanguinity and inbreeding are not thought to be the cause of excess MS prevalence.

It is possible that this excess prevalence may be explained genetically through the Northern Isles having a higher proportion of common risk alleles. If founders of the islands by chance had higher frequencies of these risk alleles, this may cause additional cases of MS. Alternatively, one or more MS risk variants which are rare in mainland Scotland may be present at an elevated frequency in the Northern Isles.

It is possible that the Northern Isles may have unique rare variants. If rare alleles were the cause of the excess of MS prevalence in Orkney, this has the potential to provide a new insight into new pathways or identify novel candidate genes for research. This has been the case for other diseases, for example a rare non-polyposis colon cancer gene was discovered, resulting in a novel molecular mechanism identified for malignancy (Bronner *et al.*, 1994).

Other genetic reasons can include variations in copy number variants, but from other complex disease studies (Craddock *et al.*, 2010) these are unlikely candidates for explaining MS heritability in the Northern Isles.

Epigenetic variations, for example histone modifications, inherited expression of non-coding RNA and DNA methylation may explain some of the excess of MS prevalence. Very few epigenetic studies have been performed for MS, and those that have been conducted do not provide strong evidence to support the occurrence of large transgenerational epigenetic risk factors (Petronis, 2010; Grossniklaus *et al.*, 2013; Westerlind, Ramanujam, *et al.*, 2014).

Alternatively, MS in the Northern Isles may be influenced by gene-environment (GxE) and gene-gene (GxG) interactions (Zuk *et al.*, 2012; Beecham *et al.*, 2013). However, to detect typical GxG interactions, a very large sample size is needed; as the population of Orkney is approximately 21,850 and Shetland is around 23,080, this would be a huge limiting factor for these studies (Zuk *et al.*, 2012). Very few reported GxE analyses have been conducted for MS, again due to a large required sample size with corresponding environmental data (Lill, 2014).

Regardless, studying Multiple Sclerosis in the Northern Isles is nearly always going to be limited by sample size, given the small population of the islands and an even smaller number of MS cases. However, a number of analysis methods can still be performed on subsets of each population. Genetic and phenotypic data has been gathered for approximately 10% of each population (the ORCADES and VIKING cohorts, described in the next chapter). The research in this thesis aims to use this data and perform a number of genetic analysis methods to elucidate why the prevalence of MS in these Northern Isles remains so high.

1.4 Aims of the study

1.4.1 Complex disease research

With the advent of modern genetic disease research, an initial focus was placed on Mendelian conditions. Caused by one major genetic defect, Mendelian diseases such as Huntington's disease result in a few cases within the population which are often characterised by a specific transmission pattern (for example, dominant or recessive inheritance) (Lander and Schork, 1994). These studies were expensive, often requiring extensive family-based data collection, and commonly resulted in a low power for complex diseases (Altmüller *et al.*, 2001; Pearson and Manolio, 2008). As analysis methods progressed, thousands of individuals were able to be interrogated at a previously unreachable resolution and the focus shifted from Mendelian to complex diseases (Hirschhorn and Daly, 2005).

The complexity of many common diseases develops from the presence of both influential genetic and environmental factors (Hirschhorn *et al.*, 2002). These factors contribute to an individual's susceptibility to a particular complex disease in a probabilistic manner (as opposed to the deterministic risk alleles that are found in monogenic disorders), although no one single pattern of inheritance is followed for all complex disease. The number, frequency, size and type of associated variants will differ between diseases (Pritchard, 2002). However, in general selection acts to reduce the frequency of high-effect alleles to prevent individuals with extreme allelic effects from becoming commonplace in a population (Gibson and Wagner, 2000; Gibson, 2009). Therefore, it is more likely that common variants with low to moderate effect sizes contribute towards the genetics of complex diseases.

Understanding the genetic architecture of a disease is essential to progressing in a clinical setting. Complex disease research can lead to a greater understanding of the pathology and cellular mechanisms that contribute to disease risk and development. Disease diagnosis, patient treatment, disease management (for example, enabling a patient to make beneficial lifestyle choices), disease prediction and treatment response prediction (including the onset, severity and response to treatment) can all be improved from insight into the genetics of a complex disease (Manolio *et al.*, 2009).

There are numerous genetic analysis methods which have the potential to reveal insight into complex disease mechanisms. This thesis seeks to apply several of these methods in understanding Multiple Sclerosis and answer the question as to why the Northern Isles of Scotland have the highest rates of MS in the world.

1.4.2 Research objectives

Multiple Sclerosis is a multifactorial disease of autoimmune origin which is increasingly common at higher latitudes including Scotland. It has previously been shown that the Northern Isles of Scotland have the highest prevalence of MS in the world. Various risk factors, both genetic and environmental, are implicated in MS, but the reasons for the peak in Orkney and Shetland are not well understood. This thesis seeks to better understand these very high rates through a number of approaches using the data of the Northern Isles Multiple Sclerosis study, Orkney Complex Disease Study and the Viking Health Study - Shetland.

The specific objectives of this research are as follows:

- establish a heritability estimate for MS in Orkney and Shetland;
- identify any novel SNPs which may contribute to MS in the Northern Isles, and;
- determine the contribution of common risk variants to the excess risk in the Northern Isles.

It is hoped that the findings of this thesis will contribute to a greater understanding of MS both within the Northern Isles and to MS research as a whole.

CHAPTER 2: STUDY DATA

This chapter aims to explore the datasets used in this thesis by describing their content and structure and summarising the quality control procedures implemented prior to obtaining the data. Further quality control checks that I performed on receiving the data will be described fully.

2.1 Introduction

2.1.1 Cohort populations

This thesis focuses on the genetics of Multiple Sclerosis in the Northern Isles of Scotland, specifically Orkney and Shetland. Orkney is an archipelago of 70 islands, 10 miles north of Scotland, with a population of 21,850 (National Records of Scotland, 2018). Shetland lies 50 miles north of Orkney and has a population of 23,200 people that inhabit 15 islands (National Records of Scotland, 2018). The ancestral history of the people of Orkney and Shetland differs from that of mainland Scotland: both have been inhabited for at least 6000 years but experienced an influx of Norse settlers during the 9th century (Wilson *et al.*, 2001; Capelli *et al.*, 2003; Goodacre *et al.*, 2005). It was not until the 15th century that both island groups became part of Scotland; thus, the ancestral genetics of the current population has a significant Scandinavian influence (Goodacre *et al.*, 2005). Scandinavian mtDNA and Y DNA lineages are found in 30% of individuals in Orkney, and 44% of individuals in Shetland, significantly high than the 15% found in the North West coast of Scotland (Goodacre *et al.*, 2005).

In addition to a distinctive ancestral history, the Northern Isles also have remained largely isolated from the rest of the United Kingdom. The geographical position of the islands combined with limited transport links (which are often reliant on weather conditions) means access to and from the islands is limited; this was particularly true in the past, before boat and air services were operational. Therefore, both sets of islands remain considerably genetically isolated from mainland Scotland (Vitart *et al.*, 2005; Weiss *et al.*, 2016).

The unique genetics and isolated nature of the islands have led the Northern Isles to become of focus to genetic health researchers. Two cross-sectional, family-based

cohorts have been established in Orkney and Shetland and have become platform resources for the study of complex disease in Scotland. For Multiple Sclerosis specifically, the Northern Isles Multiple Sclerosis Study (NIMS) was also created. These three datasets, NIMS, The Orkney Complex Disease Study (ORCADES) and Viking Health Study - Shetland (VIKING) are the primary datasets that are used in this project.

In addition to ORCADES and VIKING, **Chapter 5** brings the introduction of the Generation Scotland dataset to allow a comparison of the Northern Isles populations to mainland Scotland (Smith *et al.*, 2013). Generation Scotland is a family-based genetic epidemiology cohort gathered for 23,960 individuals across Scotland. For the research purposes in this study, only individuals from Glasgow and Dundee were included for analysis (n=8787). Individuals from other regions within Generation Scotland were excluded (for example, Aberdeen); Northern Isles residents frequently travel to and reside in Aberdeen as this region holds the principal ferry route to and from the Northern Isles. Therefore, to have a clear representation of mainland Scotland, only individuals from southern Scotland were included.

2.1.2 Data collection

Data collection consisted of gathering phenotypic and genotypic data from ORCADES, VIKING and NIMS. This data collection was entirely carried out by other researchers and their teams prior to this thesis.

ORCADES

Ethical approval for data collection was granted in 2004, with the collection period running from 2005 to 2011. Participation was voluntary and it was required that each participant had at least two grandparents born in Orkney, although the great majority had three or four grandparents from the archipelago. High density genotype data was collected from 2080 participants, and up to 500 disease-related phenotypes were recorded, including data gathered from venepuncture and cardiovascular measurement clinics. Most subjects gave measurements and participated in dual energy X-ray absorptiometry (DEXA) scans, eye clinics and cognitive function testing. There is a high degree of kinship within ORCADES participants, which includes both nuclear families and further relations (McQuillan *et al.*, 2008; McWhirter *et al.*, 2012).

VIKING

The Viking Health Study – Shetland (VIKING) is a family-based epidemiology cohort based in Shetland. The recruitment period lasted from 2013 to 2015 and each participant required at least two grandparents from the island. Genotypic and phenotypic information were collected from 2105 participants. Like ORCADES, there is a high degree of kinship among participants.

NIMS

The Northern Isles Multiple Sclerosis (NIMS) study provided a useful resource to improve case numbers in the ORCADES dataset (McWhirter *et al.*, 2012). This study recruited MS patients and controls specifically from the Northern Isles: of the 266 participants, 88 were cases. These individuals were born between 1937 and 1939 and were selected on the basis that their risk of developing MS would be lower (as they are over 70). All four grandparents were required to come from the same location.

Generation Scotland

Generation Scotland was established with the aim of creating a family-based cohort to represent the general population across Scotland for studying the genetics of health traits. A key feature of this cohort is the ability to link both past and future health records to individual data. Potential participants were invited at random based from collaborating medical practices, with the criteria that they were aged 18-65 and had one-first degree relative (> 18 years of age) who would also participate. A total of 6665 of individuals were recruited directly through invitation, along with 1288 individuals who volunteered without invitation, and 16 007 of their family members. This gave a cohort total of 23 960.

2.2 Cohort Data

2.2.1 Data summary

Data from ORCADES, VIKING and Generation Scotland was provided at the start of this thesis.

Table 1 - Table 4 tabulate the cohort information (

Table 1), sample quality control measures (Table 2), SNP quality control measures

(Table 3) and imputation information (Table 4). The quality control procedures described here were performed by other individuals prior to the beginning of this thesis. Following QC, 188 NIMS individuals were merged with 2027 ORCADES individuals. However, it is important to have a full understanding of the procedures to ensure that results produced in this thesis would not be inaccurate or biased.

	ORCADES	NIMS	VIKING	Generation Scotland
Region of Origin	Orkney	Orkney and Shetland	Shetland	Glasgow and Dundee
Subjects Genotyped	2080	266	2105	20,195
Genotyping platform	Illumina Hap300; Illumina Omni1; Illumina OmniX	Illumina Omni1	HumanOmni ExpressExome8 v1-2_A	HumanOmni ExpressExome8 v1-2_A
Genotyping calling algorithm	Beadstudio using Hap300v2 cluster files; Genome Studio using Illumina cluster files	Beadstudio using Hap300v2 cluster files; Genome Studio using Illumina cluster files	Beadstudio-Gencall v3.0	Beadstudio-Gencall v3.0

Table 1: Cohort information for ORCADES, NIMS, VIKING and Generation Scotland
Cohort information for Orkney Complex Disease Study (ORCADES), Northern Isles Multiple Sclerosis Study (NIMS), Viking Health Study – Shetland (VIKING) and Generation Scotland cohorts. This information was provided at the start of this thesis.

	ORCADES and NIMS (merged)	VIKING	Generation Scotland
Call rate	< 97%	< 97%	< 97%
Heterozygosity	FDR< 1%	FDR< 1%	FDR< 1%
Exclusions	Ethnic outliers; duplicates; gender mismatches; genomic kinship incompatible with pedigree	Ethnic outliers; duplicates; gender mismatches; genomic kinship incompatible with pedigree	Ethnic outliers; duplicates; gender mismatches; genomic kinship incompatible with pedigree
Final number of subjects	2215	2105	20,032

Table 2: Sample Quality Control information for ORCADES, NIMS, VIKING and Generation Scotland
Quality control information for Orkney Complex Disease Study (ORCADES), Northern Isles

Multiple Sclerosis Study (NIMS), Viking Health Study – Shetland (VIKING) and Generation Scotland cohorts. This QC work was carried out by other researchers prior to this thesis.

	ORCADES	NIMS	VIKING	Generation Scotland
MAF	< 0.01	< 0.01	< 0.01 for OMNI markers; < 0.0001 for Exome Chip markers	< 0.01 for OMNI markers; < 0.0001 for Exome Chip markers
HWE	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
Call rate	< 98%	< 98%	< 98%	< 98%

Table 3: SNP Quality Control information for ORCADES, VIKING and Generation Scotland
SNP quality control information for Orkney Complex Disease Study (ORCADES), Northern Isles Multiple Sclerosis Study (NIMS), Viking Health Study – Shetland (VIKING) and Generation Scotland cohorts. This QC work was carried out by other researchers prior to this thesis.

	ORCADES	VIKING	Generation Scotland
Post-QC SNP N	287,208	611,836	519,798
Phasing Software	Shapeit v2-r644	Shapeit v2-r837 and duohmm	Shapeit v2-r837 and duohmm
Imputation Panel	1000 Genomes Phase 1 integrated variant set v3	Haplotype Reference Consortium r1-1	Haplotype Reference Consortium r1-1
Imputation software	Impute v2.2.2	PBWT Sanger server	PBWT Sanger server
Imputed SNPs	~ 37,000,000	39,131,578	24,111,857

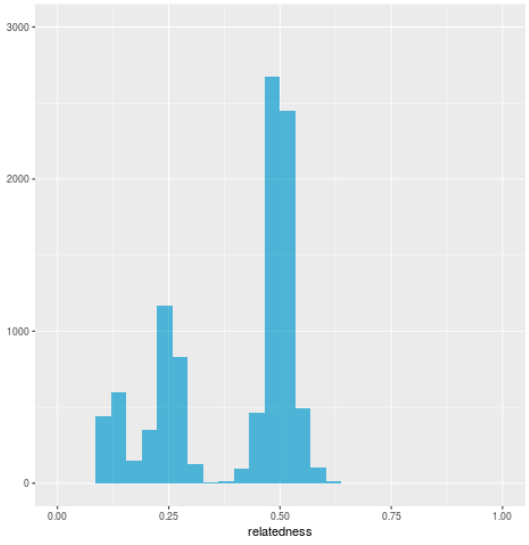
Table 4: Imputation information for ORCADES, VIKING and Generation Scotland
Imputation information for Orkney Complex Disease Study (ORCADES), Viking Health Study – Shetland (VIKING) and Generation Scotland cohorts. All imputation work was carried out by other researchers prior to this thesis. Imputation quality control was carried out by removing monogenic and low imputation quality (INFO < 0.4) variants.

2.2.2 Data description

In total, 2215 ORCADES individuals, 2015 VIKING individuals and 20,032 Generation Scotland individuals were available for analysis (Table 5). However, only 8787

individuals from Generation Scotland (from the Glasgow and Dundee regions) were taken forward for analysis; the data description here will apply only to these individuals.

There were 97 MS cases in ORCADES, 15 cases in VIKING and 30 cases in Generation Scotland; of these, the majority (n=69, n=12 and n=25, respectively) were female. The age distribution plots show generally normal distributions, with slightly uneven distributions with lesser sample sizes (Figure 3). There is some underlying population structure within the data (Figure 4), although there is no observed distribution bias of cases. Population structure may be due to reported or cryptic relatedness within the island groups. There are clear cluster differences between the Northern Isles groups and mainland Scotland. There are a number of closely related individuals within both



cohorts (Figure 5); these include first cousin relationships (relatedness coefficient ≅ 0.125), grandparent-grandchild and uncle/aunt-nephew/niece relationships (relatedness coefficient ≅ 0.25) and parent-offspring and sibling relationships (relatedness coefficient ≅ 0.50). It is therefore necessary to account for relatedness in appropriate analyses (such as GWAS).

		Count			Mean Age *(standard deviation)*		
Population	Sex	Case	Control	Total	Case	Control	Total
ORCADES	Male	28	843	871	54.30 *(9.40)*	54.85 *(15.20)*	54.83 *(15.04)*

	Female	69	1275	1344	49.13 *(12.36)*	53.85 *(15.54)*	53.61 *(15.43)*
	All	97	2118	2215	50.64 *(11.76)*	54.25 *(15.41)*	54.09 *(15.28)*
VIKING	Male	3	839	842	60.95 *(8.83)*	51.34 *(15.47)*	51.37 *(15.46)*
	Female	12	1251	1263	53.73 *(13.10)*	48.93 *(15.06)*	48.97 *(15.05)*
	All	15	2090	2105	55.28 *(12.39)*	49.90 *(15.27)*	49.93 *(15.26)*
Generation Scotland	Male	5	3574	3579	45.80 *(7.92)*	45.89 *(15.26)*	45.89 *(15.26)*
	Female	25	5134	5159	50.80 *(9.53)*	46.48 *(14.78)*	46.50 *(14.76)*
	All	30	8708	8738	49.97 *(9.35)*	46.23 *(14.98)*	46.25 *(14.97)*

Table 5: Summary statistics for ORCADES, VIKING and Generation Scotland
Count and mean age for the Orkney Complex Disease Study (ORCADES), Viking Health Study – Shetland (VIKING) and Generation Scotland cohorts, split by gender and (MS) status.

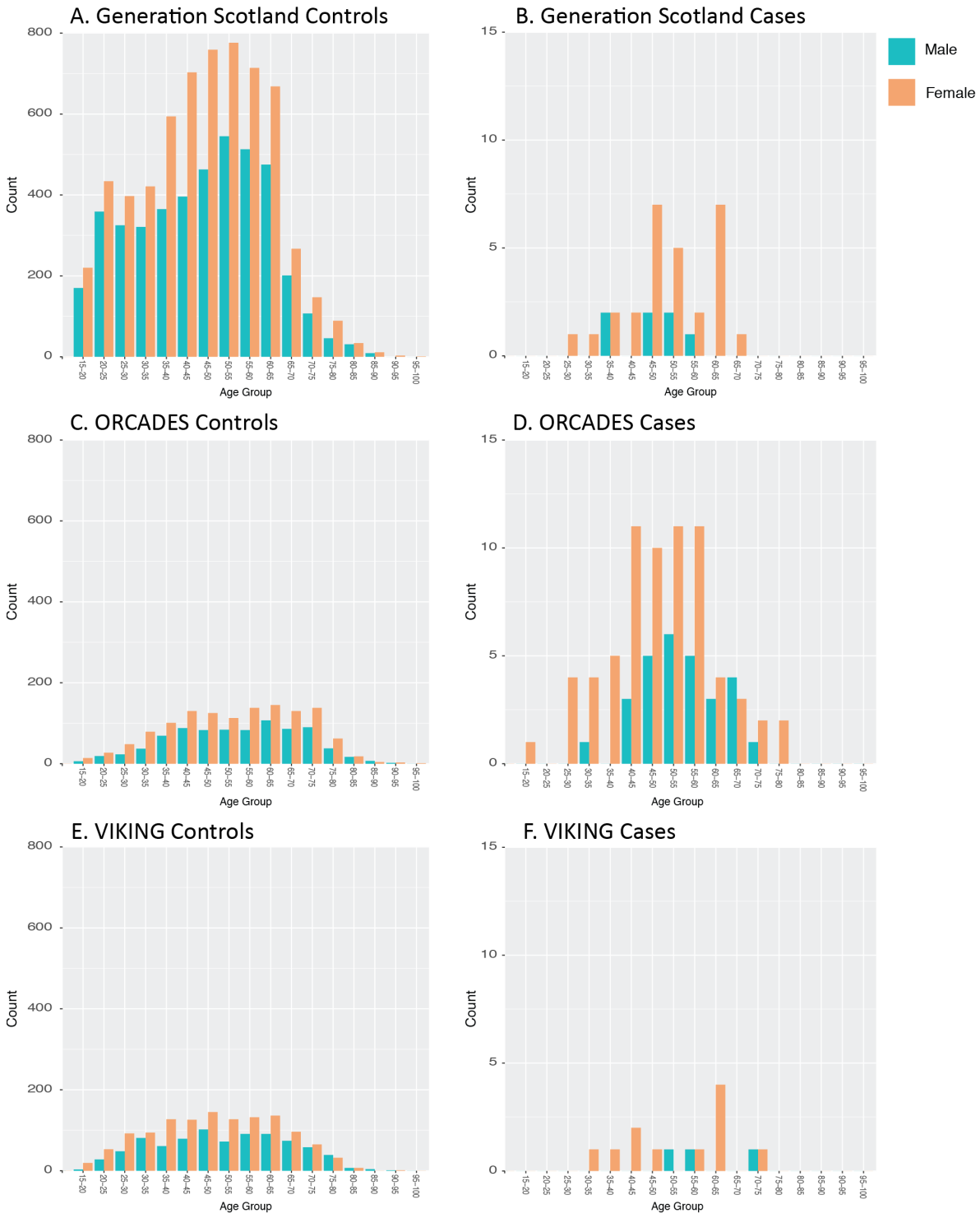


Figure 3: Age distribution plots in Generation Scotland, ORCADES and VIKING

Age distribution plots for Generation Scotland, ORCADES and VIKING, split by MS status; total number of individuals within each group can be found in Table 5. Note that cases and controls are on separate x-axis scales (controls: 0-800, cases: 0-15).

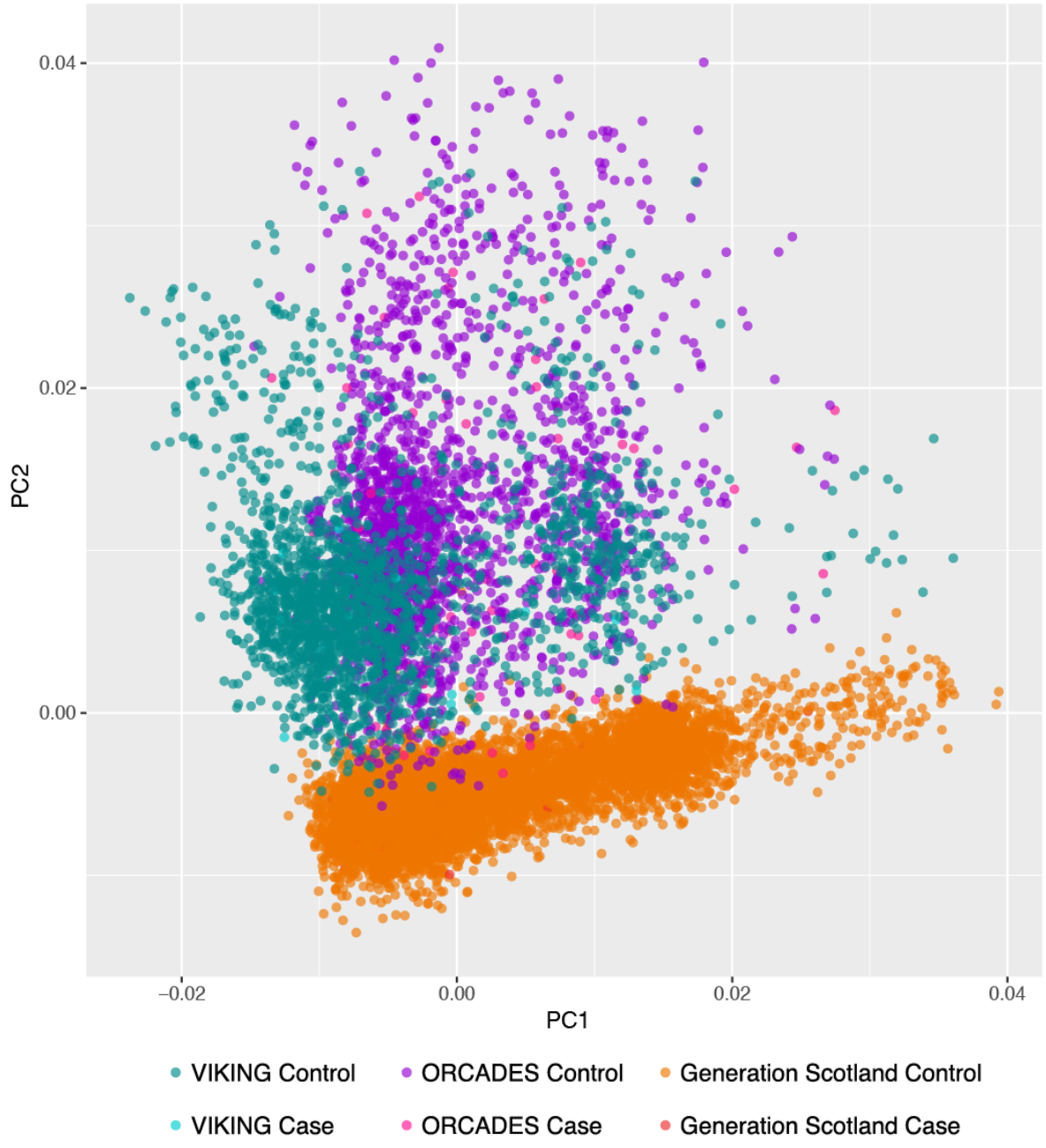


Figure 4: Principal component plot for VIKING, ORCADES and Generation Scotland
Principal component plot containing individuals from Orkney Complex Disease Study (ORCADES) (97 cases / 2118 controls), Viking Health Study – Shetland (VIKING) (15 cases / 2090 controls) and Generation Scotland (30 cases / 8708 controls) cohorts, using PC1 and PC2. Multiple sclerosis cases and controls are plotted in separate colours.

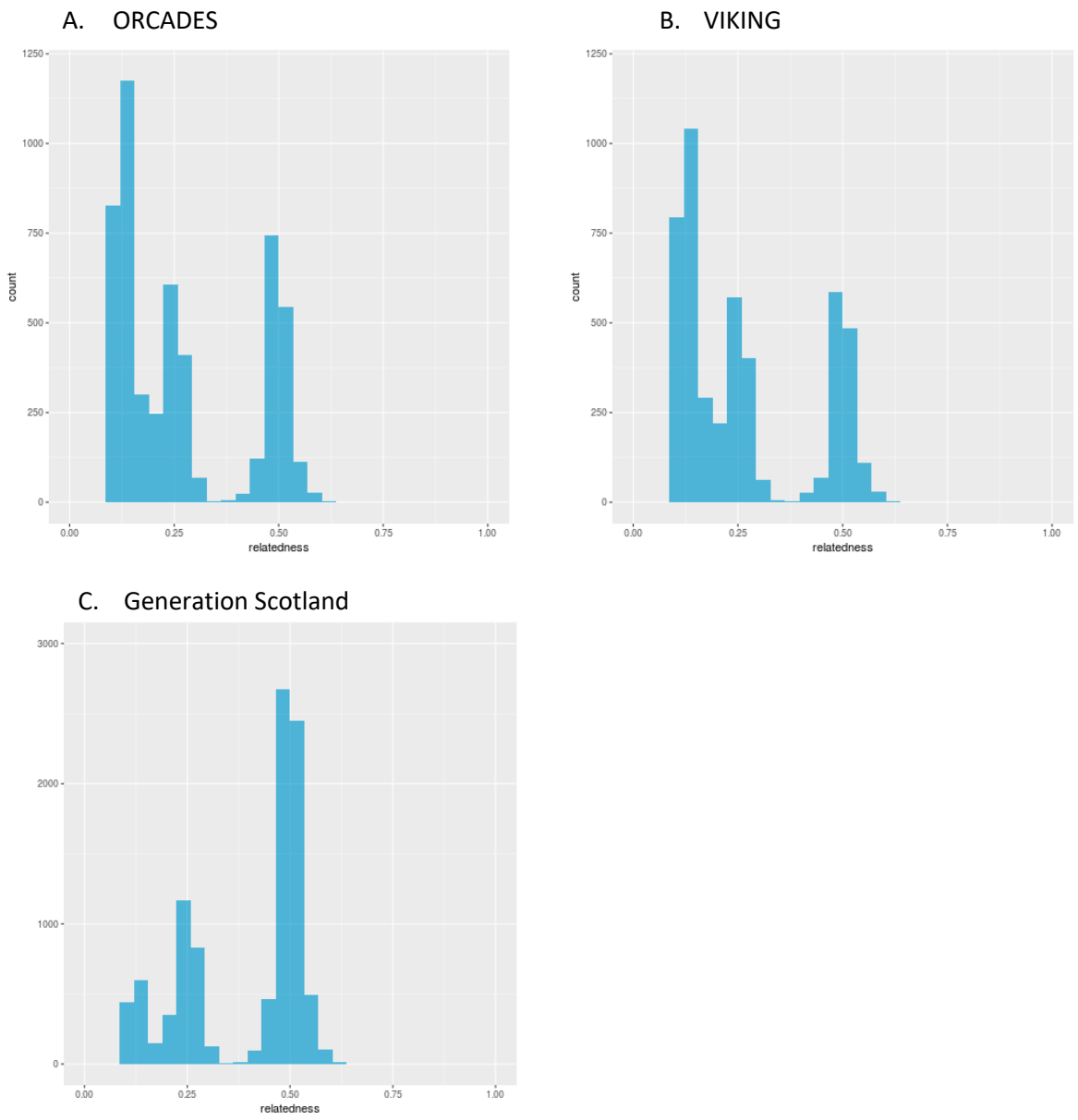


Figure 5: Relatedness in ORCADES, VIKING and Generation Scotland

Relatedness coefficients for pairs in ORCADES, VIKING and Generation Scotland. For clarity, this plot has been restricted to including individuals with relatedness coefficient > 0.10, as the majority of individuals below this point will be unrelated (Turner et al., 2011). Therefore only 5222 pairs are shown out of 2,449,791 in ORCADES, 4702 pairs out of 2,368,576 in VIKING and 9998 out of 38,171,953 in Generation Scotland.

2.3 Discussion

Data for ORCADES, VIKING and Generation Scotland was presented to me at the beginning of this thesis, having undergone stringent quality control procedures at both a sample and SNP level for genotypes and phenotypes. I carried out several sense checks on this data, which included counting the number of MS cases and controls and calculating their mean age, plotting principal component values and plotting relatedness between pairs of individuals. The motivation to carry out these checks was to determine the quality of data, and to ensure that the quality procedures had produced a reliable and accurate data set for all three cohorts. To this extent, my sense checks confirmed this; no unknown abnormalities were revealed in the data, and individuals in both case and control groups appear to be in similar age groups with a mostly normal distribution. This is particularly important for diseases such as MS that have a mean age of onset that occurs later in life; in the case of MS this is 30 years old (Hauser and Oksenberg, 2006). Although there was no distribution bias of cases in the principal component plots, there appears to be some underlying population structure within ORCADES and VIKING; both ORCADES and VIKING do not appear as one homogenous cluster. ORCADES had a largely scattered appearance with one principal cluster, while VIKING had one main cluster and one secondary smaller cluster. Generation Scotland had an oblong cluster, which was separate from the Northern Isles groups (although a small number of individuals from the Northern Isles overlapped with this group). The scattering and separate clusters within ORCADES and VIKING may be due to different island groups in each dataset. This uneven data clustering, combined with the high number of related pairs within the data, suggests it is therefore important to correct for relatedness in downstream analyses that would be influenced by underlying population structure (such as genome wide association analyses). Batch effect has previously been checked; relative pairs share exactly what is expected, therefore it is unlikely batch effect exists.

In general, there are several advantages and disadvantages to using ORCADES and VIKING. Both datasets are the largest collection of genetic information from each group of islands; therefore, they allow an insight into these populations that is not possible from phenotypic data alone. However, as both islands have relatively small populations, the datasets will not yield a high number of MS cases, despite the

prevalence of MS within the islands being among the highest worldwide. Although the small case numbers will be a major limitation when carrying out subsequent analyses in this study, both datasets have high quality genotypic and phenotypic information, and are most likely the best genomic dataset these regions will have in the foreseeable future. Therefore, full advantage must be taken of the information available.

2.4 Conclusion

Quality control of genetic data is important for several reasons. Primarily, it ensures that the data can be used to produce accurate information in subsequent analyses and allows researchers to successfully interpret results to a reliable conclusion. Carrying out quality control procedures is a balance between retaining as much information as possible and retaining information that will be of high quality. The methods discussed in this chapter have been designed to maintain this balance, and as such, they have processed both ORCADES and VIKING into useable and accurate datasets which can be used confidently for the analyses in this thesis.

CHAPTER 3: HERITABILITY OF MULTIPLE SCLEROSIS IN THE NORTHERN ISLES OF SCOTLAND

Heritability, or how much variation in a trait is influenced by genetics, is the cornerstone to genetic investigations of a trait. Here, an estimate of heritability for Multiple Sclerosis is established for Orkney (case numbers for Shetland were too few to obtain an accurate estimate). Although heritability estimates for MS have been widely reported in the literature, heritability is specific to a population and time period, and it is therefore useful to determine an estimate for the key populations of this project.

3.1 Introduction

3.1.1 What is heritability?

Determining the genetic contribution to phenotypic variation is an important part of investigative disease studies, particularly for diseases such as Multiple Sclerosis that have a higher occurrence within families. Within the general European population, the risk of developing MS is about 0.3% (O'Gorman *et al.*, 2013). However, this is much higher within families. Second- and third-degree relatives of MS patients have a risk of ~0.5%, first-degree relatives have a risk of ~1-3%, and identical twins have a ~17% risk (O'Gorman *et al.*, 2013). These findings suggest that a proportion of variation in MS risk is due to genetic factors, which have been quantified in numerous studies and will be discussed further in this chapter (Ristori *et al.*, 2006; Patsopoulos *et al.*, 2011; Westerlind, Ramanujam, *et al.*, 2014; Patsopoulos and (IMSGS), 2016; Baranzini and Oksenberg, 2017).

The proportion of variation in a trait that can be explained by genetic factors is known as heritability (Falconer and Mackay, 1996). As heritability is a proportion, its value lies between 0 and 1. If a trait has a heritability of 0, genetics do not explain any variation in the trait and all variation comes from the environment. If a trait has a heritability of 1, genetics explain all variation in the trait and the environment has no effect on trait variation. However, these values of heritability are specific to a population at any given time period, as both genetic and environmental variance can change over time and population (Mayhew and Meyre, 2017). For example, two populations may have

differing frequencies in effect alleles that influence the trait; if one of the populations has less variation in allele frequencies between its population members, it will have to less variation in the trait due to genetics and therefore have a lower heritability estimate than the other population. In another example, two populations that have similar genetic variation but differ in the amount of environmental variance present will have different heritability estimates. Heritability can therefore be described as measurement specific to a time period and population that gives an indication of what influences a trait more: genetics or environment.

In simple terms, the relationship of genetics and environment in relation to a phenotype can be expressed as follows (Falconer and Mackay, 1996):

$$\boldsymbol{P = G + E + GE} \quad \textbf{(1)}$$

where $\boldsymbol{P}$ is the phenotype,

$\boldsymbol{G}$ is the genotype,

$\boldsymbol{E}$ is the environment and

$\boldsymbol{GE}$ is the genotype and environment interactions.

The variance of the phenotype P can then be written as (Falconer and Mackay, 1996):

$$\boldsymbol{\sigma_P^2 = \sigma_G^2 + \sigma_E^2} \quad \textbf{(2)}$$

where $\boldsymbol{\sigma_P^2}$ is the phenotypic variance,

$\boldsymbol{\sigma_G^2}$ is the genetic variance and

$\boldsymbol{\sigma_E^2}$ is the environmental variance (which includes variance due to gene-environment interactions).

The values of these components can then be determined to obtain the heritability of the trait. There are several distinct types of heritability, each corresponding to the genetic components that contribute to it.

The most encompassing type of heritability is broad-sense heritability (H^2), which is the variance in a trait explained by all genetic factors. This is represented by the equation (Falconer and Mackay, 1996):

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} \quad \textbf{(3)}$$

where H^2 is the estimate of broad sense heritability,

σ_G^2 is the variance in the trait explained by genetics (G) and

σ_P^2 is the total variance in the trait within the population (P).

This estimate of heritability makes no assumptions about the underlying genetic structure, i.e. the variance could come from hundreds of small effect variants or one large effect variant (de los Campos *et al.*, 2015). The genetic variation may come from additive genetic effects, dominant/recessive genetic effects and/or interactive genetic effects. Additive genetic effects, where the sum of the effect of a group of alleles is equal to the sum of their individual effects, are typically the largest contributor to genetic variation within complex traits (Hill, Goddard and Visscher, 2008). Hill, Goddard and Visscher examined empirical evidence for genetic variation from several species (including humans, using twin studies) and found that additive variation typically accounted for more than half of the total genetic variance. Dominant/recessive effects (where the effect of an allele is masked by a second allele) and interactive effects (where the effect of an allele is dependent on the presence of one or several modifier alleles) also contribute to genetic variation between individuals, however they play a smaller role in heritability estimates than additive effects (Zhu *et al.*, 2015). Similar findings were found by Zhu et al, who examined the contribution of dominance effects to genetic variation. They analysed 79 quantitative traits in 6715 European Americans and found that dominance genetic variance was around a fifth of that of additive genetic variance

for the majority of traits (Zhu *et al.*, 2015). The extent to which interaction effects like epistasis contribute to genetic variation is not fully understood but is thought to be small (Carlborg and Haley, 2004).

As additive, dominant/recessive and epistatic effects all contribute towards the variation of a trait that can be explained by genetics, the variance in the trait explained by genetics can be broken down into its parts and written mathematically as the following (Falconer and Mackay, 1996):

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \quad \textbf{(4)}$$

where σ_G^2 is the variance in the trait explained by genetics (G),

σ_A^2 is the variance in the trait explained by additive genetic effects (A),

σ_D^2 is the variance in the trait explained by dominant/recessive genetic effects (D), and

σ_I^2 is the variance in the trait explained by interaction genetic effects.

Of the three components of genetic heritability, additive variance is the most predictable form of variance, as it is passed from generation to generation. The amount of variation due to additive variance is known as narrow sense heritability (h^2). This is represented by the following equation (Falconer and Mackay, 1996):

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \quad \textbf{(5)}$$

where h^2 is the narrow sense heritability,

σ_A^2 is the variance in the trait explained by additive genetic effects (A) and

σ_P^2 is the total variance in the trait within the population (P).

True narrow sense heritability is smaller than true broad sense heritability as narrow sense heritability does not include dominant/recessive or interactive effects, whereas broad sense heritability does include these effects, if present.

Further types of heritability are SNP heritability (h_g^2) and GWAS heritability (h_{GWAS}^2). SNP heritability measures the contribution of all common SNPs to variation in a trait (Yang *et al.*, 2010), while GWAS heritability measures the contribution of those measured SNPs which are genome-wide significant in a specific study. The SNP heritability estimate is calculated on unrelated samples and it is independent from sample size. As GWAS heritability is dependent on variants crossing the prescribed significance threshold, it is therefore dependent on sample size. Estimates of GWAS heritability can therefore in theory approach estimates of SNP heritability if a large enough sample size has been gathered. As sample sizes increase, the power to detect small effect sizes also increases and thus novel common variants which may not have been detected with a small sample size may become significant and contribute to the GWAS heritability estimate.

Estimates of SNP heritability have become more accurate over time given the recent advances in DNA technology. Not only are SNP arrays psychically genotyping a huge number of common variants, but the effect of variants in LD with these is able to be captured. The genetic contribution of these measured SNPs (S) can therefore be estimated as SNP heritability (Yang *et al.*, 2010):

$$\boldsymbol{h_g^2 = \frac{\sigma^2_{SNPs \epsilon S}}{\sigma_P^2}} \qquad \textbf{(6)}$$

where $\boldsymbol{h_g^2}$ is the SNP-heritability,

$\boldsymbol{\sigma^2_{SNPs \epsilon S}}$ is the variance in the trait explained by the additive effects of common SNPs and

$\boldsymbol{\sigma_P^2}$ is the total variance in the trait within the population (P).

The previous measures of heritability described each capture different levels of genetic contributions to a trait. The relationship between these heritability estimates can therefore be defined as follows (Manolio et al., 2009):

$$H^2 \geq h^2 \geq h_g^2 \geq h_{gwas}^2 \quad (7)$$

where H^2 is broad-sense heritability (H^2) which captures all variance in a trait due to genetics,
h^2 is narrow-sense heritability which captures variance in a trait due to additive genetic variants alone,
h_g^2 is SNP heritability which captures variance in a trait due to SNPs and
h_{gwas}^2 is GWAS heritability which captures variance in a trait due to statistically significant SNPs.

3.1.2 Missing heritability

When calculating heritability for complex traits, it is important to be aware of the problem of "missing heritability", i.e. not all heritability has been explained by measured variants alone (Manolio *et al.*, 2009). A classic example of this in complex traits is height. Initial family-based studies measured the heritability of height as approximately 0.80 (Silventoinen *et al.*, 2003; Macgregor *et al.*, 2006; Visscher, Hill and Wray, 2008). Narrow sense heritability for height was then calculated as slightly lower than broad sense heritability, with an estimate of 0.69 (0.67 – 0.71) from Zaitlen *et al.*, 2013. The authors of this study suggest that this estimate may be upwardly biased due to non-additive effects and common environmental effects. However, when SNP-heritability estimates for height were calculated, only ~0.45 of variance was explained, with 50 significantly associated SNPs accounting for only ~0.05 (Yang *et al.*, 2010). Thus, a proportion of heritability appeared to be "missing". A gap in heritability estimates is also seen in Multiple Sclerosis. Broad sense heritability measured using twin studies was estimated at 0.64 (with a 95% confidence interval of 0.36-0.76) (Westerlind, Ramanujam, *et al.*, 2014). However, the most recent IMSGC (14,802 cases, 26,703 controls) calculated SNP heritability to be 0.19 (95% CI 0.18, 0.20)

(International Multiple Sclerosis Genetics Consortium et al., 2017). Although SNP heritability is a proportion of broad-sense heritability and so it is expected that it will be smaller, this leaves a large gap between the two estimates, even given the confidence intervals. There are several reasons as to why gaps appear between heritability estimates, particularly when comparing pedigree studies to genotype studies.

Firstly, broad sense heritability estimates calculated through familial studies using twins and siblings may be inflated. These studies often disregarded gene/environment interactions and could underestimate common familial environment, which result in an overinflated heritability estimate.

A gap may also exist between narrow sense heritability estimates and SNP heritability estimates as SNP heritability is dependent on array design and does not account for rare SNPs (Sandoval-Motta et al., 2017). Only common SNPs are used to calculate the genetic relationship matrix, and so rare variants are automatically not accounted for. This can be understood through the difference in calculating the two estimates. Narrow sense heritability estimates are calculated using a related population with a genetic relationship matrix (GRM) used to denote the relationship between each pair of individuals. These matrices are constructed by identifying allele sharing in measured loci throughout the genome, to give an estimate of the actual proportion of the genome that is identical by descent across individuals (Stanton-Geddes et al., 2013). In contrast, SNP heritability estimates are calculated on unrelated populations. The inclusion of related individuals for calculating narrow sense heritability allows information about allelic correlations across the whole genome to be inferred. For example, if an individual has allelic correlations with a parent on chromosome 1, then it can be inferred that they will be correlated on their other chromosomes: that half their genome is shared. Even if only common variants were assayed, it would be able to be inferred that you would also share half the rare variants. In other words, the relatedness of the common SNPs would be predictive of the rare SNPs. Therefore, by including a GRM and related individuals in calculating a heritability estimate, a full estimate of additive variation can be calculated.

When an unrelated population is used as in the case with calculating SNP heritability, the GRM is only measuring local allelic sharing (i.e. sharing of that particular SNP and the SNPs in LD with that SNP). Therefore, SNP heritability is far more dependent on

array design and rare SNPs are not measured, resulting in smaller SNP heritability estimates.

Another gap exists between SNP heritability and GWAS heritability. This is due to a combination of small effect size and sample size. GWAS are dependent on sample size, as small effect variants require a larger sample size to be detected and cross a significance threshold. Therefore, small-effect variants often remain hidden, although GWAS heritability has, in theory, the potential to approach SNP heritability given a large enough sample size.

3.1.3 Measuring heritability in binary traits

Many of the assumptions when calculating heritability rely on the trait being continuous. For binary traits, where individuals are either affected or not affected by a trait with no intermediate state, the variance calculation becomes more challenging.

To calculate heritability estimates for this type of phenotype, a threshold model was developed. This uses an underlying measurement of liability to determine if an individual will obtain disease status or not (Gottesman and Shields, 1967). Liability collectively describes all genetic and environmental factors that contribute to developing a binary trait and can be estimated using a group of individuals to plot a standard distribution curve (Figure 6). At some point individuals will cross a certain liability threshold and they will be affected by the trait. All individuals will have the same liability threshold level; however individuals will be more likely to exceed the threshold level depending on their exposure to environmental risk factors and their genetics (Lee *et al.*, 2011).

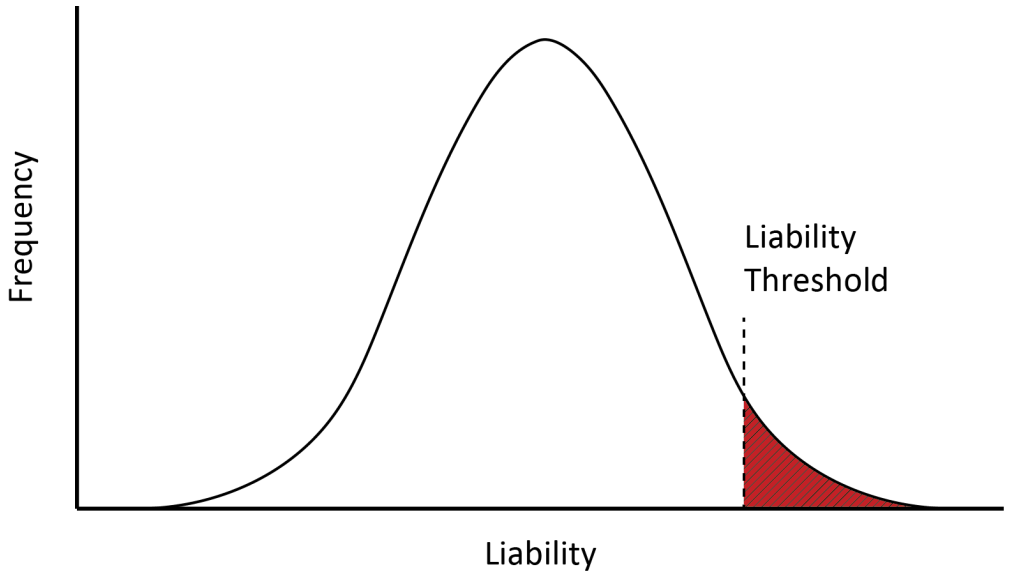


Figure 6: Liability threshold model

The liability threshold model, which describes the distribution of individuals with a. binary trait; the red area indicates individuals affected by the binary trait, with the dashed line indicating the threshold which needs to be crossed to obtain affected status.

In calculating the heritability estimate for a binary trait, the observed heritability is first calculated and is used alongside the population prevalence and sample prevalence to estimate the heritability of the liability distribution. Observed heritability (or heritability on the observed scale) is the ratio of phenotypic variance due to additive effects (as described in Equation 5). Converting heritability on the observed scale to heritability on the liability scale is done using the following formula (Lee *et al.*, 2011):

$$h^2_{liability} = h^2_{observed} \frac{K(1-K)}{\varphi(\Phi^{-1}[K])^2} \frac{K(1-K)}{P(1-P)} \quad \textbf{(8)}$$

where $h^2_{liability}$ is the heritability on the liability scale,

$h^2_{observed}$ is the heritability on the observed scale,

K is the frequency of the binary trait within the population,

P is the frequency of the binary trait in the observed data, and

$\boldsymbol{\varphi}(\boldsymbol{\Phi}^{-1}[\boldsymbol{K}])$ is the probability density function, evaluated at the $\boldsymbol{K}$ quantile of the inverse cumulative density function of the standard normal distribution.

The heritability estimate is calculated as quantitative on an observed scale, and then subsequently converted into the liability scale. As the liability scale calculation considers sample and population prevalence, it allows for comparisons across populations to be made.

3.1.4 Research aims

Establishing estimates of heritability provides a quantification of the burden of disease variance attributable to genetic factors. Estimates of broad sense heritability of MS have been estimated using twin studies as 0.64 (with a 95% confidence interval of 0.36 - 0.76) (Baranzini and Oksenberg, 2017), and SNP heritability estimates using over 200 MS risk genes have been estimated at 0.19 (International Multiple Sclerosis Genetics Consortium et al., 2017). An estimate of heritability in the Northern Isles has not yet been published, and so this chapter seeks to investigate if the heritability estimates for MS in Orkney and Shetland are of a similar value to those published in the literature.

3.2 Methodology

SNP heritability estimates were calculated separately for ORCADES and VIKING using the software package Genome-wide Complex Trait Analysis (GCTA), v1.91.7 beta (Yang *et al.*, 2011). The datasets were those described in **Chapter 2**.

GCTA was first used to estimate genetic relationships between individuals and construct a GRM. Individuals with a relationship coefficient greater than 0.05 were removed from the analysis. GCTA was then used to conduct a mixed linear model analysis of variance explained by SNPs. GCTA works by fitting the effects of all SNPs (with a MAF greater than 0.05) in the analysis as random effects using a mixed linear model. The model can be written as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{\varepsilon}$$

(9)

where $\boldsymbol{y}$ is an n x 1 vector of phenotypes (in this case MS status) and n is the sample size,

$\boldsymbol{X}$ is an incidence matrix for fixed effects,

$\boldsymbol{\beta}$ is a vector of fixed effects (for ORCADES: sex, age and array; for VIKING: sex and age),

$\boldsymbol{g}$ is an n x 1 vector of the total genetic effects of the individuals and

$\boldsymbol{\varepsilon}$ is a vector of residual effects.

The variance in **y** can be denoted as follows:

$$\boldsymbol{V} = \boldsymbol{A}\sigma_{\boldsymbol{g}}^2 + \boldsymbol{I}\sigma_{\boldsymbol{\varepsilon}}^2$$

(10)

where $\boldsymbol{V}$ is var (**y**),

$\boldsymbol{A}$ is the genetic relationship matrix,

$\sigma_{\boldsymbol{g}}^2$ is the variance explained by all included SNPs,

$\boldsymbol{I}$ is an $n \times n$ identity matrix and

$\sigma_{\boldsymbol{\varepsilon}}^2$ is the variance explained by residual effects.

Using the above equation, SNP heritability was estimated using a genome-based restricted maximum likelihood (GREML) approach; a type of maximum likelihood estimation where the parameters of a statistical model are estimated given a set of observations (Patterson and Thompson, 1971). GCTA uses an AI-GREML approach, which uses less computational power than standard GREML calculations (Gilmour, Thompson and Cullis, 1995). The GREML approach is useful as it is not biased by sample size or individually significant associations.

The estimate of variance that was produced was on the observed scale, however GCTA used prevalence information to convert this to a liability scale using the linear transformation described in the chapter introduction. Prevalence information is necessary to account for ascertainment bias that is typically found in case-control studies. For the ORCADES calculation, prevalence information of 0.00402 was used, and for VIKING the prevalence 0.00295 was used (Visser *et al.*, 2012).

3.3 Results

SNP heritability estimates were calculated for both Orkney and Shetland. Orkney had an estimate of 0.307 (95% CI 0.129, 0.485), while the very small number of cases in Shetland resulted in an estimate of 0, with confidence intervals spanning the full gamut from not at all heritable to completely heritable (95% CI 0, 1). Heritability estimates from key sources in the literature were gathered as a comparison (Table 6). These sources include estimates of broad-sense heritability (H^2) from a twin study and sibling study (Ristori *et al.*, 2006; Westerlind, Ramanujam, *et al.*, 2014), and estimates of SNP heritability (h_g^2) from the most recent and largest IMSGC study (International Multiple Sclerosis Genetics Consortium *et al.*, 2017; Zheng *et al.*, 2017). These were compared to the estimates produced in this thesis (
Figure 7 and Figure 8).

Estimates of broad heritability (
Figure 7) are higher than those of SNP heritability (Figure 8) which would be expected, with the twin study estimate at 0.48 and the sibling study estimate at 0.64. However, the confidence intervals for these are wide, particularly the twin study estimate (95% CI 0.06, 0.86). The SNP heritability estimate had much smaller confidence intervals; the largest IMSGC study gives an estimate of 0.19 (95% CI 0.18, 0.20).

The heritability analysis conducted in this study for Shetland was underpowered, and as such the SNP heritability estimate produced in this study for Shetland is not meaningful. The SNP heritability for Orkney is higher than that of the consortium estimate, however the confidence intervals for the Orkney estimate overlaps that of the consortium study estimate.

Study type	Heritability Type	Heritability Estimate (95% CI)	Study size
Twin study* (Ristori *et al.*, 2006)	H^2	0.48 (0.06, 0.86)	216 twin pairs
Sibling study** (Westerlind, Ramanujam, *et al.*, 2014)	H^2	0.64 (0.36, 0.76)	74,757 twin pairs (containing 315 MS cases) and 2.5 million sibling pairs
IMSGC Consortium (International Multiple Sclerosis Genetics Consortium *et al.*, 2017) ***	h_g^2	0.19 (0.18, 0.20)	14802 cases and 26703 controls
ORCADES (current study)	h_g^2	0.31 (0.13, 0.49)	97 cases and 2118 controls

* The environmental contribution for this estimate was 0.29 (95% CI 0, 0.60) for shared environmental factors and 0.23 (95% CI 0.12, 0.39) for individual environmental factors
** The environmental contribution for this estimate was 0.01 (95% CI 0, 0.18) for shared environmental factors and 0.35 (95% CI 0.24, 0.51) for individual environmental factors
*** The super-extended MHC (chromosome 6, ~24M-35M) explained 21.4% of the overall h_g^2 estimate

Table 6: Heritability estimates from published and current study

Heritability results published from the major sources of MS research, including sibling and twin studies and the largest IMSGC study (as of 2018), alongside the estimate from this thesis.

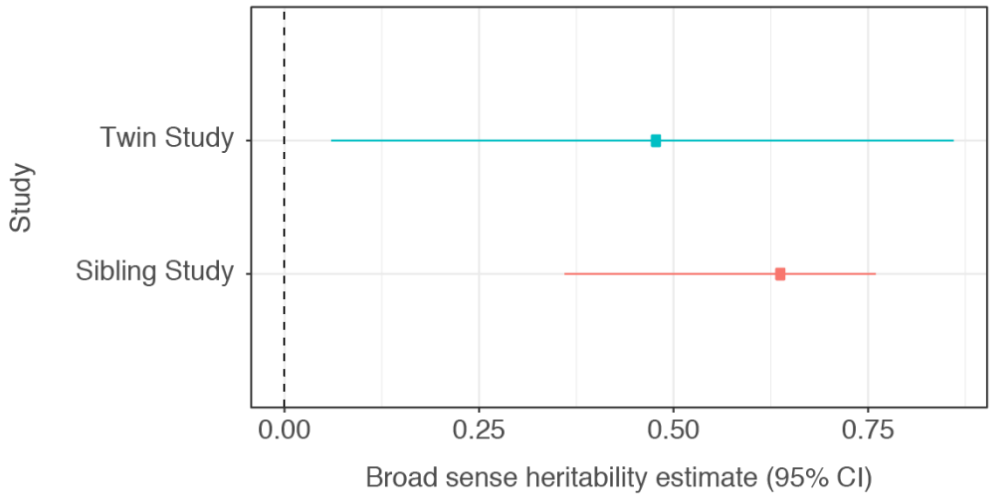


Figure 7: Comparison of broad sense heritability estimates (H^2)

Published broad sense heritability results, using a twin study (216 twin pairs; Ristori 2006) and a sibling study (74757 twin pairs (containing 315 MS cases) and 2.5 million sibling pairs; Westerlind, Ramanujam, et al., 2014a).

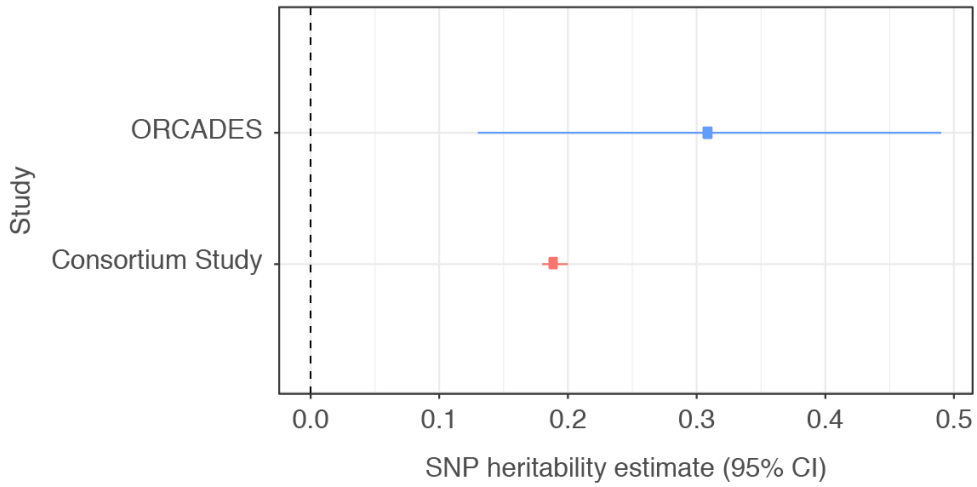


Figure 8: Comparison of SNP heritability estimates

Published SNP heritability results from IMSGC (14802 cases and 26703 controls; (International Multiple Sclerosis Genetics Consortium *et al.*, 2017)) compared with the ORCADES heritability estimate.

3.4 Discussion

Summary of Findings

SNP heritability was calculated for Orkney and Shetland using the ORCADES and VIKING cohorts, with the aim of quantifying the burden of Multiple Sclerosis variance that was attributable to genetic factors. As Multiple Sclerosis has a particularly high prevalence in Orkney (and to a lesser extent, Shetland), it is possible that excessive environmental or genetic burdens may be influencing disease prevalence in this population. Thus, it was important to determine if the heritability of MS in the Northern Isles was significantly higher or lower than the estimates published in current literature. It was found here that the SNP heritability in Orkney was 0.307 (95% CI 0.129, 0.485), while an accurate estimation of SNP heritability for Shetland was unable to be obtained due to small case numbers. The Orcadian SNP heritability estimate is higher than that predicted by the IMSGC consortium, however as confidence intervals overlap for the IMSGC estimate, it is not significantly different. Based on the limited evidence in this study, we give no evidence to indicate that our findings differ from those previously published. However, this study was somewhat underpowered so the possibility that Orkney has a higher heritability than other populations cannot be

excluded. Any difference seen in the heritability of MS in Orkney compared to the heritability of MS in other populations could be accounted for by chance, given the confidence intervals.

Although the findings here were not significant, a broader discussion can still be had regarding the variability of heritability estimates for MS. Four key studies of MS heritability were highlighted in this chapter: these estimates of heritability include two broad-sense heritability estimates (H^2) and two SNP heritability estimates (h_g^2).

Broad-heritability estimates

Two of the key twin and sibling study findings which give broad-sense heritability estimates are those by Ristori et al (2006) and Westerlind et al (2014). Ristori et al estimated H^2 using twin pairs within an Italian population while Westerlind estimated H^2 using twin pairs and siblings within a Swedish population. Both these studies have obtained estimates of broad heritability which at first glance, appear to show heterogeneity between the two populations: the Italian study (Ristori *et al.*, 2006) estimates MS H^2 as 0.48, while the Swedish study estimated MS H^2 as 0.64. However, if the confidence intervals are included, there is no significant difference between the estimates (Westerlind, Kuja-Halkola, *et al.*, 2014). In general, the low prevalence of MS has resulted in the twin study lacking power and causing large confidence intervals in comparison to the Swedish study which additionally included siblings. A key advantage of this type of study design is that it can capture the total genetic (additive, dominance and epistasis) and environmental (shared and non-shared) effects that contribute to variation in a phenotype (Røysamb and Tambs, 2016): these studies provide a "maximum" estimate of heritability of which SNP heritability makes up a proportion. Thus, the SNP heritability estimates produced in the key published studies and the estimate produced in this study, are in line with the H^2 estimates.

However, there are several limitations with twin and sibling studies. Most obviously, these estimates of heritability fail to identify any specific genes or environments that influence the estimate. Studies estimating SNP heritability and narrow heritability may have smaller estimates of heritability than twin and sibling studies, leading to the "missing heritability" gap. It is unlikely that twin studies are invalid, but rather that a combination of factors result in the heritability gap: extremely large sample sizes are needed to provide sufficient power to identify small effects (resulting in smaller

estimates of narrow-sense and SNP heritability), large confidence intervals in the twin and sibling studies and the presence of non-additive genetic effects contributing a small amount to heritability (Goldman, 2014). Additionally, twin studies cannot tell shared environmental variance from genetic variance, and this can result in an overestimation of heritability. The confidence intervals for the broad-sense and SNP heritability studies overlap, so it is difficult to quantify and estimate of missing heritability for MS.

SNP heritability estimates

A key SNP heritability estimate was included as comparison to the SNP heritability estimate produced in this study: the IMSGC estimated SNP heritability of 0.19 (95% CI 0.18, 0.20). The estimate provided from the consortium is comparable for the Orkney estimate, as they were both produced using the same program (GCTA). When compared to the Orkney SNP heritability estimate, the IMSGC estimate is lower and more accurate (lower CI) estimate, due to its larger sample size with 14802 MS cases. Although the Orkney estimate of SNP heritability is higher, overlapping confidence intervals suggest that the results in this study for Orkney are in line with these results.

Finding Implications

There are currently no published studies indicating the heritability of MS in Orkney. The findings here indicate that Orkney does not appear to have an excessively high or low heritability and the heritability estimates of the Orcadian population is in line with those in other European populations. However, region-specific heterogeneity in MS heritability cannot be ruled out due to the limited sample size of this study.

The findings here suggest that variation in genetics and environment in Orkney are both important in determining disease status. Although this is not new knowledge, it is important confirmation to the people of Orkney that both play an important role in disease variation.

Study Limitations

There has been some discussion regarding the accuracy of GCTA (Krishna Kumar et al., 2016), however the consensus appears that GCTA is a well-used and well-proven software tool that provides accurate estimates of heritability, given model assumptions are met (Speed et al., 2012; Gusev, 2015).

The largest caveat in this study is the small number of cases when performing heritability analysis, particularly for Shetland where an accurate estimate of heritability was unable to be obtained. It would be interesting to repeat this study in future providing more genetic data from individuals suffering from MS within the island populations has been gathered.

3.5 Conclusion

This chapter estimated the SNP heritability of Orkney to be 0.307 (95% CI 0.129, 0.485). An accurate SNP heritability estimate for Shetland was not able to be obtained due to a small number of cases.

The SNP heritability estimate for Orkney was based on a limited sample size and did not indicate a statistically significant difference from previously published results. Although this appears to indicate that there are not any excessive environmental or genetic burdens causing the high prevalence on the islands, the possibility of this cannot be excluded at this stage until a larger study is conducted.

Several types of heritability estimates exist, including broad-sense, narrow-sense, SNP heritability and GWAS heritability. For MS, an accurate estimation of SNP heritability has been obtained in the largest MS study to date, however the estimates for broad-sense heritability appear less precise in comparison. It is more possible that twin and sibling studies are likely to yield biased results due to the nature of these studies (inflation through unaccounted gene/environment interactions and underestimation of common familial environment). However, they provide a rough estimate as to the total genetic contribution to trait variation. Both types of heritability estimate provide useful insight regarding what influences a trait, genes or environment, and at what proportion.

However, a gap between the heritability estimates for broad sense and SNP heritability clearly indicate that missing heritability is an issue in Multiple Sclerosis. SNP heritability does not include some genetic factors that may contribute to MS, including rare variants and dominant/recessive effects. Additionally, Multiple Sclerosis is a multifactorial disease involving both genetic and environmental factors, and so the contribution of interactive effects may also contribute to the missing heritability.

As genetic studies become larger and more studies are carried out on the other genetic factors that contribute to MS, heritability estimates will move further to the true heritability of a trait, further improving the understanding of MS.

CHAPTER 4: GENOME WIDE ASSOCIATION STUDY OF MULTIPLE SCLEROSIS IN THE NORTHERN ISLES OF SCOTLAND

Within this chapter, genome wide association studies are performed using the ORCADES and VIKING datasets to determine if any novel common MS risk variants exist in the Northern Isles of Scotland. It is possible that this region has experienced the jackpot effect, where large effect MS risk variants rarer elsewhere, exist in the Northern Isles at a higher frequency through the effect of drift and low effective population size. If this is the case, it is possible that these variants contribute to the excess of MS risk in the region.

4.1 Introduction

4.1.1 What is a GWAS?

A genome-wide association study (GWAS) aims to identify common genetic variants that contribute to a specific trait (for example, disease risk) within population samples. It is a collection of simple regressions carried out on millions of SNPs across the genome to test for an association with a trait of intent. The overarching goal when conducting GWAS is to increase understanding of the biology of specific traits. When that trait is a human disease, it is the goal that this increased understanding will lead to improved prediction, prevention and treatment of the disease.

GWAS is a particularly useful method of analysis as detailed prior knowledge of physiology is unnecessary. When a GWAS is carried out, the genome is scanned using genome-wide SNP arrays; by considering all common variants equally, no hypothesis is required. If a variant in the genome is found to be significantly more frequent in people with the trait than without, it is said to be associated with that trait. This variant points to a region of linkage disequilibrium which influences the trait.

In the pre-GWAS era, genotyping was vastly more expensive and small-scale methods such as candidate gene and linkage studies were the primary method of genetic analysis. In comparison to GWAS, these studies typically had small sample sizes and included

limited numbers of variants (Long and Langley, 1999). These early studies in genetic disease research commonly focused on monogenic or oligogenic diseases, and required multiple individuals in each family tested at great expense (Pearson and Manolio, 2008). The results from these studies were frequently large effect variants which were often private to a family, and so the findings were not able to be extrapolated to a larger population where the variant would not be found. Often in candidate gene studies, multiple independent tests were carried out until a significant result was found. This multiple testing was not taken into account when determining the appropriate threshold of statistical significance; this often led to a failure to replicate results as the original finding was a false-positive obtained by chance (Hirschhorn *et al.*, 2002; Morgan *et al.*, 2007). There was an explicit need for analysis methods that focussed on multifactorial traits, that produced results at a higher genetic resolution (better pinpointing of the region of the genome that influenced the trait) and that did not require extensive family pedigrees. This came to focus following the completion of the human genome project and the HapMap Project in the early 2000s (International Human Genome Sequencing Consortium, 2001; Human Genome Sequencing Consortium, 2004; The International HapMap Consortium, Altshuler and Donnelly, 2005), which led to a vastly increased public database of SNPs. Coupled with rapid advances in technology that allowed the price of genotyping to drop, genetic analysis methods swiftly progressed to take advantage of the cheaper genotyping and thousands of individuals were able to be interrogated at a previously unreachable resolution covering every gene in the genome (Hirschhorn and Daly, 2005). Genetic analysis moved from hypothesis-driven testing at specific loci to hypothesis free, genome-wide testing, and the focus moved to variants of moderate effect sizes within complex, multifactorial traits.

4.1.2 Strengths, weaknesses, findings and prospects in GWAS

Strengths

GWAS have shown great success in aiding complex trait genetic research and improving our understanding of disease biology. However, with every genetic method there are advantages and disadvantages to its use.

In general, GWAS is a powerful method for detecting common variants associated with multifactorial diseases. They have proven that they are robust in identifying risk alleles, and subsequently novel genes and pathways. An identified gene or pathway can then go on to inform drug development trials, which have a higher chance of reaching a phase III trial (or above) with supportive genetic evidence: this could lead to large savings in an already vastly expensive industry (Nelson *et al.*, 2015). Identified associations have also proven to be very replicable, within populations and between populations (with an adequate sample size) (Marigorta and Navarro, 2013).

Practically, GWAS are good value for money, due to the low cost of genotyping arrays and standard analysis pipelines. If it is feasible to gather large enough sample sizes for common diseases or traits, then significant discoveries usually follow. Additionally, as population samples can be unrelated, GWAS avoids the difficulty and expense of recruiting family groups.

In recent years, GWAS have also proved fruitful in the production and public availability of summary statistics. Following publication of a GWAS, the summary association statistics (in the form of SNP effect sizes and their standard error or p-values) are often released into the public domain (Welter *et al.*, 2014). These summary statistics can be used for a number of subsequent analyses, including detecting new associations (Zhu *et al.*, 2016), estimates of SNP heritability (B. K. Bulik-Sullivan *et al.*, 2015) and refining disease prediction scores (Krapohl *et al.*, 2018).

Key Findings

There are several key findings which have arisen following over a decade of GWAS results that are important to highlight. Firstly, GWAS has shown that high polygenicity is found for most common diseases and complex traits (Gibson, 2018). For many traits, GWAS conducted on hundreds of thousands of unrelated individuals have discovered hundreds of loci (with thousands more suggestive associations), each contributing a fraction towards the genetic variance of that trait. Up to 5% of all common variants, and a greater percentage of genes, may associate with any one trait (Boyle, Li and Pritchard, 2017). A consequence of having multiple variants contributing towards a trait is that the proportion of variance explained by each variant is small. At an individual level, high trait polygenicity will mean a person will have a most likely unique combination of

alleles that both increase and decrease risk for that trait, given the high number of possible combinations that can occur.

GWAS has also proven less fruitful than originally expected for accounting for genetic variation, resulting in much discussion surrounding the missing heritability. However, the realisation that most genetic effects have very small effect sizes resulted in the conclusion that this heritability is hidden rather than missing; this highlights a key problem with many GWAS, that they are largely underpowered to detect these small effect variants (Gibson, 2010). In recent years, sample sizes have been growing with ever larger consortia and meta-analyses used to make new discoveries. For example, in 2009, a GWAS of schizophrenia with 3000 cases discovered the first locus associated with the disease (Purcell *et al.*, 2009). Five years later, cases numbers of 35,000 increased this number to 108 (Ripke *et al.*, 2014).

GWAS have also revealed the extent to which pleiotropy exists amongst traits: it is thought that the majority of functional variants influence more than one trait (Pickrell *et al.*, 2016). This finding was known from previous research (for example, Mendelian mutations are often associated with multiple phenotypes in affected individuals (Visscher *et al.*, 2017)). However, GWAS have shown the extent to which pleiotropy occurs across traits, with the same variants associated with multiple traits in different groups of individuals (Sivakumaran *et al.*, 2011). For example, several causal variants have been found across multiple autoimmune diseases (Ellinghaus *et al.*, 2016) AND genetic correlations have been found between a number of different traits using GWAS data (for example, anorexia nervosa and schizophrenia) (B. Bulik-Sullivan *et al.*, 2015). These findings suggest that traits and diseases should not be studied in isolation from one another, as pleiotropy may cause variants to impact different tissues or act at different times in an individual's life.

Within Multiple Sclerosis specifically, GWAS have progressed understanding of the disease. The first Multiple Sclerosis GWAS was published in 2007 by the International Multiple Sclerosis Genetics Consortium; with 931 cases, this study identified the first genome-wide association outside the MHC region (Hafler *et al.*, 2007). The variant identified lay in the first intron of *IL2RA* (OR: 1.25, $p=2.96 \times 10^{-8}$), a gene which encodes the interleukin-2 receptor α chain. Important for multiple immune-related pathways (Liao, Lin and Leonard, 2011), IL2RA is the target of the MS drug daclizumab

(Bielekova *et al.*, 2004). The current largest MS GWAS has now analysed 47,351 cases (with 68,284 controls) and identified 233 genome-wide loci that explain around 20% of MS genetic variation; each locus ranges in OR from 1.02 to 1.2 (Patsopoulos and (IMSGS), 2016).

In general, the findings that have arisen from the use of GWAS have influenced the way researchers think about the genetic architecture of traits and diseases, with the future looking to embrace huge sample sizes and holistic approaches when considering disease biology.

Weaknesses

Although the benefits of GWAS have been tremendous in progressing the understanding of disease biology, it is not a panacea for genetic research.

When examining a successful GWAS association, there is often not a straightforward or clear link between the identified SNP and causal gene or pathway. If a GWAS is successful, multiple common variants can be identified that are associated with a specific trait. However, these variants are not directly informative of the trait-influencing gene or mechanism, as the associated SNP is usually in LD with the causal variant, which may be some distance away in a different gene. Additionally, that gene's function may be unknown, particularly in the role it plays in disease development. In order to find the direct causal gene or pathway, follow up studies are required. Current methods applied to select variants include fine mapping to define GWAS hits (where probabilities of causality are assigned to candidate variants and these are connected to likely genes (Spain and Barrett, 2015)), and investigating putative functional SNPs through in vitro and in vivo experimentation to determine molecular mechanisms to identify target genes (Edwards *et al.*, 2013). Although laboratory methods have been developed to progress with GWAS findings, these are costly and time consuming (Claussnitzer *et al.*, 2015). GWAS is only the first step in a long and often expensive process of determining the biological underpinnings of a locus.

GWAS require large sample sizes to detect low frequency variants. In one association study, a signal is deemed significant if the p-value is less than a threshold of 0.05. When testing multiple SNPS (assumed to be more than 1 million), a correction on this threshold is applied (5×10^{-8}) to reduce the false positive rate. This strict threshold

means that adequate statistical power is critical to detect association, and so sample sizes need to be very large to allow for low frequency variants to be detected (Visscher *et al.*, 2017). The requirement for large sample sizes makes GWAS unsuitable for rare diseases, due to the difficulty of gathering a large enough population sample.

Another weakness of current GWAS publications is most published data are from populations of European descent. This is in part due to circumstance (a large percentage of research institutes and funders lie within Europe and North America, and so research populations in these regions), however recent attention to this issue is encouraging the future of GWAS to expand to populations in other regions (Huffman, 2018).

Prospects

The future years of GWAS will hopefully address several of the weaknesses discussed here. Firstly, it is hoped that studies will expand to focus on populations of non-European descent. By including more diverse populations in genetic studies, understanding of variation in complex disease will be improved, along with better knowledge for personalised medicine. For example, populations have different allele frequencies and LD structures; this can be used to help fine-map causal variants (Morris, 2011).

The future of GWAS is dependent on larger sample sizes: bigger samples sizes will lead to the identification of more associated variants of smaller effect. This will in turn account for more genetic variation of a trait and improve the accuracy of genetic predictors. Larger sample sizes are coming to the forefront in recent years with the advent of mega-biobanks that have over 100,000 individuals (such as UK Biobank (Sudlow *et al.*, 2015)). However, larger sample sizes (of 100,000+ individuals) bring new problems to focus. Firstly, there are problems with replication (Huffman, 2018). In a study of such large magnitude, replication can come from meta-analysis of other smaller studies, or from splitting the main sample into a discovery and replication set. However, it is argued that when splitting a sample, the two datasets are not entirely independent due to the combined data collection and processing (Huffman, 2018).

Currently, GWAS is based on common SNP arrays (relying on LD); in the future it is likely that GWAS will be based on whole genome sequencing, however the cost of WGS

is still prohibitive at very large scales and does not warrant switching from using arrays, particularly as imputation can recover genotypes at millions of ungenotyped markers when using the most up to date imputation reference panels (see below). Focus in upcoming years will also move to post-GWAS work; improving overall biological understanding and using that knowledge to improve medicine.

4.1.3 GWAS protocol

Quality Control

Genome wide association studies use high-throughput genotyping technologies to assay hundreds of thousands single nucleotide polymorphisms (SNPs) to determine if any association exists between variants and a phenotypic trait (Pearson and Manolio, 2008). These large genotyping arrays of 500,000+ SNPs are estimated to capture a significant proportion of variation in populations (67-89% in European and Asian populations and 46-66% in African populations (Frazer *et al.*, 2007). However, genome wide association studies by their design are prone to small effects of assay and selection bias if proper quality control measures are not undertaken. QC procedures have been designed to address both SNP-specific and individual-specific problems found in raw data, which generally arise from either mis-identified or unidentified data. Some QC procedures can also highlight any larger problems with the data; for example, discordant sex information in individual samples may suggest a sample mix-up (Turner *et al.*, 2011). Left unchecked, errors in data can result in both false positives (false associations that cloud real associations) or false negatives (real associations which are not detected).

Population substructure within the dataset is also checked. It is important that the individuals sampled are from an ancestrally homogeneous population. If there are several groups of individuals who differ in genetic ancestry, this has the potential to cause false associations: for example, if the groups also differed systematically in their phenotype, any detected association could be due to differences in ancestry rather than a true association of an allele causing the phenotype (Cardon and Palmer, 2003). The substructure is assessed through the calculation of principal components. The calculation of principal components reduces the dimensionality of the data: rather than

an individual being represented by values from thousands of variables (e.g. genetic markers), it can be represented by relatively few variables (the principal components) (Ringnér, 2008). The first principal component calculated shows the largest variation between samples, and the second principal component shows the largest variation that is uncorrelated to the first (Ringnér, 2008). Typically, 10 principal components are calculated (Turner *et al.*, 2011). The principal components can be plotted against one another to visually assess the differences between samples. Individuals with similar ancestry tend to cluster together, allowing the identification of ethnic outliers for removal (Ringnér, 2008). Principal components are also used as covariates to account for the minor differences in ancestry still present after removing frank outliers.

Imputation

Although genotyping chips are rapidly improving in regard to the number of SNPs that can be genotyped, it is often not financially possible to obtain maximum genetic coverage for every individual in a cohort (Herzig *et al.*, 2018). To overcome this, it is possible to take advantage of linkage disequilibrium patterns and sequence data from many individuals to create imputation reference panels; large panels of SNPs that include markers on all arrays used and many others, and which thus can be used to predict SNPs not present on the genotyping chips. By using imputed SNPs in an analysis, a causal locus can be identified even if the causal SNP is not genotyped, via an indirect association between the imputed SNP and phenotype (Hirschhorn and Daly, 2005).

Imputation uses known reference panels that have been genotyped with a much higher number of genetic variants to infer an individual's haplotypes, based on their observed SNPs (McCarthy *et al.*, 2016). There are several panels which can be used to impute genetic data, however the Haplotype Reference Consortium (HRC) panel currently has the largest coverage for European populations (McCarthy *et al.*, 2016). The HRC panel uses 64,976 haplotypes at 39,235,157 SNPs, constructed from 20 European genetic studies and allows minor allele frequencies as low as 0.001 to be imputed accurately (McCarthy *et al.*, 2016). The reference panel contains dense genetic information for the markers surrounding the haplotype, however, as the study sample haplotype may have several matches in the reference panel, the surrounding genetic information is given a match score rather than assigned a specific allele. For example, instead of assigning a

SNP as allele C, it would be reported as 0.88 C, 0.11 A, 0.01 T. This probability information is considered in analyses and allows an estimate of uncertainty to be accounted for.

Tests of association

Following data quality control and imputation, tests for association are carried out to estimate the desired fixed effects (and in mixed models, random effects). Genotypes can be grouped under several different types of models, including additive, dominant and recessive, although the additive model is most frequently used in complex trait GWAS (Visscher *et al.*, 2017). In this model, it is assumed that each additional copy of the minor allele will increase the trait or risk of disease by the same value.

A test of association is carried out on each individual SNP. The type of test used can include linear regression, logistic regression and mixed modeling, where both fixed and random effects are included in the model.

Linear regression is used as a statistical tool used to model the relationship between the genotype and phenotype of continuous traits, under the assumption that the relationship is linear. Each SNP is tested using the following equation:

$$\boldsymbol{Y = \alpha + \beta X_i + e_i} \tag{11}$$

where $\boldsymbol{Y}$ is a n x 1 vector of phenotype data,

$\boldsymbol{i}$ is the i^{th} individual,

$\boldsymbol{\alpha}$ is the baseline phenotype (the intercept),

$\boldsymbol{\beta}$ is the fixed effect size,

$\boldsymbol{X}$ is an n x 1 vector of the nonreference allele count at the SNP and

$\boldsymbol{e}$ is the random error term of individual *i*.

The distributional properties of the random error term *e* must be independent and identically distributed with a normal distribution. However, with related populations, the independence of *e* is violated.

Binary traits use logistic regression instead of linear regression, which constrains predicted probabilities to a range of 0 to 1. The basic logistic regression model for each SNP is as follows:

$$ln\left(\frac{q_i}{(1-q_i)}\right) = \alpha_i + \beta X_i \quad \textbf{(12)}$$

where $\boldsymbol{q}$ is is a n x 1 vector of phenotype data,

$\boldsymbol{i}$ is the i^{th} individual,

α is the baseline phenotype (the intercept),

$\boldsymbol{\beta}$ is the fixed effect size and

$\boldsymbol{X}$ is an n x 1 vector of the nonreference allele count at the SNP.

These methods can be applied successfully with the condition that there is no population structure causing stratification within the data. Although QC procedures try to detect and remove population structure (if possible), some structure may still exist within the data. Population stratification may be present in seemingly homogeneous populations, for example a population that has experienced several isolated migratory events from various source populations may exhibit stratification, or populations that contain family groups (Hirschhorn *et al.*, 2002). A solution for population stratification in a population that has substructure is by using linear polygenic mixed effects models, which consider the relationship between individuals within the population sample. These models combine both fixed effects and random effects, in this situation using a relationship matrix, to control for kinship structure. A relationship matrix is estimated from the genome-wide SNP data. By using a kinship matrix to account for random effects, samples which have similar kinship-values have stronger random effects correlations. The model can be written using the equation (Meyer and Tier, 2012):

$$\boldsymbol{Y} = \boldsymbol{W\beta} + \boldsymbol{G\gamma} + \boldsymbol{g} + \boldsymbol{e} \quad \textbf{(13)}$$

where $\boldsymbol{Y}$ is the n x 1 vector of phenotype data,

$\boldsymbol{W}$ is a matrix of covariates,

$\boldsymbol{\beta}$ is a matrix of covariate effects,

$\boldsymbol{G}$ is an n x 1 vector of genotype data,

$\boldsymbol{\gamma}$ is a matrix of genotype effect size,

$\boldsymbol{g}$ is length n random vector of polygenic effects akin to heritability, where $\boldsymbol{g} \sim \boldsymbol{N(o,\ \sigma_g^2\ \Psi)}$ and σ_g^2 represents additive genetic variance and Ψ is the genetic relatedness matrix and

$\boldsymbol{e}$ is the random error term (residuals).

However, applying LMM to binary traits is more complex than when using fixed-effects models, as in LMM the normally distributed vector is unobserved and as such it is more difficult to apply a logit transformation. Additionally, cases are generally oversampled: this can lead to several issues when using LMM, for example the genetic effect and error vector stop being independent.

This problem applies to the analysis in this chapter, as GWAS are conducted on a binary trait in two population samples with family substructure. To overcome this, a linear fixed effects model is fitted to the data, using MS status as the phenotype and including covariates (discussed further in the methodology section). The residuals from this model are then used, along with a kinship matrix, to fit a linear mixed model. The resulting residuals from this model are then used as a phenotype for a GWAS, where a linear regression is performed on each SNP to produce estimates of effect.

As this regression is run for every SNP in the imputed panel, the significance level threshold of GWA studies needs to be very rigorous to ensure significant associations that appear by chance due to multiple testing are not classed as meaningful. For example in a study of 1 million SNPs the conventional level of $P < 0.05$ would lead to 50 000 SNPs being classed as associated with the trait of interest (although the SNPs are not independent so the 50K, while true, is misleading) (Pearson and Manolio, 2008). To avoid this, a Bonferroni correction is applied (Pearson and Manolio, 2008). Under the same conditions as given previously, the significance level would decrease from

$p=0.05$ to $p= 5 \times 10^{-8}$, with SNPs having to cross this threshold to be deemed significant. Although this commonly used method does remove the risk that the association is present by chance, studies with small sample size will struggle to detect significant variants due to a lack of power to allow small effect disease-associated SNPs to reach the significance threshold (Pearson and Manolio, 2008).

Why might GWAS results fail to replicate?

If findings from an initial study cannot be replicated, this is usually due to one or more factors, including chance, population stratification (although if a similar population is used this is likely to wrongly replicate), study bias, genotyping error, winner's curse or jackpot effect in the initial study (Khoury *et al.*, 2006; Chanock *et al.*, 2007).

The winner's curse results in overestimation of genetic effect sizes in initial studies. This was first described in 1983 from an auction theory context: within an auction, the winning bid on an item is likely to overestimate the true value of that item as it was the highest of all the bids (Bazerman and Samuelson, 1983). This can be applied to genetic association studies, where the initial positive association is the winning bid. The genetic effect size at this locus are determined from that "winning bid", and thus tend to be upwardly biased (Palmer and Pe'er, 2017). It is commonly found among GWA studies as these studies can easily be underpowered to detect small genetic effects. As the initial effect size is upwardly biased, this leads to an underestimation for the subsequent sample size required to replicate the result (Hirschhorn *et al.*, 2002).

The jackpot effect arises within family studies and population isolates; these groups may have higher frequencies of rare, moderate effect variants by chance, and so novel associations can be discovered (Feng *et al.*, 2015). However, the allelic architecture of these populations is often unique to that isolate, and so results are unlikely to be replicated in subsequent studies (Feng *et al.*, 2015).

Although the winner's curse and jackpot effect produce association results which often fail to replicate in subsequent studies, the results are true associations (despite statistical inflation). However, association results may also fail to replicate as they are false positives; this can arise in association studies due to population stratification or by chance.

4.1.4 Research aims

The aim of the research in this chapter is to carry out a GWAS on the ORCADES and VIKING cohorts to assess if any novel common variants exist in these populations that contribute to the excess prevalence of MS in these regions. Both Orkney and Shetland are small and relatively isolated populations. It may have been possible that rare MS risk variants, by chance, drifted to higher frequencies in these regions, and thus the allelic enrichment of these variants is causing the high rates of MS found in the islands.

4.2 Methodology

4.2.1 Creating a merged dataset

When conducting a GWAS, it is important to maximise the sample size as much as possible (in particular, case numbers) to increase the power to detect associations and improve estimates of effect size. For more than one population, this is typically done via meta-analysis. With ORCADES and VIKING, it is especially important to maximise the number of cases given the limited population size. However, performing a meta-analysis on the data is more challenging given the nature of the two populations: there is some migration between the populations of Orkney and Shetland, and as such many Orcadian individuals have relatives on Shetland and vice versa. While this does not cause issues for independent analyses on both datasets, it can cause a problem if there are related individuals between datasets when meta analysing results. Therefore, to account for relatedness between populations, a merged dataset was created. This merged dataset contained several additional ORCADES individuals who were not genotyped within the first group.

Merging Genotypes and Imputed Files

In the genotyped data, the intersecting SNPs in the ORCADES and VIKING PLINK array files were determined (the ORCADES PLINK files contained the NIMS individuals). A new set of PLINK files containing all individuals from ORCADES and VIKING with the intersectional SNP subset was then created using PLINK (v1.90b3.29). The ORCADES and VIKING array datafiles used to create these were those that underwent the quality control procedures discussed in **Chapter 2**.

HRC-imputed files were merged on a per chromosome basis using the program QC Tool (v2).

Merging Phenotypes

The ORCADES and VIKING phenotype files were merged in R (v3.4.4). Phenotypes were quality controlled to ensure units were consistent between phenotypes.

Calculating principal components and GRM

Principal components were created using the genotype files and PLINK. PLINK generates a variance-standardised genetic relationship matrix and extracts the top PCs from that matrix.

4.2.2 GWAS

A GWAS was carried out on the merged HRC-imputed ORCADES/VIKING dataset. The phenotype used for the GWAS were GRAMMAR+ residuals; several stages were performed to take the raw MS phenotype to the GRAMMAR+ residuals to allow them to be analysed as a quantitative trait.

Stage 1: Fitting a linear fixed effects model

Here, the phenotype MS status and covariates (not including kinship) were fit in a linear fixed effect model. Age and sex were included as covariates due to their association with MS: MS has an age of onset of around 30 years of age and is twice as common in women than men. Principal components 1 through 10 were included to adjust for population structure. The following model was fit using the program GenABEL v1.8-0 (Aulchenko *et al.*, 2007):

Model 1: MS status (1,0) ~ Age + Sex + PC1..PC10

Fitting this model produced covariate-adjusted residuals, which were then passed on to the next stage of analysis. No genotype data is used at this stage.

Stage 2: Fitting a random effects model

Residuals produced from *Stage 1* along with a GRM (see **4.2.1**) were fit in a random effects model:

Model 2: Residuals (*Model 1*) ~ GRM

As previously discussed in **Chapter 2**, both cohorts have a high degree of relatedness, and so accounting for kinship as a random effect is necessary to prevent population stratification within this data. GenABEL's polygenic function was used to fit the mixed model. This model generated GRAMMAR+ residuals.

Stage 3: Genome-wide association analysis
The GRAMMAR+ residuals produced in *Stage 2* were passed on to REGSCAN (v0.5) for GWA analysis.

Model 3: Residuals (*Model 2*) ~ SNP

REGSCAN performs a linear regression for each SNP using the GRAMMAR+ residuals as a phenotype to produce estimates of effect sizes and their standard errors.

4.3 Results

4.3.1 Merged dataset (ORCADES/VIKING) summary

Both the genotyped and imputed data files for ORCADES and VIKING were merged successfully. A subset of 156,040 SNPS, found in both ORCADES and VIKING, were kept for the merged genotype files (Figure 9). It should be noted that the number of ORCADES SNPs is much lower than that of VIKINGs due to the merger of two SNP arrays when creating ORCADES (resulting in a smaller subset of SNPs), as described in **Chapter 2**. The genotype data was used to generate principal components for the data: the first two PC are plotted in Figure 10. A clear divide can be seen between the two island population groups, indicating that population structure exists in the dataset. At least one pure Shetlander exists within ORCADES. There also appears to be a small cluster of individuals between both populations which may be admixed between Orkney and Shetland.

Table 7 provides a summary of the new dataset, which has 112 cases and 4223 controls, with the majority of cases (81) being female. The mean age in this dataset is 52 years (standard deviation of 15.4 years). Both men and women have a similar age distribution amongst cases and controls (Figure 11).

Relatedness is an important feature of both ORCADES and VIKING, each of which include many extended families. The possible relatedness of individuals across both datasets was a point of concern for analysis, and the key motivation for creating a merged dataset (to allow relationships between populations to be assessed and accounted for). 9,620,692 relationship pairings were assessed – of these, 9,610,161 pairs had a relationship coefficient below 0.10 and were deemed unrelated. Pairs with a relationship coefficient above 0.1 can be seen in Figure 12. There are 2588 pairs who have a relatedness coefficient between 0.20-0.30 (approximately the equivalent of a grandparent / grandchild or uncle-niece relationship) and 2736 pairs who have relationship coefficient between 0.38-0.62 (equivalent of parent/child or siblings).

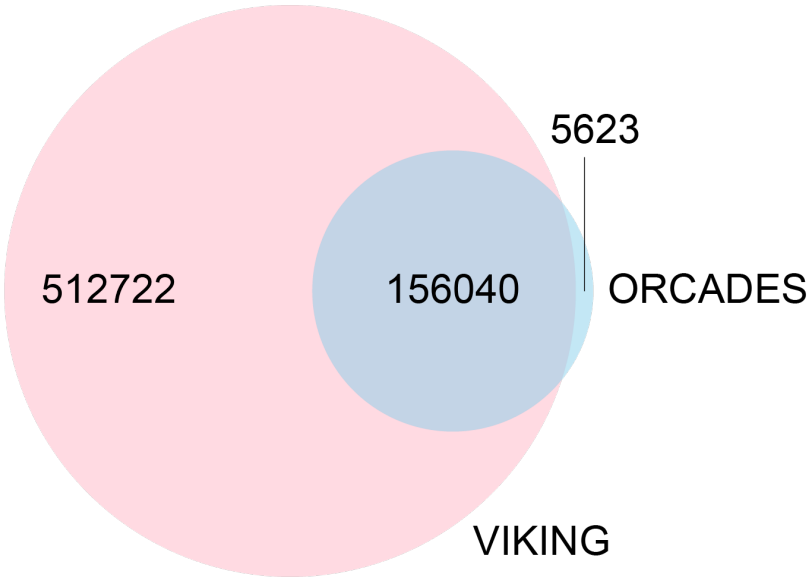


Figure 9: A summary of genotyped SNPs in ORCADES (blue) and VIKING (pink) ORCADES had a total of 161,663 SNPs while VIKING had a total of 668,762 SNPs. Of these SNPs, 156,040 were the same between both genotyped groups, and were subsequently used as the genotyped SNPs in the merged ORCADES/VIKING dataset. It should be noted that the number of ORCADES SNPs are much lower than that of VIKINGs due to the merger of two SNP arrays when creating ORCADES (resulting in a smaller subset of SNPs), as described in **Chapter 2**.

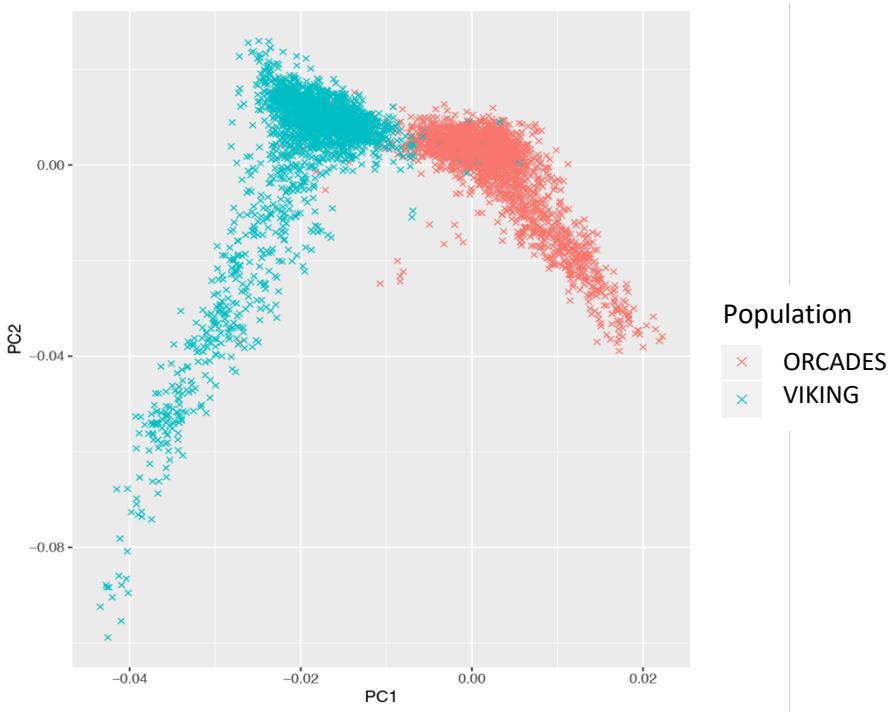


Figure 10: Principal component plots for the merged ORCADES/VIKING dataset
Principal component (PC) plots for the merged ORCADES/VIKING dataset (n=4335), using PC1 and PC2. Individuals from ORCADES are plotted in red and individuals from VIKING are plotted in turquoise.

		Count			Mean Age (standard deviation)		
Population	Sex	Case	Control	Total	Case	Control	Total
ORCVIK	Male	32	1697	1729	54.94 *(9.42)*	53.11 (15.42)	53.14 (15.33)
	Female	81	2544	2625	49.77 (12.48)	51.41 (15.50)	51.36 (15.42)
	All	112	4223	4335	51.23 (11.89)	52.09 (15.49)	52.07 (15.41)

Table 7: Summary statistics for merged ORCADES/VIKING dataset
Count and mean age for the merged ORCADES/VIKING dataset, split by gender and Multiple Sclerosis (MS) status.

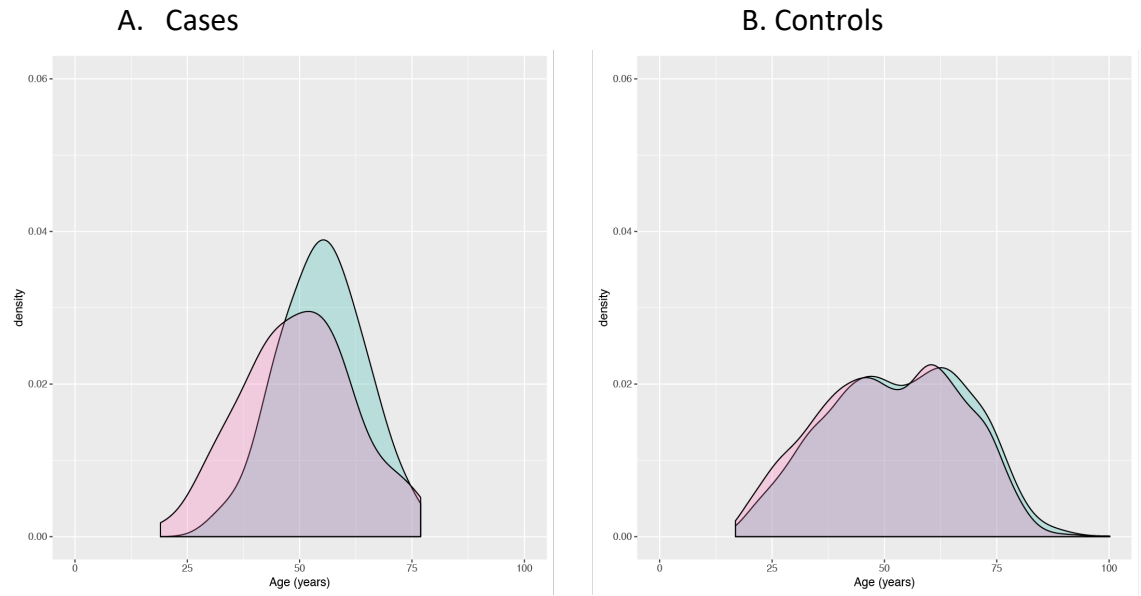


Figure 11: Age distribution plots in merged ORCADES/VIKING dataset

Age distribution plots in the merged ORCADES/VIKING dataset, split by MS status; total number of individuals within each group can be found in Table 5.

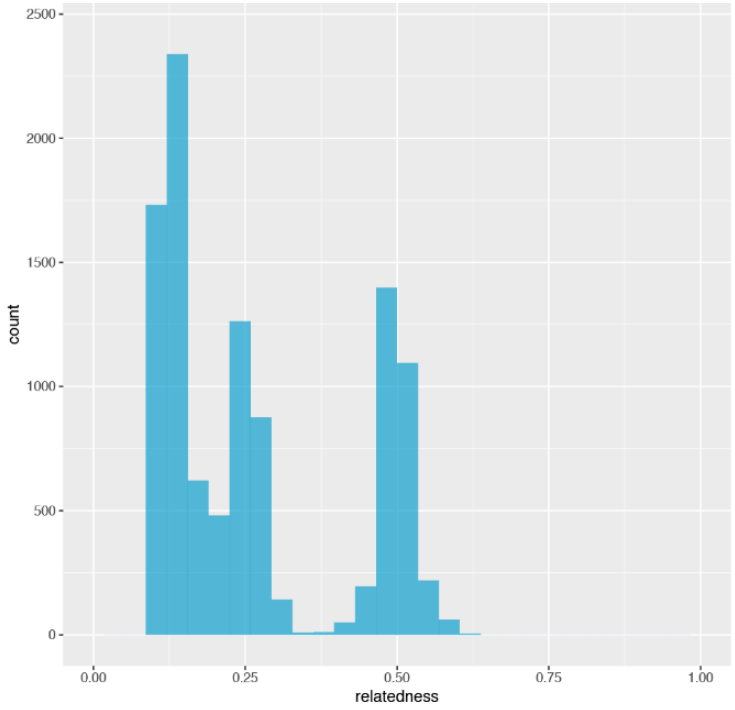


Figure 12: Relatedness in the merged ORCADES/VIKING dataset

Relatedness coefficients for 10,531 pairs in merged ORCADES/VIKING dataset. For clarity, this plot has been restricted to having a relatedness coefficient >0.10, as the majority of the 9,610,161 pairs below this point will be unrelated (Turner *et al.*, 2011).

4.3.2 ORCADES/VIKING GWAS results

A GWAS was conducted on the merged ORCADES/VIKING dataset to determine if there were any novel common MS risk variants that existed in the Northern Isles of Scotland. The ORCADES/VIKING dataset contained 112 MS cases and 4223 control individuals from the Northern Isles of Orkney and Shetland in the United Kingdom. In this results section, only SNPs with a MAF above 0.05 were considered, as SNPs with a MAF below 0.05 are more likely to be genotyping errors (Turner *et al.*, 2011), as well as being poorly powered, given our limited sample size.

There is little evidence of genomic inflation within the ORCADES/VIKING dataset (Figure 13). The quantile-quantile plot of the observed versus expected p-values found that the majority of the p-values followed the expected distribution. The lack of inflation ($\lambda = 1$) indicates that the analysis methods used have successfully corrected for any population structure within the dataset.

In this GWAS, one SNP (SNPID chr1_145044288) crossed the genome-wide significance threshold of 5×10^{-8}, however this looks unlikely to be a true signal as it is the only SNP within 1000 kb to have a p-value below 1×10^{-4}. In total, 89 SNPs had a p-value below the suggestive significance threshold of 1×10^{-5} (the threshold as defined in Björkegren *et al.*, 2015). A shortened list of these SNPs, with only the lead SNPs within 1000 kb, can be found in
Table 8 (lead SNP defined as the SNP with the lowest p-value), and a full list of these SNPs can be found in Supplementary Table 1. Although beta values and SE are listed, these should be read with caution as these SNPs (apart from chr1_145044288) were not genome-wide significant. Therefore, the results are not very informative with regards to effect size but are suggestive of possible MS susceptibility loci in the Northern Isles. Also included in this table are the p-values and associated allele frequencies from the most recent 2018 IMSGC study for comparison of results (International Multiple Sclerosis Genetics Consortium *et al.*, 2018). This study used 32,367 MS cases and 36,012 controls from Europe, Australia and the USA.

There were six regions that had two or more SNPs that had p-values below the suggestive significance threshold that were within 1000 kb of each other. These regions

can be viewed in the Manhattan plot (Figure 13) and are found on chromosomes 2, 6, 12 and three regions on chromosome 18.

Chromosome 2 had 14 SNPs within a 1000 kb region that passed the suggestive significance threshold. Lead SNP rs1398972 had a p-value of 1.49×10^{-7} in the GWAS results (Figure 14). 13 SNPs nearby were in strong LD with this SNP ($r^2 > 0.9$), along with multiple weakly correlated SNPS ($0.2 < r^2 < 0.6$) that did not pass the suggestive significance threshold. The cluster of these SNPs appears between two recombination peaks of around 20cM/Mb. rs1398972 was not significant in the 2018 IMSGC study (p-value=0.06), however the frequency of the allele in the ORCADES/VIKING dataset (0.32) is greater than that of the IMSGC dataset (0.13). This SNP is not reported to have any clinical significance and has no listed associations on the PhenoScanner website, which is a curated list of large-scale GWAS results (Staley *et al.*, 2016). The closest gene to this SNP is *LINC01117*, Long Intergenic Non-Protein Coding RNA 1117 (Bethesda (MD): National Center for Biotechnology Information, 2005).

Chromosome 6 had 27 SNPs within a 1000 kb region that passed the suggestive significance threshold, however 25 of these did not have an assigned RSID. The SNP with the lowest p-value and an assigned RSID number was rs9268154, which had a p-value of 9.7×10^{-6} (Figure 15). The SNP with the lowest p-value without an assigned RSID number was SNPID chr6_32411726 which had a p-value of 3.9×10^{-6}. As LDplot (the tool used to assess linkage disequilibrium between SNPs and plot them on the locus zoom plots) uses RSID to index SNPs, the majority of these SNPs could not be assessed for LD with the lead SNP or plotted on the locus zoom plot. However, the 2 SNPs with RSIDs that passed the suggestive significance threshold were in LD with each other ($r^2=1$). These two SNPs were moderately correlated with *HLA-DRB1*1501* tag SNP rs3135388 ($r^2=0.65$). There are also several correlations between $0.4 < r^2 < 0.6$ in the *HLA-DRB* region, which lies approximately 200 kb from rs9268154. This group of 27 SNPs was the only region to show significance in the IMSGC paper; rs9268154 had a reported p-value of 0 in the IMSGC results. The associated frequency was also higher in the IMSGC results at 0.85 (compared to the 0.78 frequency in the ORCADES/VIKING dataset). The most significant association for the SNP rs9268154 on PhenoScanner was self-reported Multiple Sclerosis, which listed a p-value of 8.9×10^{-89} from a UK BioBank study (n=337159).

Chromosome 12 had 12 SNPs within a 1000 kb region that passed the suggestive significance threshold. Lead SNP rs11055646 had a p-value of 5.6×10^{-7} in the GWAS results (Figure 16). 4 SNPs nearby were in strong LD with this SNP ($r^2 > 0.9$), and 7 SNPs had an r^2 between 0.6 and 0.7. rs11055646 was not significant in the IMSGC study (p-value = 0.14), and the SNP was at a higher frequency (0.22) in the IMSGC study than in ORCADES/VIKING (0.13). There are no listed associations for the rs11055646 SNP on PhenoScanner. The region that contained these SNPs is part of the *GRIN2B* gene, and this gene has no reported clinical significance (Bethesda (MD): National Center for Biotechnology Information, 2005).

Three regions on chromosome 18 had multiple SNPs that passed the suggestive significance threshold, at around 20 Mb, 36 Mb and 56 Mb along chromosome 18. The first region, at around 20 Mb, had 7 SNPs within a 1000 kb region that passed the suggestive significance threshold. Lead SNP rs1893251 had a p-value of 2.9×10^{-7} in the GWAS results (Figure 17), and the other 6 SNPs which passed the suggestive threshold were in strong LD with this SNP ($r^2 > 0.8$). Within the IMSGC study, rs1893251 was not significant (p-value = 0.68) and had a higher frequency (0.12) than ORCADES/VIKING (0.05). This SNP has no reported associations on PhenoScanner. The gene in nearest proximity to this SNP is *CTAGE1* (within 2000 kb; Cutaneous T Cell Lymphoma-Associated Antigen 1).

The second region on chromosome 18 at around 36 Mb had 4 SNPs within a 1000 kb region that passed the suggestive significance threshold. Lead SNP rs17602961 had a p-value of 5.3×10^{-6} in the GWAS results (Figure 18) and the other 3 SNPs which passed the suggestive threshold were in very strong LD with this SNP ($r^2 > 0.96$). There are a number of other SNPs stretching across approximately 0.2 Mb which are in strong LD with rs17602961 that did not pass the suggestive threshold; three of these are found on nearest gene, *MIR924HG* (also known as Long intergenic non-protein coding RNA 669). This SNP was not significant in the IMSGC study (p-value=0.61) but had a similar frequency (0.02) to that of ORCADES/VIKING (0.05). It has no listed associations on PhenoScanner.

The third region, at around 56 Mb, had 5 SNPs within a 1000 kb region that passed the suggestive significance threshold. Lead SNP rs62096323 had a p-value of 4.2×10^{-6} in the GWAS results (Figure 19), and the other 4 SNPs which passed the suggestive

threshold were in strong LD with this SNP ($r^2 > 0.8$). rs62096323 was not significant in the IMSGC study (p=value 0.48) and had a lower frequency (0.04) than in ORCADES/VIKING (0.10). It has no known associations on PhenoScanner; however, it is within 0.1 Mb of a known MS SNP rs7238078, found in *MALT1* (which encodes mucosa-associated lymphoid tissue lymphoma translocation protein 1). rs7238078 has a p-value of 3×10^{-9}, from the 2011 IMSGC study with 9772 MS cases and 26621 controls (Sawcer *et al.*, 2011). The MS associated SNP rs7238078 is not in LD with rs62096323 (r^2=0.002).

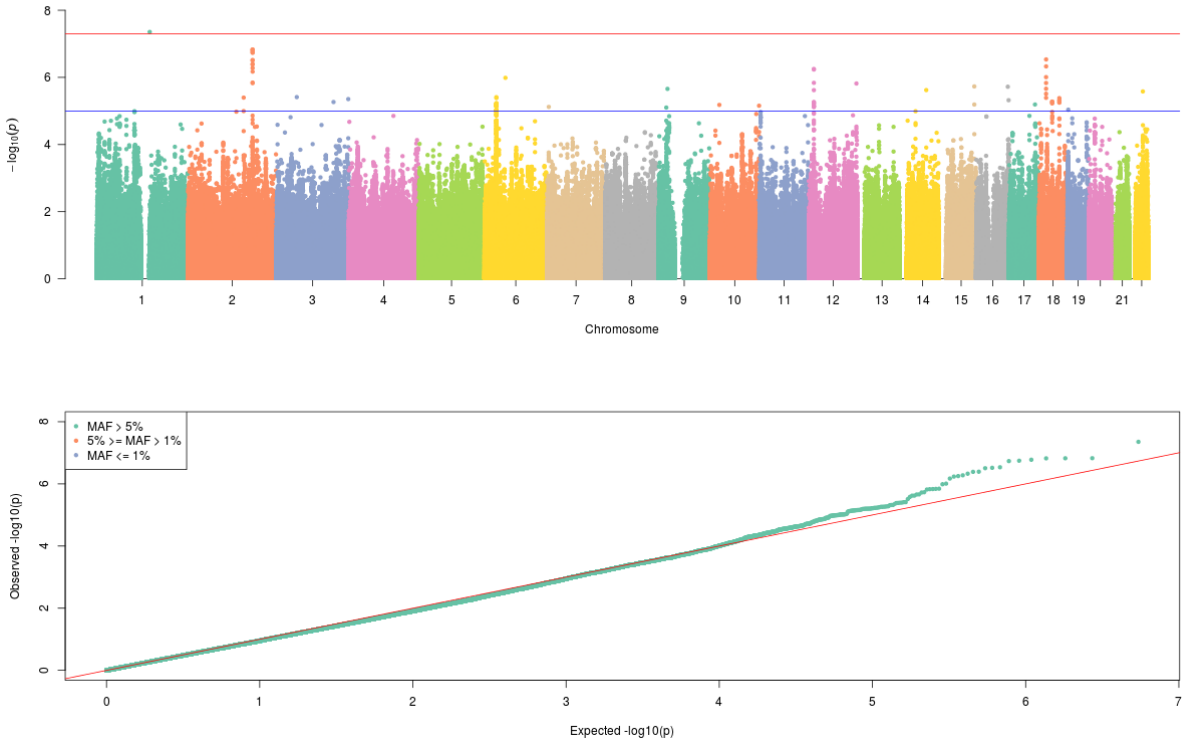


Figure 13: Manhattan and QQ plot from GWAS of Multiple Sclerosis in Shetland and Orkney

Results from a GWAS of MS using the ORCADES/VIKING dataset (cases: 112, controls: 4223). The top graph is a Manhattan plot, with chromosomes along the x-axis and $-\log_{10}$ p-values along the y-axis. The genome-wide significant threshold of 5×10^{-8} is shown in red, with the suggestive threshold of 1×10^{-5} is shown in blue. The bottom graph is a quantile-quantile plot of observed and expected p-values, $-\log_{10}$ transformed. The genetic inflation factor (λ) was 1. Only SNPs with a MAF above 0.05 are included in both plots.

RSID/SNPID	Chr	Position	A1	A0	Freq	Beta	SE	p-value	Info	IMSGC p-value	IMSGC Freq
chr1: 145044288	1	145044288	C	T	0.09	0.15	0.03	4.4×10^{-8}	0.50	NA	NA
rs34584371	2	153037808	A	G	0.12	0.08	0.02	4×10^{-6}	0.99	0.76	0.1
rs1398972	2	177569251	C	G	0.32	0.06	0.01	1.5×10^{-7}	0.99	0.06	0.13
rs7635898	3	55408967	C	A	0.09	0.09	0.02	3.9×10^{-6}	0.84	NA	NA
rs816545	3	156092170	T	C	0.05	0.11	0.02	5.4×10^{-6}	0.93	0.06	0.03
rs73221623	3	196370840	T	C	0.10	0.08	0.02	4.4×10^{-6}	0.97	0.76	0.04
rs9268154	6	32266021	A	T	0.78	-0.06	0.01	9.7×10^{-6}	1.00	0.00	0.85
chr6: 32411726	6	32411726	A	G	0.75	-0.06	0.01	3.9×10^{-6}	1.00	0.00	0.87
rs12209200	6	57217336	T	C	0.08	0.10	0.02	1×10^{-6}	0.84	NA	NA
rs4720446	7	4982335	C	G	0.88	-0.07	0.02	7.6×10^{-6}	0.98	0.60	0.96
rs10965046	9	21518275	G	A	0.14	0.07	0.02	7.9×10^{-6}	0.98	0.50	0.16
rs76585251	9	24750193	T	G	0.07	0.10	0.02	2.2×10^{-6}	0.95	0.76	0.02
rs78870428	10	25662804	G	A	0.10	0.09	0.02	6.6×10^{-6}	0.86	0.03	0.03
rs117374511	10	134496477	T	C	0.06	0.10	0.02	7×10^{-6}	0.91	0.63	0.02
rs11055646	12	14005719	C	T	0.13	0.08	0.02	5.6×10^{-7}	1.00	0.14	0.22
rs73162646	12	130464398	C	T	0.13	0.08	0.02	1.5×10^{-6}	0.99	0.66	0.13
rs8008067	14	72941207	A	G	0.07	0.10	0.02	2.4×10^{-6}	0.88	NA	NA
rs8041424	15	97359414	T	C	0.06	0.11	0.02	1.9×10^{-6}	0.95	0.89	0.02
rs7199663	16	86608899	C	A	0.16	0.07	0.01	1.9×10^{-6}	0.95	NA	NA
rs62048038	16	88727961	A	T	0.08	0.09	0.02	4.8×10^{-6}	0.92	0.71	0.03
rs2567473	17	70738077	A	G	0.75	-0.06	0.01	6.4×10^{-6}	0.99	0.17	0.56
rs1893251	18	19991432	T	C	0.05	0.12	0.02	2.9×10^{-7}	0.98	0.68	0.12
rs17602961	18	36627629	A	C	0.05	0.11	0.02	5.3×10^{-6}	0.99	0.61	0.02
rs62096323	18	56304224	G	C	0.10	0.08	0.02	4.2×10^{-6}	0.97	0.48	0.04
rs75479243	19	2739091	T	C	0.08	0.10	0.02	9.2×10^{-6}	0.85	0.37	0.02

Table 8: Key results from MS GWAS in ORCADES/VIKING dataset

Key results from a GWAS of MS on the ORCADES/VIKING dataset (cases: 112, controls: 4223). SNPs with a MAF below 0.05 were not considered. The SNPs listed here have passed the suggestive significance threshold of 1×10^{-5} and have the lowest p-value within a 1000 kb region. Two SNPs are included for the chromosome 6, 32-33 Mb region to allow the top hit *without* an assigned RSID and a top hit *with* an assigned RSID to be shown. Within this results table, it should be noted that only one of these SNPs passed the genome-wide significance threshold (SNPID chr1_145044288), and so beta values should be read with caution. The RSID is the Reference SNP cluster ID; if no RSID number has been assigned, the SNPID is listed. Chr is the

chromosome the SNP is located on. Position is the base pair position the SNP can be found at GRCh37. A1 is the reference allele, and A0 is the non-reference allele. Freq is the frequency of the A1 allele within the dataset. Beta is the effect size estimated in the analysis. SE is the standard error of the effect size. The p-value is the calculated probability or level of significance of the effect. The IMSGC p-value is the p-value from the SNP from the 2018 IMSGC summary statistics (International Multiple Sclerosis Genetics Consortium *et al.*, 2018). The IMSGC is the associated (Europe-wide) frequency for the IMSGC summary statistics (International Multiple Sclerosis Genetics Consortium *et al.*, 2018). Info is a measure of quality for imputation.

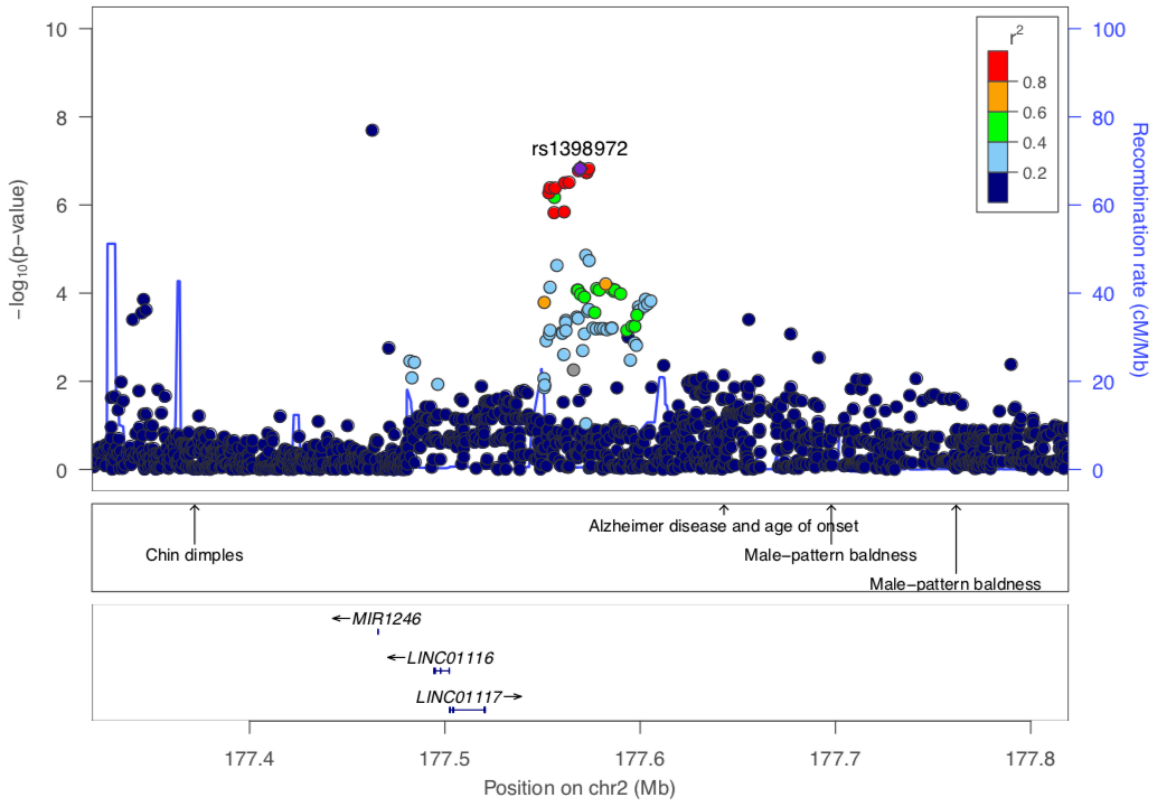


Figure 14: Locus zoom plot for chromosome 2 SNP rs1398972

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs1398972 (p-value = 1.5 x 10^{-7}). This SNP passed the suggestive significance threshold of 1 x 10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 2 in Mb. The y-axis on the left-hand side shows the -$\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs1398972 (indicated by a purple

diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region.

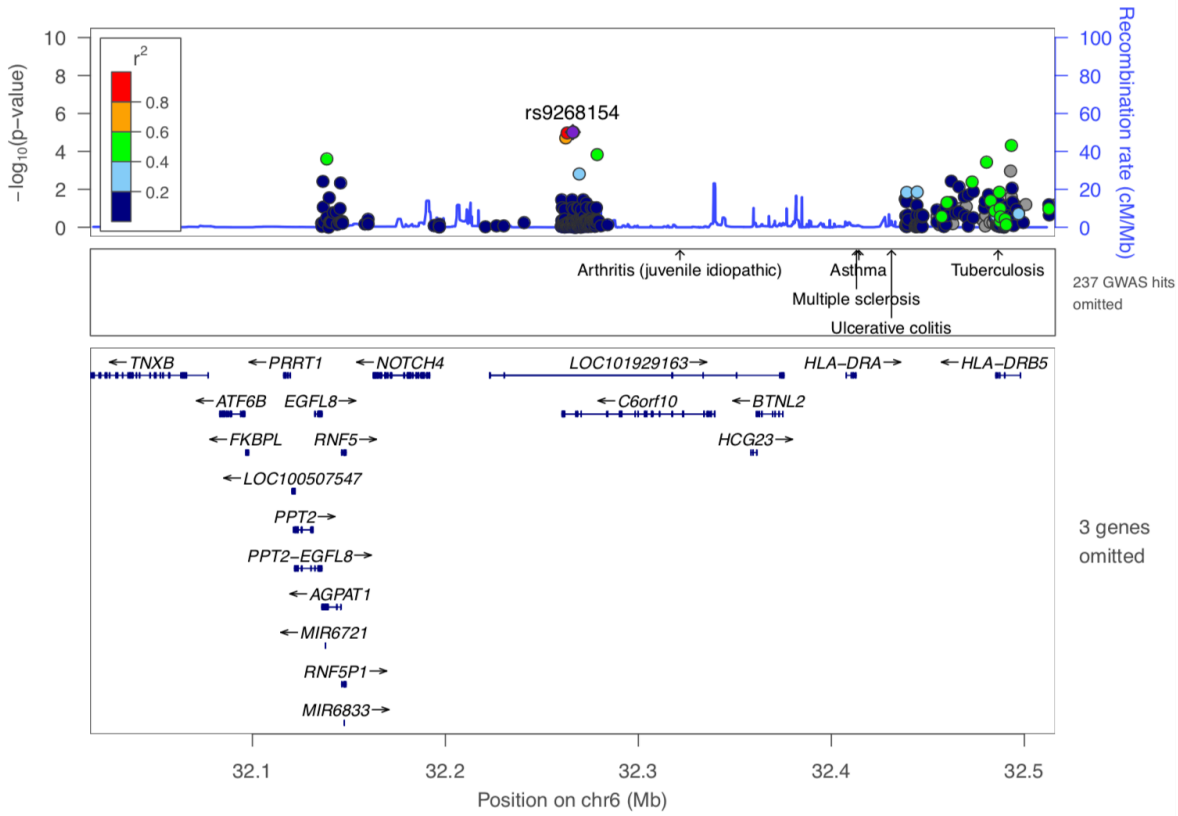


Figure 15: Locus zoom plot for chromosome 6 SNP rs9268154

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs9268154 (p-value = 9.7×10^{-6}). This SNP passed the suggestive significance threshold of 1×10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 6 in Mb. The y-axis on the left-hand side shows the $-\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs9268154 (indicated by a purple diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region. A large proportion of this graph is not plotting SNPs; however, this is not due to genetics. Rather, a large number of SNPs in this region do not have assigned RSIDs. When LDPlot is used to index SNPs for plotting, it uses RSIDs; if a SNP does not have an assigned RSID, it is not plotted. Therefore, although there are SNPs present, they are not shown on the plot.

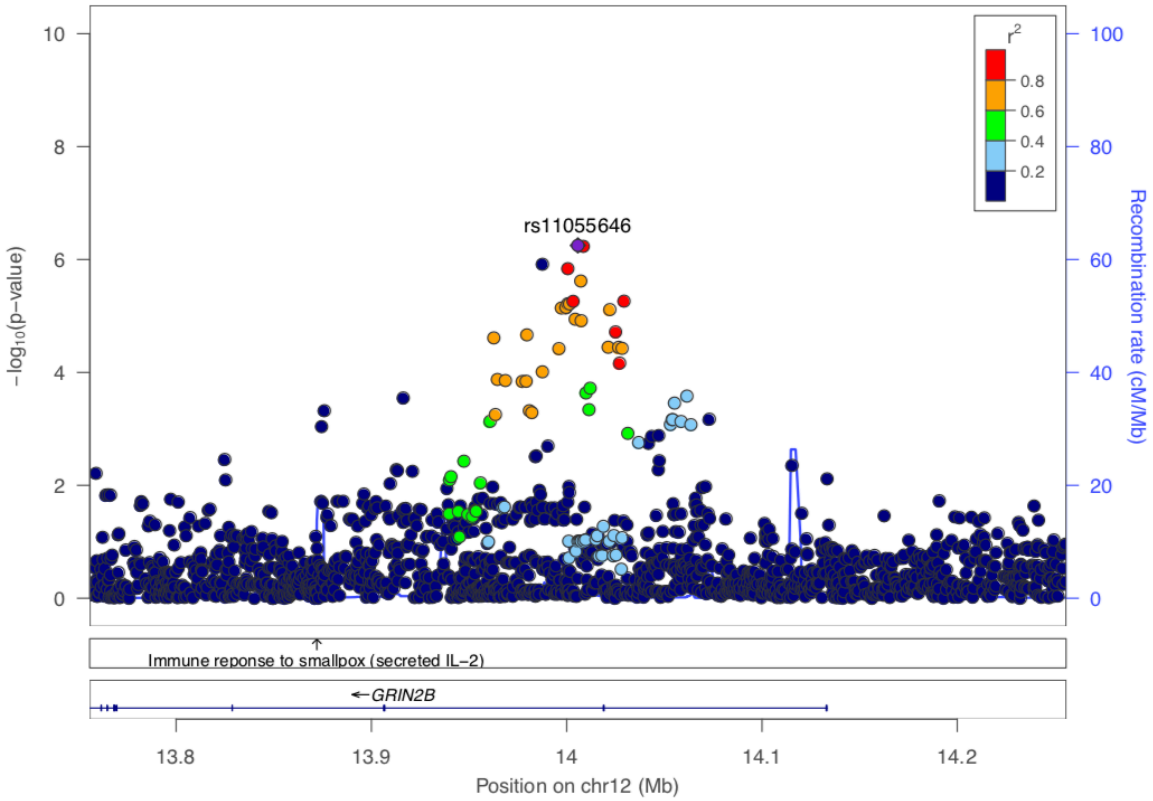


Figure 16: Locus zoom plot for chromosome 12 SNP rs11055646

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs11055646 (p-value = 5.2×10^{-7}). This SNP passed the suggestive significance threshold of 1×10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 12 in Mb. The y-axis on the left-hand side shows the $-\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs11055646 (indicated by a purple diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region.

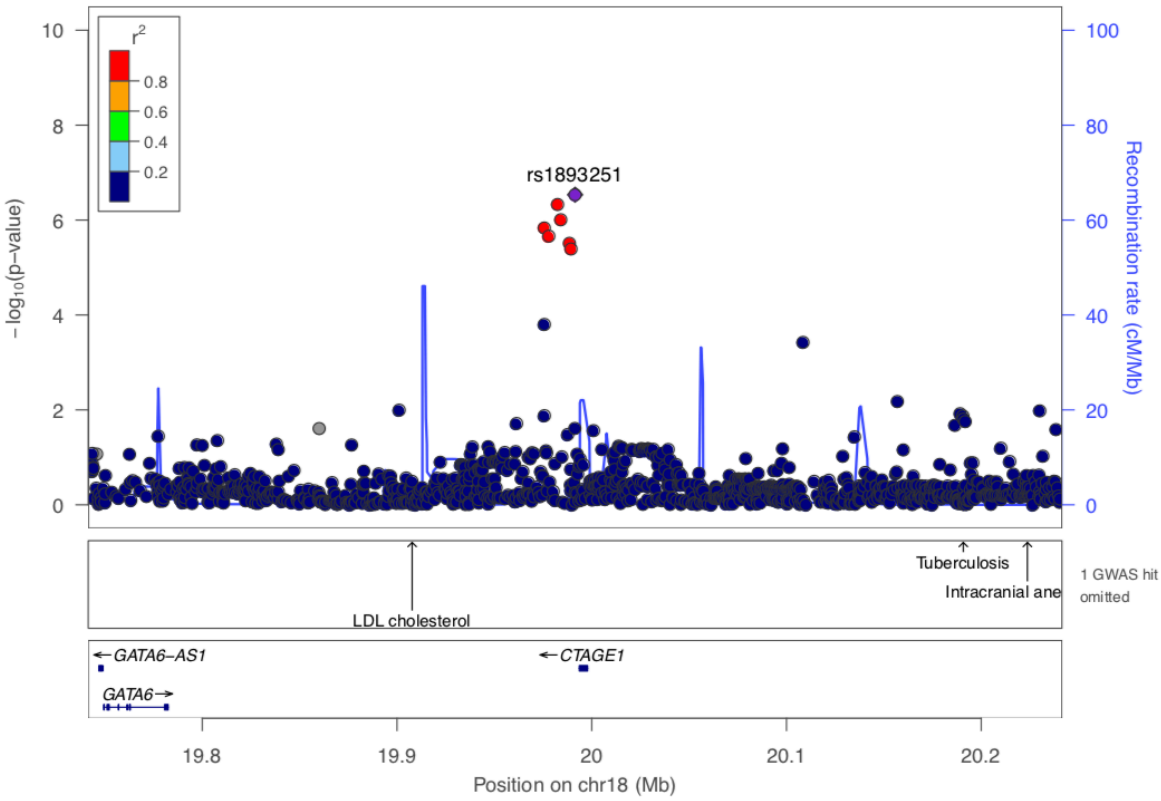


Figure 17: Locus zoom plot for chromosome 18 SNP rs1893251

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs1893251 (p-value = 2.9 x 10^{-7}). This SNP passed the suggestive significance threshold of 1 x 10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 18 in Mb. The y-axis on the left-hand side shows the -$\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs1893251 (indicated by a purple diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region.

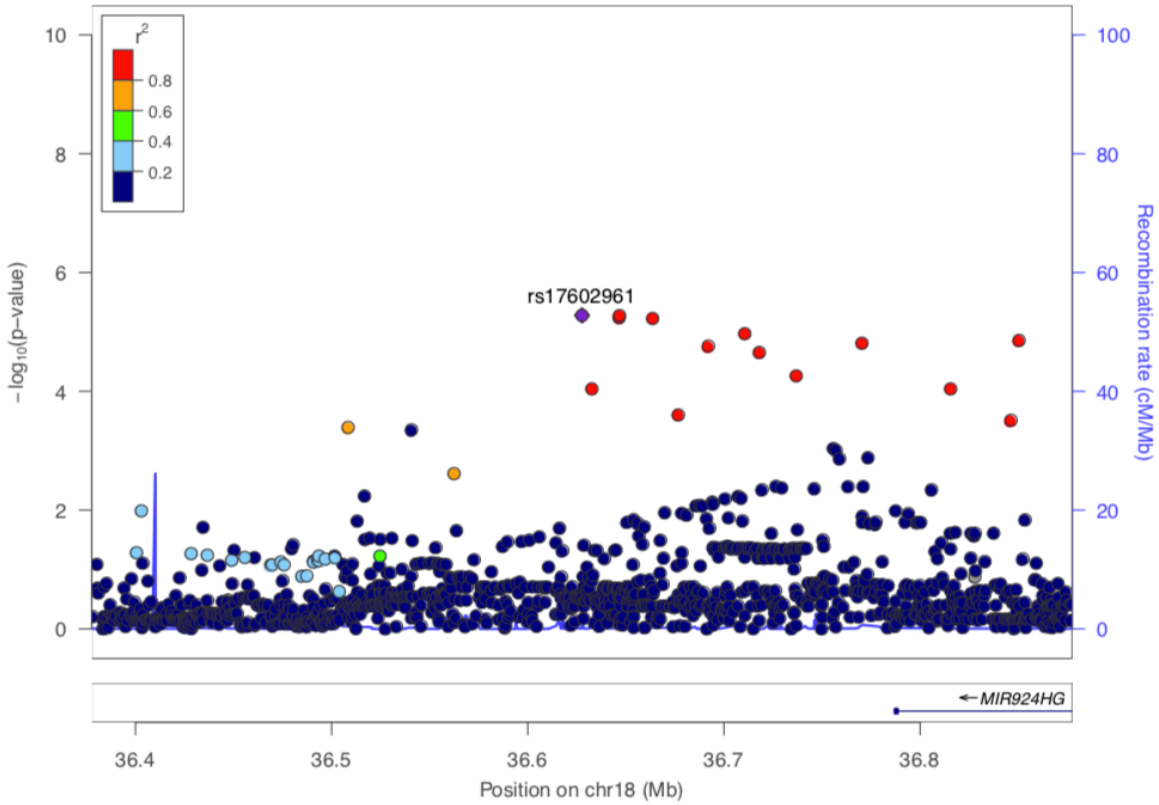


Figure 18: Locus zoom plot for chromosome 18 SNP rs17602961

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs17602961 (p-value = 5.3×10^{-6}). This SNP passed the suggestive significance threshold of 1×10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 18 in Mb. The y-axis on the left-hand side shows the $-\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs17602961 (indicated by a purple diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region.

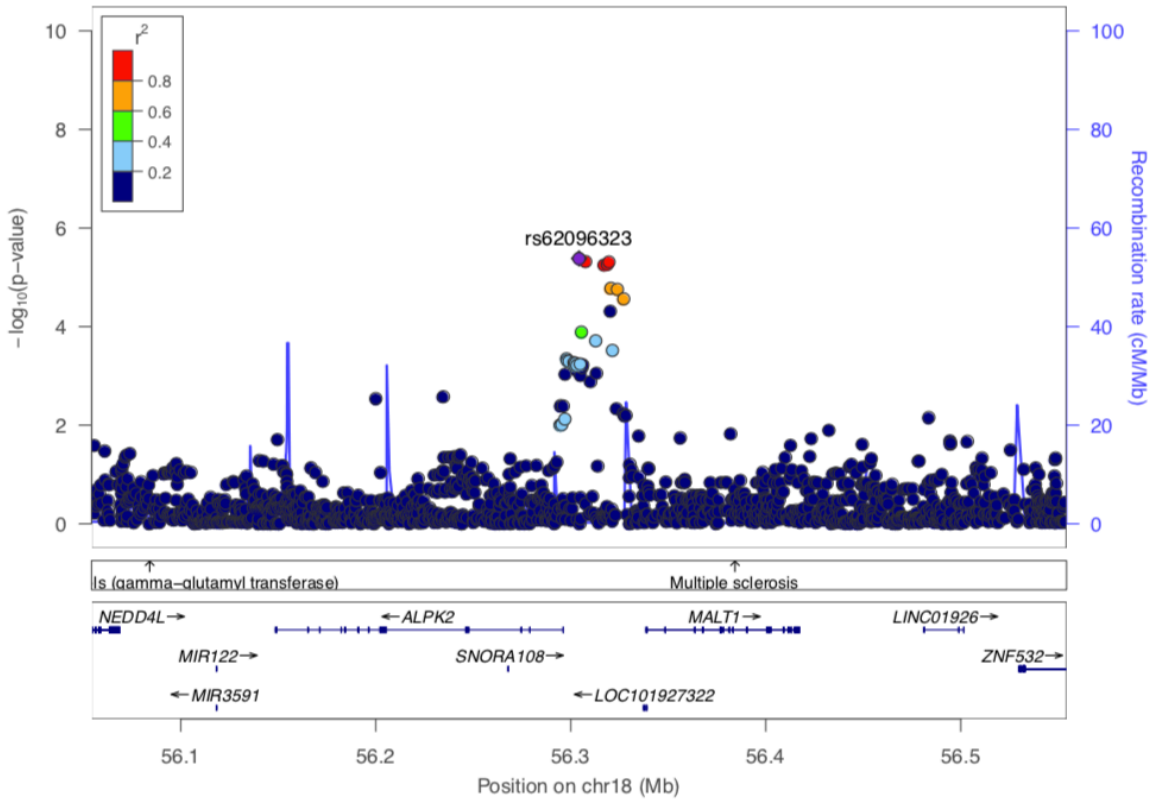


Figure 19: Locus zoom plot for chromosome 18 SNP rs62096323

This locus zoom plot is showing the results from a GWAS on MS in ORCADES/VIKING (cases: 112, controls: 4223) for the SNP rs62096323 (p-value = 4.2×10^{-6}). This SNP passed the suggestive significance threshold of 1×10^{-5} and is the SNP with the lowest p-value within 1000 kb in this locus. On the x-axis is the position on chromosome 18 in Mb. The y-axis on the left-hand side shows the $-\log^{10}$ p-value, while the axis on the right in blue shows the recombination rate in cM/Mb. SNPs are coloured based on the degree of LD (r^2) to rs62096323 (indicated by a purple diamond). Previous significantly associated traits from the GWAS catalogue are shown underneath the locus zoom plot, along with gene locations within the plotted region.

4.4 Discussion

Summary of Findings

In this chapter, a genome-wide association study was performed on a dataset containing 112 Multiple Sclerosis cases and 4223 controls from Orkney and Shetland. A common

MS risk variant, unique in frequency to the Northern Isles of Scotland, may exist; this study was undertaken to detect if such a variant existed.

In this study, six separate regions (separate defined as 1000 kb apart) were detected with suggestive significance. One of these regions was previously associated with MS, therefore we were powered to detect a real MS variant at suggestive significance. However, there was no evidence from larger population studies that other suggestive variants detected were associated with MS. These variants are therefore false positives or are in LD with an unusual variant in Orkney.

A chromosome 6 region in this study (lead SNP rs9268154, p-value of 9.7×10^{-6}) was previously associated with MS in the IMSGC study (International Multiple Sclerosis Genetics Consortium *et al.*, 2018). This SNP was in LD ($0.4 < r^2 < 0.65$) multiple SNPs in the HLA region, including *HLA-DRB1*1501* SNP rs3135388 ($r^2 = 0.65$). Significant variants within the HLA region are well known to associate with increased MS susceptibility, for example MS cases carry *HLA-DRB1*1501* twice as frequently as healthy controls (Sawcer *et al.*, 2011).

The detection of a previously identified MS risk variant shows that the methodology in this study works to detect a variant at the suggestive significance level. However, as this is one of the strongest known MS associations and it is one of the most significant variants in this study, it suggests that there isn't a Northern-Isles-specific common variant that is dominating the phenomenon of excess MS cases in the region. If a common variant of stronger effect size were present in the Northern Isles population, it would have been detected with a better significance in this study.

However, the other variants that are detected at suggestive significance may be MS-associated variants unique to the Northern Isles. Although they may be due to chance, it is possible that with greater power these regions could be genome-wide significant. It is important to note that the SNP results from this study are discussed tentatively, with the knowledge that none of the association results considered below achieved genome-wide significance. Discussion is speculative based on suggestive ($5 \times 10^{-8} < p < 1 \times 10^{-5}$) findings, and these variants did not reach significance in the larger IMSGC study (International Multiple Sclerosis Genetics Consortium *et al.*, 2018).

Discussion of suggestive variants

One of the six suggestive hits, chromosome 2 SNP rs1398972 (p-value 1.49 x 10^{-7}), was found to have a frequency of 0.32 in the Northern Isles dataset, which is higher than that of the IMSGC study (frequency 0.13). There would thus be about three times more power in the Northern Isles for a given sample size, however the IMSGC study is much more than three times the size of the present study and so is the more powerful study even for this variant. The SNP itself did not have any reported clinical significance, however it was within 100 kb of a locus associated with Alzheimer disease, another neurodegenerative disease. The nearest gene to rs1398972 was *LINC01117*, Long Intergenic Non-Protein Coding RNA 1117. Long non-coding RNAs (lncRNA) have been extensively studied in recent years, and have associations with many complex diseases, including Multiple Sclerosis (Cipolla *et al.*, 2018). For example, lncRNA *GAS5* was found to be differentially expressed in individuals with MS, indicating it has a role in the function and regulation of the immune system (Mayama, Marr and Kino, 2016). LncRNAs themselves do not translate into proteins, and are involved in the regulation of transcription processes, post-transcription processes (such as splicing and translation) and epigenetic processes. It is possible that this locus may have a causative role in Multiple Sclerosis, or that *LINC01117* may be have a role which influence immune system function.

Another suggestive hit, rs11055646 on chromosome 12, had a p-value of 5.6 x 10^{-7}. There were several SNPs around this 14Mb region on chromosome 12 with a strong correlation to rs11055646 ($r^2 > 0.6$), all within the *GRIN2B* gene. This gene codes for GluN2B, a protein found in brain neurons primarily in development before birth. This protein makes up part of the NMDA (N-methyl-D-aspartate) receptors on the neuron, which works as a channel to allow the flow of cations: the flow of cations excites the neurons to allow them to pass on cell signals and helps in the process of neuron maturation (Zarate *et al.*, 2006). This gene has previously been associated with immune response to smallpox (Kennedy *et al.*, 2012) and several neurodevelopmental disorders including attention deficit hyperactivity disorder and schizophrenia (Kim *et al.*, 2018). De-novo mutations in this gene are known to cause *GRIN2B*-related neurodevelopmental disorders, characterised by developmental delay and intellectual disability in known cases (Platzer and Lemke, 1993). As this gene is strongly tied to

brain development and neuron function, it is possible that an allele may be associated with Multiple Sclerosis development.

The final suggestive hits were found on chromosome 18. Chromosome 18 SNP rs1893251 had a p-value of 2.9×10^{-7}. The most closely located gene to this SNP was *CTAGE1* (within 2000 kb). *CTAGE1* produces the tumour antigen Cutaneous T Cell Lymphoma-Associated Antigen 1 (Koch *et al.*, 2003). Cutaneous T cell lymphomas are a group of non-Hodgkin's lymphomas that are derived from T cells. The development of this type of cancer has been noted as a side effect in treatment of the MS drug Fingolimod (Kappos *et al.*, 2015; Papathemeli *et al.*, 2016). Fingolimod is an immunomodulator and prevents the migration of T and B cells from the lymph nodes to circulate in the bloodstream. Although the association in this study is only suggestive, it may indicate a potential link between MS, Fingolimod and cutaneous T cell lymphoma.

Chromosome 18 SNP rs17602961 (p-value 5.3×10^{-6}) was in strong LD with several SNPs in nearby gene *MIR924HG* (also known as Long intergenic non-protein coding RNA 669), the only gene within the region. This gene has no known association with any disease in Northern Europeans, and so it is likely this result is due to chance.

The final chromosome 18 SNP, rs62096323, had a p-value of 4.2×10^{-6}. This SNP is within 0.1 Mb of (but not in LD with) known MS SNP rs7238078, on *MALT1*, which is essential for regulating the NF-κB pathway and known to have a role in Multiple Sclerosis development (Juilland and Thome, 2018). Mice which are deficient in Malt1 have impaired B and T cell responses and problems in the development of lymphocyte subsets, and are resistant to the induction of experimental autoimmune encephalitis, a mouse model of MS (Brüstle *et al.*, 2012; Mc Guire *et al.*, 2013). The dysregulation of *MALT1* can result in immunodeficiencies and autoimmunity, as well as psoriasis and cancer. As it has enzymatic activity, it has the potential to be targeted by drugs; a *MALT1* inhibitor has been used to relieve paralysis symptoms of MS (Jaworski *et al.*, 2014; Bornancin *et al.*, 2015). This study provides further evidence, although suggestive, of the involvement of *MALT1* in MS.

However, the discussion of these suggestive associations in this study remains speculative. In investigating a specific, small population such as Orkney and Shetland, identifying genome-wide significant associations will be challenging. In GWAS, one of

the most common ways to increase power is by enlarging the sample size (Sawcer *et al.*, 2011): Orkney and Shetland have small populations (~ 21,000 and 23,000 respectively), of which a small fraction have Multiple Sclerosis. As many of these cases are already present within the study, it is impossible to increase the sample size without waiting decades for more cases to arise.

Originally, this study set to identify if there were any SNPs which may have experienced a jackpot effect within the Northern Isles, where a large effect MS risk variant rarer elsewhere exists in the Northern Isles at a higher frequency, due to the effect of drift and low effective population size. These results suggest that because a real MS variant was detected at suggestive significance, there isn't a Northern-Isles specific common variant that is dominating the phenomenon of excess MS cases in the region. However, the lack of evidence produced here does not discount the idea that the jackpot effect may have increased the frequency of a rare MS risk allele or alleles to greater frequency. It may be the case that rare MS risk SNPs exist in this population at an increased frequency but were not detected due to small case numbers in this study. In general, the results were underpowered due to the limited number of MS cases.

Finding Implications

The GWAS did produce several suggestive results which may, given more cases, have been significantly associated with MS: at the least, this study has highlighted some potential regions of the genome for further investigation which have the potential to be validated in future studies. The loci identified have potential functions within the immune system; previous enrichment methods in MS have strongly implicated immune system cells, and it is thought that a large percentage of MS variants alter the regulation of immune-related genes (Patsopoulos, 2018). However, the clarification of these will need to utilise methods that go beyond GWAS. A principal goal of GWAS is to identify risk variants, and subsequently genes and causal pathways in disease risk and development. This has been achieved with relative success over the past 14 years with the use of moderate to large population samples, particularly with Multiple Sclerosis (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). However, the future of GWAS lies in using the genetics of hundreds of thousands of individuals with mega-biobanks. This will lead to the discovery of more variants (particularly those of smaller effect size and rare variants), account for more genetic variation and will give a

more complete picture of genetic architecture and understanding of biological pathways. In turn, genetic predictors will have improved accuracy and the ability to assess disease and develop diagnosis methods based on genetic testing will improve. Although this progression of GWAS is good for many aspects of understanding disease, small populations with unique or rare variants at a higher frequency will not benefit from this. As such, analysis methods for these populations should apply other methods which are more suited to a limited sample size. For example, whole genome sequencing is an expensive method of analysis, but could be applied to small populations to identify unique rare variants. Additionally, a regional heritability approach could be used by treating known susceptibility loci as the genomic region. This would capture more of the genetic variance and identify additional loci in comparison to a traditional GWAS.

The discovery of genetic associations is not the end goal with MS research; it is to identify causal genes and their functional consequences. Unfortunately, with small populations in which unique, rare variants may play an important role, GWAS is not always the most effective method to achieve this end goal.

Study Limitations

This GWAS was principally limited by sample size. In order to maximise sample size, there were two options: i) conduct a GWAS separately on each population and meta-analyse the results or ii) create a merged dataset and conduct one large GWAS. Both options had positive and negative aspects. The first option would avoid any errors that may arise in merging both datasets, however the second option would allow the GRM to give an overarching view of relatedness between ORCADES and VIKING. There is some migration between the two population groups, and it is possible that individuals have relatives within the other island group. The overlapping PCA plots from Chapter 2 give further evidence to support the second method. Therefore, it was decided that a merged dataset would be created, to account for this relatedness and reduce any potential bias in meta-analysing results from two non-independent populations. However, there is some issues surrounding this method. Primarily, ORCADES and VIKING were genotyped on different arrays. By pooling imputed results from the two cohorts, any heterogeneity between studies is amplified; even small individual effects due to platform specific errors such as genotyping batch effects may result in false positive

results. However, to counteract this, stringent QC methods (both before and after imputation) were carried out to remove any potential sources of bias.

4.5 Conclusion

This study carried out a genome-wide association study on 112 Multiple Sclerosis cases and 4223 controls from the Northern Isles of Scotland. This study is not able to provide substantial evidence towards the understanding of the high prevalence of MS in the Northern Isles, due to the low case numbers used in this study. However, it is unlikely that there is a dominating common variant beyond those that are already known in the literature (primarily those within the HLA region). There are several suggestive loci which do not have a previous association with MS. These may have arisen by chance or may be in LD with an unusual variant in Orkney. These suggestive loci are: i) a chromosome 2 SNP (rs1398972), found near lncRNA gene *LINC01117* (some lncRNAs have previously been linked to Multiple Sclerosis); ii) a chromosome 12 SNP (rs11055646) found within the *GRIN2B* gene, an important contributor to neurodevelopment; iii) a chromosome 18 SNP (rs1893251) found near *CTAGE1*, suggesting a possible link between *CTAGE1*, cutaneous T cell lymphoma and MS (cutaneous T cell lymphoma has been reported as a side effect of taking MS drug fingolimod) and iv) a chromosome 18 SNP rs62096323 which was found near *MALT1*, a known MS-associated gene. A further group of SNPs was found on chromosome 18 (lead SNP rs17602961), however there was no evidence of involvement in any related disease. Although these results remain speculative, they may be of interest for further research using alternative methods to GWAS.

Orkney and Shetland have some of the highest rates of Multiple Sclerosis in the world, and to date, a GWAS has not been conducted on this region as a whole. Although it was unlikely that any significant findings would be discovered in the dataset, given the number of cases, it was necessary to carry out this GWAS to rule out this possibility. To date, GWAS have been successful in reporting results for thousands of complex traits, including common diseases, quantitative traits, behavioural traits (such as educational attainment) and genomic measures (such as gene expression). However, this method of analysis favours large sample sizes in order to detect small effect and rare variants. The

results here do not rule out the possibility of unique, rare or jackpot variants contributing towards the high prevalence of MS in the islands; it merely highlights the limitations of the method applied. As such, the future for MS research in Orkney and Shetland lies with analysis methods more suited to small, isolated populations.

CHAPTER 5: CONTRIBUTION OF COMMON RISK VARIANTS TO MULTIPLE SCLEROSIS IN THE NORTHERN ISLES OF SCOTLAND

This chapter looks at the contribution of common risk variants to the prevalence of Multiple Sclerosis in Orkney, Shetland and mainland Scotland. The Northern Isles, particularly Orkney, have an excess of MS prevalence and this chapter seeks to determine the role common risk variants play in that excess risk. The hypothesis is that Orkney, and to a lesser extent Shetland, will have an increased burden of common risk variants for Multiple Sclerosis when compared to mainland Scotland.

5.1. Introduction

5.1.1 What are common risk variants?

Common risk variants are found in all human populations and can individually influence the risk of developing a complex disease by conferring a relatively small additive or multiplicative effect on a disease phenotype (Gibson, 2012). They remain prevalent in the population due to their small effect size, as it is typically only large effect alleles that are purged from the population due to selection (Reich and Lander, 2001).

The odds ratios of common risk variants typically lie between 1.1 and 2 (Bodmer and Bonilla, 2008), and can go down to 1.02 or less. When examined as a group, a collective assessment of their effect on disease risk can be determined. This combined effect can significantly impact disease risk; therefore, it is important to determine the genetic effects and cumulative risk of these variants when examining the genetic components of a complex disease.

The cumulative risk of common risk variants on developing a disease phenotype can be assessed using a polygenic risk score (PGRS) (Tesli *et al.*, 2014). Polygenic risk scores provide a method to summarise disease risk by using the estimated effect size to weigh the genetic dosage of each allele and produce an aggregate risk score including many susceptibility loci (Coleman *et al.*, 2016). This condensation of genome-wide SNP data

into a singular summary measure means that the detection of a genetic signal can be achieved using a lower sample size than would be required in a GWAS (Peyrot *et al.*, 2014): this is particularly useful when using a smaller dataset such as in this research.

To calculate a PGRS, an initial GWAS is carried out on a discovery sample and the odds ratios produced are then used to construct a score for every individual in an independent sample. The initial discovery sample GWAS determines the likelihood that each marker is associated with a specific disease phenotype (Dudbridge, 2013). It is conventional to use a previously conducted GWAS or GWAMA provided that there is no overlap of individuals between the GWAS and the population that the risk score is calculated for, as this may inflate PGRS results. Continental ancestry must also remain the same between the discovery sample and the independent sample to avoid population stratification, as discussed in **Chapter 2**. A large discovery sample size is favourable, as this will yield more accurate effect sizes for the markers and it is more likely that a PGRS calculated using these values will explain more disease variance (Dudbridge, 2013).

After choosing a set of results from an appropriate discovery GWAS, the SNPs are clumped to remove linkage disequilibrium. When a SNP is in linkage disequilibrium with another SNP in a sample, one locus is, in effect, represented more than once. This can bias results and lead to artificially inflated PGRS scores. Clumping can remove SNPs in linkage disequilibrium with another SNP by assessing blocks of SNPs within a specified window, identifying those which are in linkage disequilibrium with one another, then removing the least significant SNPs until only one SNP in the clump remains. Linkage disequilibrium is assessed using the r^2 value: the square of the correlation coefficient between two variables which denote the absence or presence of an allele at two loci (So and Sham, 2017). Typically, an r^2 value between 0.2 and 0.5 is the minimum value used to classify a pair of SNPs as in LD with one another (Ware *et al.*, no date; Mak *et al.*, 2017; So and Sham, 2017). More stringent thresholds are typically applied to larger (>1000) sets of SNPs where there is a higher chance of having SNPs in LD with one another.

Following the finalisation of a list of markers and their corresponding odds ratios, risk scores are calculated by multiplying the effect size by the number of risk alleles at a locus (none, one or two). The values at all loci are then summed for each individual,

resulting in a polygenic risk score unique to that subject. This is represented by the following equation:

$$PGRS_j = \sum_{i=1}^{n} log_{10}(OR_i) AD_{ij} \quad \textbf{(14)}$$

where ***PGRS*** is the polygenic risk score,

i is the susceptibility locus,

j is the individual,

n is the total number of SNPs,

OR is the OR of the risk allele and

AD is the allelic dosage of the risk allele.

Polygenic risk scores can be calculated over multiple p-value thresholds; for example, PGRS have the potential to include several SNPs which have previously reached a genome wide significance level, or hundreds and thousands to millions of SNPs across the genome which have not previously reached statistical significance, but nevertheless associate with the phenotype to some degree (Tesli *et al.*, 2014).

Risk scores can traditionally be used in two separate ways - to assess an association between the disease of interest and score in an independent sample, or to predict an individual's disease risk (Dudbridge, 2013). The former is carried out to determine if the markers used to construct the score are truly associated with the disease, and the latter is used to provide a more accurate, discriminatory predictor for disease by considering small effect markers as a whole (Wray, Goddard and Visscher, 2007), and to add to traditional, clinical risk factors. To assess disease prediction in binary traits such as Multiple Sclerosis, the area under the receiver-operator characteristic curve is calculated, as binary traits are assumed to come from a liability threshold model where a threshold exists that divides the group of individuals into the two separate trait categories; cases and controls (Neale, Neale and Ben, 2014). The ROC curve plots the

true positive fraction (also known as the sensitivity or probability of detection) against the false positive fraction (or 1-specificity or the probability of false alarm) (Hajian-Tilaki, 2013). The greater the value of the area under the curve, the more accurately the diagnostic test (in this case, PGRS) can discriminate between cases and controls (Hajian-Tilaki, 2013).

Each application for the PGRS has variable considerations to take into account: the power of association testing between PGRS and a disease phenotype depends on the sample size of the replication analysis, whereas determining predictions for individuals disease risk depends on the discovery sample size (Dudbridge, 2013).

5.1.2 The contribution of common risk variants to disease

The idea that common risk variants could explain the high levels of MS found in Orkney and Shetland was first suggested in 1981 by Compston, who implied that Orcadians in general may have a higher frequencies of common risk variants (Compston, 1981). Citing *HLA-B7*, *Dw2* and *DR2* specifically, he stated that controls in Orkney had a higher frequency of these variants than controls in the UK, Northern Europe and United States (Compston, 1981). This hypothesis had not previously been studied in Orkney, however common risk variants have previously been looked at in other populations for MS, with the principal study carried out by The International Multiple Sclerosis Genetics Consortium in 2010 (International Multiple Sclerosis Genetics Consortium, 2010). This study used two previously published datasets in their analysis: the IMSGC GWAS from 2007 (a study which revealed novel susceptibility locus *IL2RA* and confirmed *IL7R*) (Hafler *et al.*, 2007) and the Partners MS Centre GWAS data (De Jager, Jia, *et al.*, 2009). These datasets were chosen for their sample sizes (931/2431 and 806/2077 cases/controls, respectively), high quality control procedures and standardised MS diagnosis criteria (McDonald *et al.*, 2001; Hafler *et al.*, 2007; De Jager, Jia, *et al.*, 2009). The IMSGC GWAS was used as the discovery dataset, and the aggregate PGRS generated were used to conduct a logistic regression analysis to assess the disease status / score relationship (International Multiple Sclerosis Genetics Consortium, 2010). Varying SNP sets (determined by p-value threshold) significantly associated with disease status, suggesting that common risk variants within the genome

contribute to the polygenic inheritance of MS. Specifically, around 3% of the total variance for MS risk (as estimated by Nagelkerke's pseudo R^2) was explained by the combined effect of 12, 627 SNPs where all SNPs had $p < 0.2$ from the discovery sample. The Nagelkerke's pseudo R^2 value had a significant fit, where $p= 9.9 \times 10^{-19}$. This reaffirmed the idea that MS is partly caused by modest effect variants spread across the genome (International Multiple Sclerosis Genetics Consortium, 2010).

Polygenic risk scores have also been used to study other complex diseases, such as schizophrenia (Purcell *et al.*, 2014); case-control status for schizophrenia has been successfully predicted to a high level of statistical significance using PGRS (Purcell *et al.*, 2014). Previously, the common variants for this disease remained largely unknown with only a small number of significant markers identified (Purcell *et al.*, 2009). However, an initial GWAS identified many common variants, and an independent follow-up PGRS study was able determine an association between almost half the markers from an initial GWAS and the disease (Purcell *et al.*, 2009). These markers also associated with bipolar disorder, establishing a common polygenic basis for both diseases (Purcell *et al.*, 2009).

Polygenic risk scores are useful to highlight the potential burden that carrying multiple risk alleles can have on your disease susceptibility. For example, in prostate cancer, 36 validated genetic variants have been associated with disease risk, with the odds ratios varying from 1.1 to 1.6 (with the majority residing closer to an OR of 1) (Aly *et al.*, 2011). Many researchers have speculated against the clinical use of identifying these genetic markers, however it has been shown that carrying an increasing number of these markers can lead to an increased risk of developing a disease (Zheng *et al.*, 2008). PGRS allow genetic risk profiles to be calculated for individuals, which has the potential to lead to an improved accuracy of disease prediction. This was shown with prostate cancer: a PGRS was constructed using the most significant 35 SNPs associated with the disease (Aly *et al.*, 2011). When the PGRS was used to predict prostate cancer risk, 480 biopsies (22.7%) carried out could have been avoided than if the model had been used, with a cost of 3% of patients with an aggressive form of the disease being missed in diagnosis. By using the PGRS in risk prediction, unnecessary invasive biopsies could be avoided (Aly *et al.*, 2011). Using SNPs to inform a genetic risk profile in this manner

has a number of benefits; they are an affordable method of testing, easy to analyse, and they only need be measured once (Aly *et al.*, 2011).

Polygenic risk scores have multiple important uses, both within research and in a clinical setting. Most importantly for this study, they can be used to quantify and compare the genetic burden of common risk variants within and between populations.

5.1.3 Research aims

Extensive studies have proven that Multiple Sclerosis risk is heavily influenced by genetic variation, from the initially identified *HLA-DRB1*1501* to now over 200 loci associated with MS susceptibility. However, many aspects of Multiple Sclerosis, including the excess risk of MS in the Northern Isles of Scotland, are not well understood, despite Orkney having the highest confirmed prevalence of MS globally.

Populations such as Orkney and Shetland can remain relatively genetically independent from other populations through geographic isolation (Hatzikotoulas, Gilly and Zeggini, 2014). This can be compounded by many factors, including having multiple founding events (for example in Finland, which had 500- and 2000-year-old isolates in one region) (Peltonen, Palotie and Lange, 2000), variation in founder numbers, the occurrence of genetic bottlenecks, age of isolation events and endogamy levels (Hatzikotoulas, Gilly and Zeggini, 2014). These events uniquely shape the genetics of population isolates and can alter disease prevalence. This typically results in unique rare disease alleles; however, it is also possible for isolated populations to have different frequencies of common genetic variants. If the founders of Orkney, by chance, had higher frequencies of common risk variants for Multiple Sclerosis, this may explain the excess MS prevalence seen in the islands.

If the effect size of a discovered variant is small, the effect of this variant alone is often not significant to an individual. However, the cumulative effect of many variants can often be more meaningful to a person in the context of their own individual risk. This can bring understanding the genetic architecture of Multiple Sclerosis within Orkney and Shetland has the potential to influence and improve healthcare on the islands, and the progression of MS research and therapy. By identifying the structure of the

underlying genetic architecture of complex traits, both short and long-term research and healthcare benefits can be gained, both locally within the Northern Isles and for global disease understanding. If common risk variants are not the cause of the excess prevalence of MS in the Northern Isles, research can be focused on other genetic causes such as identifying unknown rare variants. Identification of rare variants can direct the search for novel therapeutic targets in drug development or diagnostic testing. Conversely, confirmation that common risk variants are a large contributor to excess Multiple Sclerosis prevalence in Orkney has the potential to improve disease risk and susceptibility prediction, leading to attention on prevention and early detection tactics. Generally, increased knowledge in the underlying causes of the genetics of Multiple Sclerosis can influence the design of future studies to further understand and implicate causal genes. This is particularly evident in drug development pipelines, where the majority of failures (less than 5% of potential drugs come to fruition) are due to insufficient disease models and improper knowledge of biology (Plenge, Scolnick and Altshuler, 2013).

The first chapter of this thesis discussed the various methods that have been used to understand the genetic architecture of MS in Orkney and Shetland. Both island groups have an excess of MS prevalence that has yet to be explained. This research in this chapter aims to uncover the impact of common risk variants for MS to elucidate if this is the reason why the prevalence of MS is significantly higher in the two island groups. Broadly, this chapter aims to answer the following question: how do common risk variants for MS in the Northern Isles compare to those in mainland Scotland?

Common risk variants will first be selected to compile the PGRS. These will be assessed individually between populations by calculating the allele frequencies of each individual variant and determining if there is a significant difference in any one variant between populations. The hypothesis is that Orkney, and to a lesser extent Shetland, should have a higher overall frequency of common risk variants than mainland Scotland.

These variants as a collective will then be used to construct a PGRS. It is expected that the tag SNP for the *HLA-DRB1* locus, which has the most significant association with MS, will have the largest impact on the PGRS. Therefore, PGRS will be constructed with and without this allele to assess its effect in contributing to the collective effect of common risk variants.

As a control measure, PGRS will be compared between cases and controls, with the hypothesis that cases will have a higher score than controls.

The contribution of common risk variants to explaining variance in MS risk will then be assessed, along with the ability of common risk variants to predicting MS status. This will determine the importance of common risk variants collectively in explaining MS risk and will indicate how useful these variants can be in a clinical setting, as well as providing an indicator of quality for the constructed PGRS.

Finally, the common risk variants will be compared at a population level between Orkney, Shetland and mainland Scotland. If the Northern Isles by chance have higher frequencies overall of the common MS risk variants, due to the population founders carrying more of these variants, it would be expected that they would have a significantly higher PGRS than the PGRS of mainland Scotland. The collective effect of the common risk variants may then help in explaining the increased prevalence seen in the Northern Isles. If the Northern Isles do not have a significantly higher burden from common risk variants when compared to mainland Scotland, then it may suggest specific (rare) genetic variants or environmental conditions, or some interaction of these, may be the cause.

5.1 Methods

5.2.1 Study population quality control

In addition to the standard quality control methods carried out on the study populations (see **Chapter 2**), several specific quality control measures for this analysis were implemented. Along with ORCADES and VIKING, the Generation Scotland (GS) cohort is used as a sample of individuals from mainland Scotland. This cohort contains over 24,000 individuals from regions including Glasgow, Ayrshire, Tayside and the North-East of Scotland (Smith *et al.*, 2013). There is potential for overlap between ORCADES and GS, as many individuals from Orkney live in the North-East of Scotland and could possibly be included in the Generation Scotland dataset. To ensure the GS dataset was representative of mainland Scotland only, in terms of latitude as well as

genetics, only individuals from Glasgow and Dundee were kept for this study, with the other individuals excluded.

5.2.2 Selecting common risk variants for MS risk

To assess the contribution of common risk variants to Multiple Sclerosis risk and compare this burden across populations (mainland Scotland, Orkney and Shetland), a comprehensive list of variants was compiled along with previously reported odds ratios. Within this list, it was important that the probability of linkage disequilibrium between variants was minimised, as LD allows the preferential retention of variants with more significant p-values, leading to biased and inaccurate results (Pasaniuc and Price, 2017). Therefore, strict quality control thresholds were applied to the list of variants.

Sourcing Common Risk Variants

To compile a list of common risk variants for Multiple Sclerosis, two key sources were used: the GWAS Catalogue (Welter *et al.*, 2014) and the 2011 International Multiple Sclerosis Genetics Consortium *Nature* GWAS (International Multiple Sclerosis Genetics Consortium, 2011). The GWAS Catalogue is a curated collection of all published GWAS results, and was the primary resource used to compile the list of SNPs. However, the GWAS Catalogue only curates SNPs with a minimum p-value threshold $<1 \times 10^{-5}$; therefore, the results from the largest MS GWAS were also examined to add any additional SNPs whose p-values did not pass this threshold. The 2011 IMSGC GWAS was chosen as this was the most recent and largest IMSGC GWAS that did not contain ORCADES individuals (ORCADES individuals were used in the 2013 IMSGC GWAS (Beecham *et al.*, 2013) and so using these results to generate a PGRS in ORCADES would have led to inflated results).

The raw GWAS Catalogue was downloaded (v1.0_e88_r2017-03-30) and SNPs were kept if the disease trait was listed as "Multiple Sclerosis" and the study that the data originated from only included European individuals (n SNPS = 175). Summary statistics for the 2011 IMSGC were obtained online at ImmunoBase (ImmunoBase, 2019); of the 102 SNPs listed, 82 were already listed in the GWAS Catalogue. The 20 remaining SNPs from the 2011 IMSGC paper which were not already present in the GWAS Catalogue were added to the discovery SNP set (n SNPs = 195).

Quality Control of the Common Risk Variant List

A total of 195 SNPs were compiled from both the GWAS Catalogue and the 2011 IMSGC GWAS results. However, several strict quality control procedures were applied to ensure that PGRS results produced from this list were not biased or inaccurate.

SNPs that did not have reported odds ratios and / or risk alleles were removed (leaving n SNPs =166). There were 12 pairs and 1 trio of duplicated SNPs, either due to being listed more than one time in the GWAS Catalogue or listed both in the GWAS Catalogue and in the IMSGC results. Inverse variance meta-analysis was performed within each duplicated group of SNPs, resulting in one SNP with a meta-analysed odds ratio and p-value to replace each group of duplicate SNPs (leaving n SNPs = 152). The inverse variance meta-analysis was calculated by first finding the log of the odds ratio for all SNPs within the duplicate group (the β). The inverse of the normal cumulative distribution was then calculated, using the following formula:

$$f = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \textbf{(15)}$$

where f is the inverse of the normal cumulative distribution,

μ is the mean of the distribution (in this case, 0),

σ is the standard deviation (in this case, 1) and

x is the independent variable for evaluation (the p-value determined from the source study).

The absolute value of this result was used to divide β for each SNP to find the standard error, which was in turn used to divide β to determine the z-score. The p-value for this z-score was obtained by applying the normal probability density function, with x defined as the absolute value of the z-score, and μ=0, σ=1, with the value multiplied by 2 to get the p-value. The precision estimate for each SNP was calculated by squaring the inverse of the standard error. These calculations were carried out for every SNP within each duplicate group. Finally, to obtain the meta-analytical values, the beta for each SNP was multiplied by the precision estimate. These were then summed within each

duplicated group then divided by the summed total of the precision estimate. Meta analysed standard errors were taken from the inverse of the square root of the precision estimate. Meta analysed z-scores were taken by dividing the meta analysed β by the meta analysed standard error. Finally, the meta analysed p-value was calculated as described above, with *x* equal to the meta-analysed z-score. The meta analysed odds ratios were taken by finding the exponent of the meta analysed β.

Of the 152 SNPs remaining, 15 SNPs were not present in the HRC-imputed data for ORCADES, VIKING and Generation Scotland. LDLink v3.3 (Machiela and Chanock, 2015), a tool designed to assess linkage disequilibrium using genotypic data from Phase 3 of the 1000 Genomes Project, was used to search for proxies for these SNPs. Each missing SNP was queried using its associated RSID and a specified 1000 Genomes population group, in this case Europe (which included the 1000 Genomes Project populations CEU; Northern Europeans from Utah, TSI; Tuscans from Italy, FIN; Finnish in Finland, GBR; British in England and Scotland and IBS; Iberian Population in Spain) (Clarke *et al.*, 2012). This returned a list of proxy variants for each SNP, ranked by r^2. If a proxy variant was present in the HRC-imputed datasets and had both an r^2 and D' above 0.90, the proxy variant was used in replacement of the missing SNP: 11 out of the 15 missing SNPs were replaced in this manner (n SNPs =148). Four SNPs were not replaced as they did not have a proxy SNP that matched the threshold requirements.

Finally, the remaining compiled SNPs were clumped for linkage disequilibrium, with a cut-off threshold of r^2=0.25 within a 200-kb window or clump, removing a further 21 SNPs (n SNPs =127). The LD values were calculated by determining the deviation of the observed haplotype frequency from the allele frequencies that were expected under equilibrium. Different r^2 thresholds were trialled in clumping (r^2=0.25, r^2=0.40, r^2=0.50) based on r^2 values used as cut-offs in previously published polygenic risk score studies; the effect of increasing the r^2 threshold from 0.25 to 0.50 allowed an additional 6 SNPs to be included in the risk score calculation, however the effect of this was not significant in the results, and so a stricter r^2 threshold was applied for the results.

This quality control procedure narrowed the compiled list of common risk variants from 195 to 127 and ensured that bias was kept to a minimum while retaining as many SNPs as possible.

5.2.3 Assessing the effect of the HLA-DRB1 *locus*

In assessing the effect of an aggregate collection of common risk variants, it was important to determine the contribution from the *HLA-DRB1* locus, the most strongly associated MS risk variant with by far the largest effect size (OR: 2.77, p-value: 3.2×10^{-199}).

To investigate the effect of *HLA-DRB1*, PGRS were calculated both with and without rs9271069 (an *HLA-DRB1* tag SNP; $R^2 = 1$, D' = 1), and risk scores were calculated for rs9271069 alone.

5.2.4 Calculating common risk variant frequencies

To provide more information when assessing the difference in PGRS between populations, the frequencies of the common risk variants in each population were calculated. At a population level, the mean PGRS ultimately depends on the frequency of each allele within the population. For example; if the PGRS of two hypothetical populations were compared and population A was found to have a mean PGRS that is significantly higher than population B, this could potentially indicate that there is a small number of high effect alleles that have a significantly higher frequency in population A (with the other alleles in a lower frequency), or that all alleles have a slightly higher frequency in population A: not enough to be significantly different, but enough to increase the PGRS overall. Calculating the frequencies of the common risk variants therefore gives more information as to the construction of the PGRS.

The frequency of all common risk variants were calculated in individuals without MS, allowing the general population frequency to be assessed without bias. Allelic count data was calculated for Generation Scotland, ORCADES and VIKING genotypic data using qctool v1.4 (Band and Marchini, 2018). The allelic counts were used to determine risk allele frequencies (RAF) using the formula:

$$\boldsymbol{RAF = \frac{2AA + AB}{2BB + 2AB + 2AA}} \quad \textbf{(16)}$$

where ***RAF*** is the risk allele frequency,

AA is the count of individuals homozygous for allele *A,*

AB is the count of heterozygous individuals and

BB is the count of individuals homozygous for allele *B*.

A comparison of RAF was made between Generation Scotland and both ORCADES and VIKING using a Pearson's chi-squared test, using the formula:

$$x^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad \textbf{(17)}$$

where x^2 is the chi-squared statistic,

O_i are the observed count of allele *i* and

E_i are the expected counts of allele *i*.

A Bonferroni-corrected threshold of 0.0004 (0.05 / 126) was used to determine significance following multiple testing at 126 loci.

5.2.5 Calculating polygenic risk scores

Polygenic risk scores were calculated using the R package PRSice (v1.25), which used the HRC-imputed dosage data for ORCADES, VIKING and GS as the target SNP set along with the effect sizes from the discovery SNP set to calculate risk scores for all individuals. Risk scores were calculated using Equation 14: an individual's dosage of a risk allele is weighted using the previously calculated log odds ratio to weight the dosage, and this is summed for every SNP for the individual.

Different groups of SNPs can give more accurate risk scores than others, and so PGRS were calculated for several groups of SNPs based on different thresholds for the reported p-value that accompanied the reported OR. Polygenic risk scores were

calculated at the following p-value thresholds (pT): pT 5 x 10^{-8} (n SNPs=61), pT 5 x 10^{-6} (n SNPs =101), pT 0.0005 (n SNPs =126) and pT 0.05 (n SNPs =127). Two main criteria for selection of a p-value threshold were in place: i) success in distinguishing between MS cases and controls (as determined by p-value significance) and ii) the value of Nagelkerke's pseudo R^2 when PGRS were used to predict MS status along with appropriate covariates (age, sex, principal components 1 and 2; full model description outlined in section 5.2.6). A better separation of cases and controls would indicate a more effective risk score, while a higher Nagelkerke's pseudo R^2 (a measure of explained variance) would indicate that the PGRS explains more variance in MS risk. All procedures listed in section 5.2.6 were carried out at each pT threshold group to determine which group would be the most accurate in estimating the role of common variants on Multiple Sclerosis status: descriptive statistics for each pT group can be found in Supplementary Table 2. All groups were successful in statistically differentiating between cases and controls at each population level, however pT 0.0005 had the lowest p-value score in Generation Scotland, ORCADES and VIKING for distinguishing between cases and controls (although this was a very minor difference). Additionally, when used to model PGRS as a predictor of MS status, all groups had a Nagelkerke's pseudo R^2 value within 0.004% of each other (both with and without covariates included). p-value threshold 0.0005 was consequently chosen as it had the lowest p-value in distinguishing cases and controls, however all thresholds were largely the same and the results produced by this p-value threshold group would likely be comparable to the results produced by any of the other three threshold groups. Therefore, the results in the subsequent sections will refer to the results produced from the SNPs included in this p-value threshold (n=126). As described in 5.2.3, PGRS were also produced for pT 0.0005 without *HLA-DRB1* tag SNP rs9271069 and for rs9271069 alone.

5.2.6 Determining the contribution of common risk variants to MS risk

From stage 5.2.5, PGRS were produced for Generation Scotland, ORCADES and VIKING for the p-value threshold group (pT) 0.0005 (n SNPs = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 alone (n SNPs = 1). Several statistical analysis methods were carried out to assess the validity and quality of these scores and

to quantify the effect of common risk variants on MS risk in the Northern Isles when compared to mainland Scotland.

Differentiating MS cases and controls with common risk variants

As a quality control measure to determine if cases and controls could be distinguished using polygenic risk scores, a case-control comparison was conducted within each dataset (Generation Scotland, ORCADES and VIKING) for each SNP group (pT 0.0005, pT 0.0005 without rs9271069 and rs9271069 alone). It was expected that MS cases would have significantly higher PGRS than MS controls.

Within each SNP group, the three datasets were first standardised by z-scoring the PGRS to allow comparison between populations. To z-score each individual's risk score, the following formula was applied:

$$\boldsymbol{z = \frac{(x - \mu)}{\sigma}} \tag{18}$$

where $\boldsymbol{z}$ is the z-score statistic,

$\boldsymbol{x}$ is an individual's risk score,

$\boldsymbol{\mu}$ is the mean PGRS for Generation Scotland, ORCADES and VIKING and

$\boldsymbol{\sigma}$ is the standard deviation for Generation Scotland, ORCADES and VIKING.

The mean z-scored PGRS value was calculated separately for each dataset, along with 95% confidence intervals, which were calculated using the formula:

$$\boldsymbol{CI = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}} \tag{19}$$

where $\boldsymbol{CI}$ is the confidence interval,

$\boldsymbol{\bar{x}}$ is the PGRS mean,

$\boldsymbol{\sigma}$ is the standard deviation of the PGRS

z is the critical value (here: 1.96) and

$\boldsymbol{n}$ is the sample size.

The mean z-scored PGRS values for cases and controls were then compared within each dataset using two-sample t-tests.

Cases and controls were compared between populations (and within SNP set groups) using two-sample t-test formula:

$$t = \frac{\bar{\bar{x}}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \textbf{(20)}$$

where $\boldsymbol{t}$ is the t statistic,

$\boldsymbol{1}$ refers to the first sample (here: cases),

$\boldsymbol{2}$ refers to the second sample (here: controls),

$\bar{x}$ are the sample means,

$\boldsymbol{\Delta}$ is the hypothesized difference between the case / control means (here, 0; to determine if cases and control means were equal)

$\boldsymbol{s}$ are the sample standard deviations and

$\boldsymbol{n}$ are the sample sizes.

Finally, a meta-analysis for cases and controls was performed to determine an estimate for the overall case / control PGRS. This was calculated using the formula:

$$\bar{\bar{T}}_{\cdot} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \quad \textbf{(21)}$$

where ***T*** is the effect size (here: the mean PGRS),

i is the dataset (here: Generation Scotland, ORCADES or VIKING),

k is the total number of datasets (here: 3) and

w is the weight (here: the sample size).

The corresponding meta-analysed 95% confidence intervals would be calculated as stated in Equation 19, with n equal to the sum of w_i.

Developing a PGRS MS risk model

Following on from confirmation of the validity of the PGRS in discriminating between cases and controls, a logistic regression model was developed to analyse the quality of the PGRS by assessing how much of MS risk variance could be explained by common risk variants and if common risk variants could be used to predict MS.

To investigate the association between PGRS and MS status, a logistic regression model was developed. However, logistic regression models do not account for genetic relatedness within the dataset, which can bias standard error values of the beta by incorrectly adjusting for type I error. To resolve this, related individuals were removed before fitting the data to a logistic regression model.

To remove related individuals from the dataset, the program KING v2.1 (Manichaikul *et al.*, 2010) was used to estimate kinship coefficients from pairwise relationships. The kinship coefficient is defined as $2\varphi_{ij} = \pi_{1ij}/2+\pi_{2ij}$ for two individuals *i* and *j*, where π_{1ij} and π_{2ij} respectively represent the probability that the two individuals share one and two alleles identical by descent (IBD). KING provides a robust algorithm to calculate the kinship coefficient by inferring pair-wise relationships by using allele frequencies to model the genetic distance between a pair of individuals (Manichaikul *et al.*, 2010). A kinship coefficient between 0.088-0.177 indicates a second degree relative, while a kinship coefficient between 0.044 and 0.088 indicates a third degree relative. A kinship coefficient cut off threshold of 0.05 was chosen for this study to maintain a balance between retaining information and removing related individuals, with any relationship which had a coefficient above 0.05 had an individual removed; cases were retained where possible to limit the loss of power.

The dataset containing polygenic risk scores was then merged with relevant covariate data and any individuals with missing data entries were removed. Following removal of related individuals and individuals with missing data, cases and controls in each dataset were as follows: 29/ 8341 Generation Scotland individuals, 14/642 VIKING individuals and 80/645 ORCADES individuals.

The following logistic regression model was fitted separately to each dataset for each SNP set using the R function *glm*:

$$\boldsymbol{log}\,\frac{\boldsymbol{p(x)}}{\boldsymbol{1-p(x)}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_3} + \boldsymbol{\beta_4 X_4} + \boldsymbol{\beta_5 X_5} \quad \textbf{(22)}$$

where $\boldsymbol{log}\,\frac{\boldsymbol{p(x)}}{\boldsymbol{1-p(x)}}$ is the logit,

$\boldsymbol{\beta_o}$ is the Y intercept,

$\boldsymbol{\beta_{1..5}}$ are the regression coefficients,

$\boldsymbol{X_1}$ is the first predictor variable; the polygenic risk score,

$\boldsymbol{X_2}$ is the first covariate; age,

$\boldsymbol{X_3}$ is the second covariate; sex,

$\boldsymbol{X_4}$ is the third covariate; principal component 1 and

$\boldsymbol{X_5}$ is the fourth covariate; principal component 2.

Age, sex and the first two principal components were included as covariates. Age can be predictive of MS as the disease onset typically occurs between the ages of 30 and 40. Sex was included as the sex ratio of women to men with MS is 2:1. Inclusion of the first two principal components adjusted for population structure, as discussed in **Chapter 2** (Wu *et al.*, 2011). The same model but without covariates was fitted to the data to serve as a comparison measure in later analyses, described below.

Determining how much variance in MS risk is explained by common risk variants
After fitting the logistic regression model to the data, the model results were used to assess how much variance in MS risk common risk variants could explain.

Nagelkerke's pseudo R^2 value was calculated for each dataset and SNP group using the model results to determine the proportion of variance explained by the common risk variants. This calculation is based on the Cox-Snell's R^2 calculation (Nagelkerke, 1991), which compares the log likelihood of the current model to the null model. The log likelihood is the natural logarithm of the likelihood function, which is the joint density of the variables but viewed as a function of the parameters of the statistical model. Nagelkerke's pseudo R^2 calculation follows the same formula, however it divides the Cox-Snell formulae by the upper bound to adjust the scale of the results to span from 0 to 1. Nagelkerke's pseudo R^2 was therefore calculated using the following formula:

$$\boldsymbol{Nagelkerke's pseudo\ R^2 = \frac{1 - (L_0 / L_M)^{2/n})}{1 - [p^p(1-p)^{(1-p)}]^2}} \quad \textbf{(23)}$$

where L_0 is the value of the likelihood function from the null model,

L_M is the value of the likelihood function from the full model,

$\boldsymbol{p}$ is the marginal proportion of cases with events and

$\boldsymbol{n}$ is the sample size.

Using common risk variants to predict MS
To determine if common risk variants could be a predictor of an individual's MS status, the model results were used to calculate a receiver-operator curve (ROC) and the area under this curve (AUC). Using common risk variants as a predictor for MS status not only suggests a potential use of common risk variants in a clinical setting, but it also gives an indication of quality for the PGRS by evaluating the model results from a different facet.

Receiver-operator curves were plotted for each dataset and SNP group, for the model with covariates and the model without covariates, to assess how well both the PGRS alone and the PGRS with covariates predicted MS diseases status.

The predictive capacity of each corresponding model was used to plot the fraction of true positive results against the fraction of false positive results (sensitivity versus l-specificity). This allowed a visual evaluation of the performance of the predictor in distinguishing between cases and controls, with a curve arching further from the centre line being more favourable.

In conjunction with the ROC curve, the area under the curve (AUC, or AUROC) was calculated to quantify the predictive ability of each model. The AUC was calculated as follows:

$$\boldsymbol{AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{1}_{p_i>p_j}} \quad \textbf{(24)}$$

where $\boldsymbol{AUC}$ is the area under the curve,

$\boldsymbol{i}$ iterates over every m data point with true class label 1 (MS case),

$\boldsymbol{j}$ iterates over every n data point with true class label 0 (MS control) and

$\boldsymbol{p_i}$ and $\boldsymbol{p_j}$ represent the probability score that was assigned by the model to data point $\boldsymbol{i}$ and $\boldsymbol{j}$, respectively.

The AUC values for each dataset were meta-analysed (following Equation 21) to give an overall predictive indicator for each SNP set.

Comparing common risk variants between mainland Scotland and the Northern Isles
Finally, the common risk variants were compared at a population level between Generation Scotland, ORCADES and VIKING, with the aim of determining i) if there was a significant difference at a general population level between the collective effect of common risk variants and ii) if there was a difference, how this compared to observed prevalence differences in the populations.

To assess the difference in common risk variants between each population, mean PGRS between individuals without MS were compared between Generation Scotland, ORCADES and VIKING. Cases were not used so as not to bias the results due to inflated numbers of cases included in each study. A comparison between populations was carried out visually via probability density plots of the mean PGRS and statistically through two-sample t-tests, with the calculation as described above (mean PGRS comparison between Generation Scotland and ORCADES, Generation Scotland and VIKING and ORCADES and VIKING, with Bonferroni corrected p-values of 0.016 applied and 95% confidence intervals calculated).

Finally, a comparison was made between the expected calculated difference in common risk variants between datasets and the observed prevalence differences seen in mainland Scotland, Orkney and Shetland. This comparison was carried out in several stages.

In the first stage, the mean PGRS for each dataset was calculated. For ease of comparison, the Generation Scotland PGRS mean was then set to 0, and the difference in means between Generation Scotland and VIKING, then Generation Scotland and ORCADES was calculated.

Secondly, the beta value from each of the dataset's model with covariates was meta-analysed and 95% confidence intervals were calculated, using the formula described in Equation 19. The beta produced from these models is the ratio between the effect from the initial GWAS and the actual effect on the population; as it is a polygenic risk score, if the initial beta values were correct it would be expected that the same effects in the population would be observed, and thus the beta would be 1. However, since the model provides an estimation of beta, this value can appear to fluctuate slightly. Therefore, a one-sample t-test was used on the meta-analysed beta values to verify that there was no significant deviation from 1. This was calculated using the following formula:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \quad \textbf{(25)}$$

where t is the t statistic,

$\bar{x}$ is the meta-analysed beta value,

s is the standard deviation,

n is the sample size and

μ is the specified mean (here: 1).

As there was no significant deviation from 1 in all the meta-analysed beta values, the beta value was set to 1 (Supplementary Figure 1).

The third stage was to work out the expected risk of MS from the common risk variants. This could be calculated from the confirmed beta value (the log of odds) and the mean difference in PGRS between both ORCADES and VIKING with Generation Scotland. If the beta value was 1, that would indicate that the log of odds of developing MS increased by 1 for every additional increase in 1 in the PGRS above the Y-intercept baseline value (this translates to an odds ratio of 2.7 ($\exp(\beta) = OR$; $\exp(1) = 2.7$)). To determine the expected risk of developing MS from common risk variants, this beta value was multiplied by the difference in mean PGRS values between ORCADES and VIKING with Generation Scotland; the proportion of risk that can be attributed to the difference in means. Thus, given the frequencies of the common risk variants that have been looked at in the PGRS, these values would be the expected increase in MS risk for the Northern Isles populations, using Generation Scotland (Glasgow/Dundee) as the baseline. This would reflect the genetic difference due to common risk variants.

The fourth stage was to determine how this expected difference in MS risk explained by common risk variants compared to the observed MS risk between populations. Observed MS prevalence data was obtained from Visser et al (Visser *et al.*, 2012). The Visser prevalence data is as follows: 145 per 100,000 for individuals in Glasgow, 295 per 100,000 in Shetland and 405 per 100,000 in Orkney. This data was used to create contingency tables to compare i) Glasgow and Shetland (
Table 9) and ii) Glasgow and Orkney (
Table 10).

	MS Case	MS Control
Shetland	295	99 705

Glasgow	145	99 855

Table 9: Contingency table for odds ratio calculations in Shetland and Glasgow
The contingency table for odds ratio calculations of Multiple Sclerosis in Shetland compared to Glasgow. The table uses prevalence data directly taken from Visser et al (Visser *et al.*, 2012).

	MS Case	**MS Control**
Orkney	402	99 598
Glasgow	145	99 855

Table 10: Contingency table for odds ratio calculations in Orkney and Glasgow
The contingency table for odds ratio calculations of Multiple Sclerosis in Orkney compared to Glasgow. The table uses prevalence data directly taken from Visser et al (Visser *et al.*, 2012).

The odds ratio was calculated based on the following formula:

$$\boldsymbol{OR} = \frac{(\boldsymbol{ad})}{(\boldsymbol{bc})} \tag{26}$$

where $\boldsymbol{OR}$ is the odds ratio and

$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{d}$ are based on the standard contingency table format (Table 11).

	Case	**Control**
Population 1	*a*	*b*
Population 2	*c*	*d*

Table 11: Standard contingency table format for odds ratio calculations
The standard contingency table format for odds ratio calculations of a disease (case / control) between the treatment group (population 1) and control group (population 2). Odds ratio calculations (see Equation 26) are taken directly from this table.

The odds ratio values were then directly converted into log of odds ratios by taking the natural logarithm of each value. Standard error values (Equation 27), and then 95% confidence interval values (Equation 28), were calculated as using the following formulae:

$$\boldsymbol{SE\{ln(OR)\}} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{a}} + \frac{\mathbf{1}}{\boldsymbol{b}} + \frac{\mathbf{1}}{\boldsymbol{c}} + \frac{\mathbf{1}}{\boldsymbol{d}}} \quad \textbf{(27)}$$

where $\boldsymbol{SE\{ln(OR)\}}$ is the standard error of the log of odds ratio (beta) and

$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{d}$ are based on the standard contingency table format (Table 11).

$$\mathbf{95\%}\ \boldsymbol{CI} = \boldsymbol{ln(OR)} \pm \mathbf{1.96} \times \boldsymbol{SE\{ln(OR)\}} \quad \textbf{(28)}$$

where $\mathbf{95\%}\ \boldsymbol{CI}$ is the 95% confidence interval,

ln(OR) is the natural log of odds ratio (beta value) and

SE{ln(OR)} is the standard error values of the beta value, as calculated in Equation 27.

To convert the expected log of odds risk due to each PGRS SNP set into prevalence data, the log of odds was converted into an odds ratio and substituted into Equation 26, setting *c* equal to 145 and *d* equal to 99,855 (the observed prevalence data for mainland Scotland), and solving for *a* (with *b*=(100,000-*a*)).

5.3 Results

5.3.1 How do individual common risk variants differ between populations?

The frequencies of all common risk variants were calculated in individuals without MS, with the aim of providing more information as to the composition of the PGRS calculation. A Pearson's chi-square test was then performed on risk allele frequencies between populations to determine if SNPs had a significantly higher frequency in any of the three population groups (with the null hypothesis that allele frequencies are the same, and the alternate hypothesis that allele frequencies are different).

The SNP with the highest associated MS risk, *HLA-DRB1* tag SNP rs9271069 (OR=2.77), had a significantly higher frequency in Orkney (RAF=0.23) and Shetland (RAF 0.21) than mainland Scotland (RAF 0.17: respective p-values of 8×10^{-13} and 2.3×10^{-6},

Table 12).

From the 126 common risk variants included in the study, Shetland and Orkney had 18 and 21 SNPs respectively that had a significantly higher frequency than mainland Scotland. Mainland Scotland had 20 and 12 SNPs higher than Shetland and Orkney, respectively (

Table 12). A full list of results (including non-significant results) can be found in Supplementary Table 3.

				RAF			Chi-Squared p-value	
RSID	**CHR**	**RA**	**OR**	**GS**	**ORC**	**VIK**	**GS / ORC**	**GS / VIK**
rs4648356	1	C	1.16	0.67	0.68	0.70	0.16	4.89×10^{-5}
rs11810217	1	A	1.15	0.27	0.26	0.24	0.17	8.86×10^{-5}
rs1323292	1	A	1.12	0.82	0.81	0.79	0.04	1.23×10^{-8}
rs7522462	1	G	1.11	0.70	0.68	0.66	3.66×10^{-3}	4.61×10^{-7}
rs6718520	2	A	1.17	0.46	0.40	0.47	7.72×10^{-14}	0.13
rs7592560	2	A	1.1	0.54	0.58	0.57	6.79×10^{-5}	0.01
rs17174870	2	G	1.1	0.74	0.74	0.71	0.98	9.00×10^{-7}
rs882300	2	C	1.19	0.53	0.51	0.50	0.03	2.97×10^{-5}
rs10201872	2	A	1.13	0.17	0.16	0.14	0.06	3.08×10^{-6}
rs9821630	3	G	1.09	0.28	0.30	0.25	1.79×10^{-3}	1.55×10^{-4}

rs11129295	3	T	1.11	0.35	0.42	0.37	2.33×10^{-20}	0.02
rs669607	3	C	1.13	0.47	0.43	0.43	$\mathbf{4.86 \times 10^{-8}}$	$\mathbf{7.04 \times 10^{-7}}$
rs1500710	3	A	1.09	0.58	0.53	0.58	$\mathbf{1.64 \times 10^{-11}}$	0.92
rs9657904	3	T	1.4	0.79	0.75	0.75	$\mathbf{2.28 \times 10^{-8}}$	$\mathbf{5.95 \times 10^{-9}}$
rs228614	4	G	1.09	0.54	0.59	0.54	6.71×10^{-8}	0.65
rs6821894	4	T	1.08	0.62	0.59	0.60	$\mathbf{3.09 \times 10^{-5}}$	0.01
rs6879677	5	A	1.08	0.40	0.36	0.37	$\mathbf{5.11 \times 10^{-6}}$	3.64×10^{-3}
rs1062158	5	A	1.09	0.62	0.59	0.65	2.48×10^{-4}	1.16×10^{-4}
rs10866713	5	A	1.17	0.21	0.20	0.18	0.02	$\mathbf{2.59 \times 10^{-6}}$
rs9260119	6	A	1.21	0.45	0.42	0.43	$\mathbf{1.97 \times 10^{-4}}$	0.01
rs9271069	6	A	2.77	0.17	0.23	0.21	3.36×10^{-5}	2.05×10^{-11}
rs12212193	6	G	1.09	0.45	0.53	0.47	5.53×10^{-23}	0.05
rs11962089	6	G	0.69	0.11	0.11	0.08	0.44	$\mathbf{4.65 \times 10^{-10}}$
rs802734	6	A	1.1	0.67	0.65	0.63	4.19×10^{-3}	$\mathbf{7.24 \times 10^{-8}}$
rs17066096	6	G	1.14	0.24	0.28	0.28	7.58×10^{-10}	8.07×10^{-8}
rs2066992	7	C	1.18	0.95	0.97	0.95	2.05×10^{-8}	0.89
rs354033	7	G	1.1	0.74	0.78	0.79	9.43×10^{-10}	4.33×10^{-14}
rs2019960	8	C	1.1	0.23	0.23	0.20	0.65	$\mathbf{1.77 \times 10^{-4}}$
rs2150702	9	G	1.16	0.49	0.54	0.53	4.18×10^{-8}	4.92×10^{-6}
rs3780792	9	G	1.6	0.33	0.34	0.36	0.15	7.43×10^{-5}
rs12722489	10	C	1.24	0.84	0.82	0.80	0.06	$\mathbf{1.24 \times 10^{-10}}$
rs793108	10	A	1.09	0.49	0.54	0.45	2.00×10^{-11}	$\mathbf{1.09 \times 10^{-4}}$
rs7912269	10	A	1.16	0.94	0.94	0.96	0.08	9.53×10^{-11}
rs7923837	10	G	1.1	0.62	0.66	0.64	2.80×10^{-7}	0.01
rs650258	11	C	1.12	0.64	0.66	0.61	0.02	$\mathbf{3.35 \times 10^{-5}}$
rs694739	11	A	1.08	0.62	0.62	0.56	0.75	$\mathbf{7.96 \times 10^{-13}}$
rs4409785	11	G	1.11	0.19	0.22	0.14	4.72×10^{-8}	$\mathbf{1.20 \times 10^{-13}}$
rs10466829	12	A	1.09	0.50	0.46	0.51	$\mathbf{1.40 \times 10^{-7}}$	0.26
rs17594362	13	T	1.12	0.12	0.12	0.09	0.39	$\mathbf{5.23 \times 10^{-7}}$
rs9596270	13	T	1.35	0.93	0.90	0.92	$\mathbf{8.14 \times 10^{-16}}$	1.81×10^{-3}
rs2300603	14	T	1.11	0.76	0.73	0.71	$\mathbf{4.75 \times 10^{-6}}$	$\mathbf{3.11 \times 10^{-12}}$
rs11864333	16	A	1.09	0.50	0.54	0.53	6.11×10^{-7}	1.31×10^{-4}
rs386965	16	G	1.09	0.21	0.21	0.27	0.70	5.93×10^{-18}
rs17445836	16	G	1.25	0.77	0.76	0.79	0.29	7.88×10^{-5}
rs9891119	17	C	1.1	0.35	0.33	0.39	0.01	2.65×10^{-7}
rs1373089	17	A	1.08	0.49	0.54	0.52	3.84×10^{-10}	9.49×10^{-5}
rs8081176	17	C	1.09	0.31	0.33	0.34	0.01	7.22×10^{-5}
rs12456021	18	A	1.1	0.19	0.23	0.18	4.70×10^{-9}	0.18
rs7238078	18	T	1.11	0.77	0.83	0.76	1.79×10^{-20}	0.38
rs1077667	19	C	1.16	0.78	0.82	0.80	3.82×10^{-7}	3.62×10^{-3}
rs2278442	19	A	1.08	0.65	0.62	0.69	$\mathbf{7.01 \times 10^{-5}}$	6.09×10^{-9}
rs8112449	19	G	1.1	0.67	0.70	0.65	1.15×10^{-4}	0.03
rs10411936	19	A	1.16	0.30	0.32	0.34	0.01	7.23×10^{-7}
rs874628	19	A	1.12	0.68	0.71	0.73	3.91×10^{-5}	9.26×10^{-12}
rs7255066	19	C	1.1	0.25	0.22	0.24	$\mathbf{6.29 \times 10^{-6}}$	0.11
rs281380	19	G	1.08	0.33	0.39	0.32	6.55×10^{-18}	0.17
rs2762932	20	G	1.15	0.14	0.15	0.17	0.19	3.41×10^{-10}
rs2248359	20	C	1.12	0.59	0.65	0.58	6.96×10^{-13}	0.30

Table 12: Risk allele frequencies in mainland Scotland, Orkney and Shetland
Risk allele frequencies (RAF) and Pearson's chi-squared test p-values for SNPs which have at least one significant result when comparing PGRS between populations. The chi-squared p-values are shown for two RAF comparisons: Generation Scotland (GS) and ORCADES (ORC); Generation Scotland and VIKING (VIK). Significant results, corrected for multiple testing at 126 loci over two population comparisons (corrected p-value significance level of $<1.98 \times 10^{-4}$), are shown in bold. Significant p-values from RAF that are higher in Generation Scotland than ORCADES or VIKING are highlighted in dark red, and significant p-values from RAF that are higher in ORCADES or VIKING than Generation Scotland are highlighted in light blue. *NB: RAF values are rounded to 2 decimal places. This can result in values (such as rs4149584) that appear the same due to rounding but give differing p-values between populations.*

5.3.2 Can common risk variants differentiate between MS cases and controls?

The common risk variants used in this study were all previously associated with MS risk: therefore, as a quality control measure mean PGRS in cases were compared to mean PGRS in controls, with the expectation that individuals with MS would have significantly higher risk scores than individuals without MS.

The mean z-scored PGRS probability densities of cases and controls are distinguishable in the full pT 0.0005 SNP set, the pT 0.0005 SNP set without rs9271069 and in rs9271069 alone (Figure 20). Within the full pT 0.0005 SNP set, cases across all three populations display a small frequency peak in the lower end of the distribution, indicating that there is a proportion of cases who have the same or lower mean PGRS as controls. VIKING has a less even distribution of cases, due to small sample size. When rs9271069 is removed from the pT 0.0005 group, the distribution of ORCADES and VIKING cases and controls remains similar, however the distribution for Generation Scotland significantly changes, with the distribution of cases displaying more individual peaks. When viewing the mean PGRS probability density of rs9271069 alone, the distribution of PGRS scores is similar between populations, however Generation Scotland has a higher frequency of controls with a raw PGRS equal to 0.

The confidence intervals for mean PGRS cases and controls do not overlap in any of the three population groups for SNP sets pT 0.0005 and pT 0.0005 without rs9271069

(Figure 21). An overlap in confidence intervals is seen between cases and controls in PGRS from rs9271069 alone. However, when mean PGRS scores in cases and controls are meta-analysed, there is no overlap in the confidence intervals in any of the SNP sets. The largest difference between meta-analysed PGRS is seen in the full pT 0.0005 SNP set, with cases having 0.67 (95% CI 0.52, 0.82) standard deviations from the mean (compared to pT 0.0005 without rs9271069 and rs9271069 alone, where mean PGRS in cases are 0.49 (95% CI 0.33, 0.64) and 0.53 (95% CI 0.35, 0.72) standard deviations from the mean, respectively).

In the full pT 0.0005 SNP set, the ORCADES controls have a PGRS mean 0.08 standard deviations above Generation Scotland and 0.02 standard deviations below VIKING. When the PGRS of rs9271069 alone are looked at, ORCADES controls have a PGRS of 0.21 standard deviations higher than Generation Scotland and 0.08 standard deviations higher than VIKING. However, when rs9271069 is taken out of the PGRS calculation, ORCADES controls have a mean PGRS score 0.02 standard deviations below Generation Scotland and 0.06 standard deviations below VIKING.

The same pattern is also seen amongst cases; in the full pT 0.0005 SNP set, the ORCADES cases have a PGRS score 0.06 standard deviations above Generation Scotland and 0.03 standard deviations below VIKING. In the PGRS of rs9271069 alone, ORCADES cases have the highest mean PGRS: 0.30 and 0.19 standard deviations above Generation Scotland and VIKING cases, respectively. However, as seen in the controls, when rs9271069 is taken out of the PGRS calculation, ORCADES cases fall to a mean PGRS score 0.08 standard deviations below Generation Scotland and 0.13 standard deviations below VIKING. (NB: confidence interval estimates have been omitted from this paragraph for ease of reading but can be viewed in Figure 21).

When MS cases and controls are compared within each dataset using two-sided t-tests (Table 13), they can be distinguished statistically in each population in the full pT 0.0005 SNP set (with p-values: Generation Scotland = 5.42×10^{-4}; ORCADES = 4.07×10^{-9}; VIKING = 2.93×10^{-3}). When rs9271069 is removed, only Generation Scotland (p-value 4.22×10^{-3}) and ORCADES (p-value 1.03×10^{-5}) are statistically different. When rs9271069 is considered alone, only ORCADES cases and controls are statistically different (p-value 1.53×10^{-4}). This reflects the number of cases; less cases leads to reduced power in each dataset.

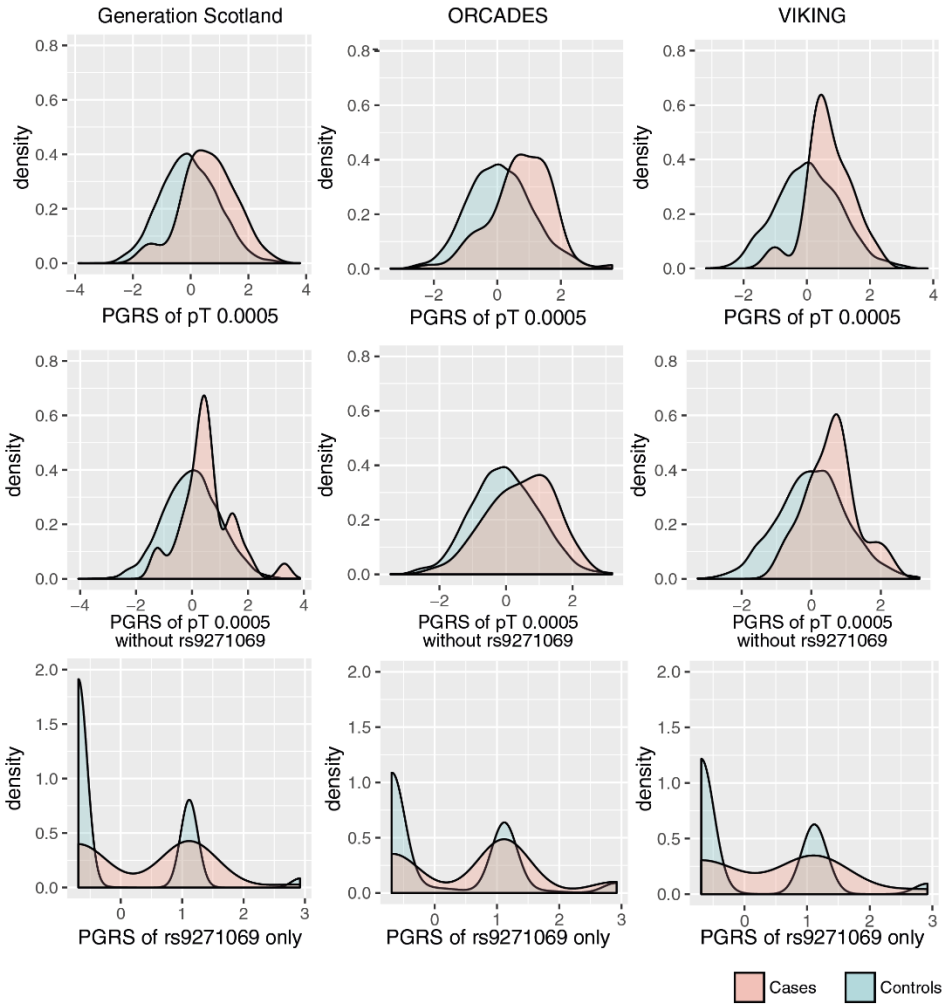


Figure 20: Probability density plots of z-scored polygenic risk scores

The probability density plots use z-scored polygenic risk scores (PGRS) for MS cases (n = 30) and controls in Generation Scotland (n cases = 30, n controls = 8708), ORCADES (n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158). Three SNP sets are used for comparison: pT 0.0005 (n=126); pT 0.0005 without rs9271069 (n=125); rs9271069 alone (n=1).

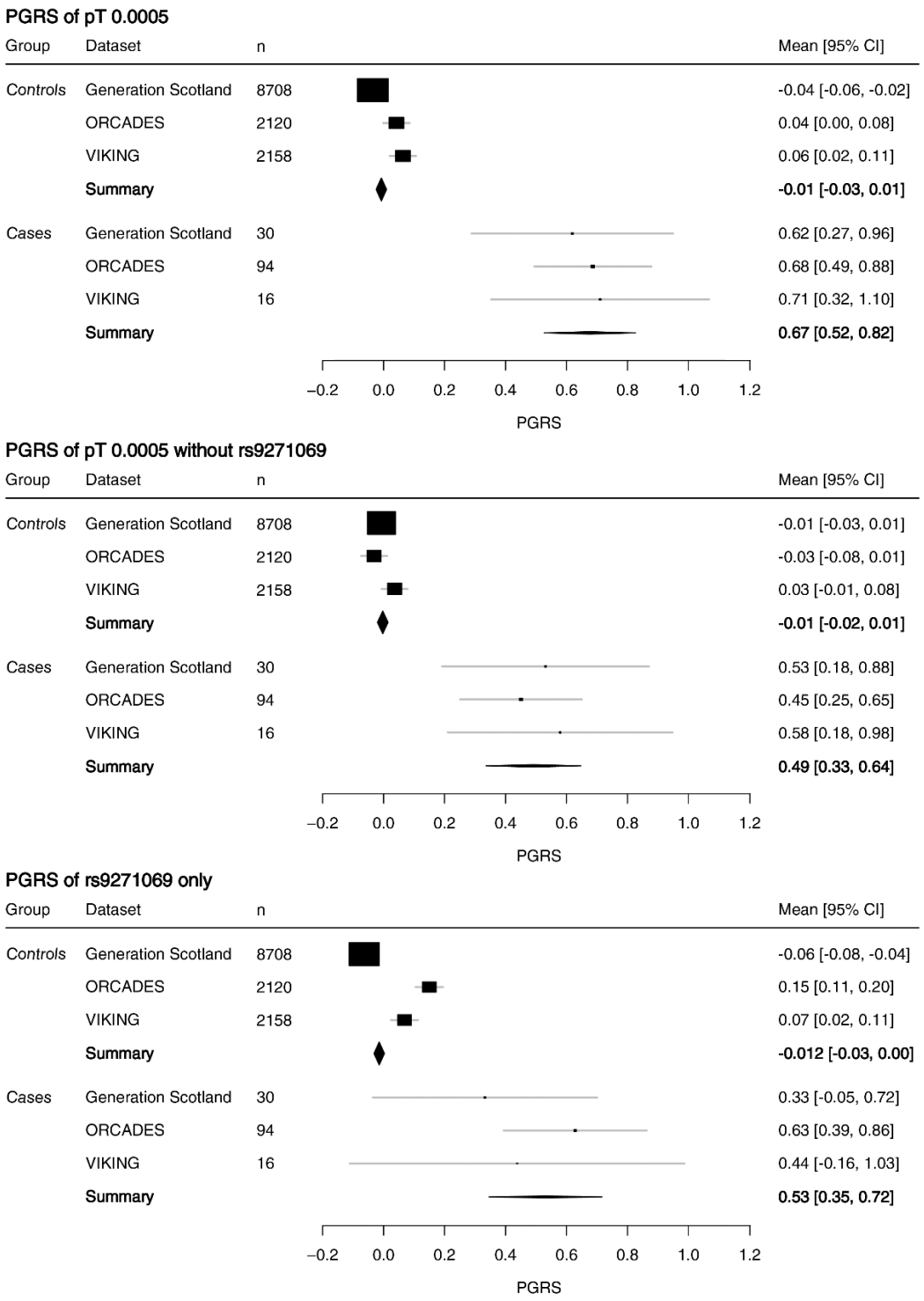


Figure 21: Forest plots of z-scored polygenic risk scores (PGRS)

The forest plots use z-scored polygenic risk scores (PGRS) for Multiple Sclerosis cases and controls in Generation Scotland, ORCADES and VIKING. Three SNP sets are used for comparison: pT 0.0005 (n=126); pT 0.0005 without rs9271069 (n=125); rs9271069 alone (n=1).

SNP set	n (SNP)	Dataset	n (cases)	n (controls)	t test statistic	p-value
pT 0.0005	126	GS	30	8708	-3.88	**5.42×10^{-4}**
		ORCADES	94	2120	-6.43	**4.07×10^{-9}**
		VIKING	16	2158	-3.53	**2.93×10^{-3}**
pT 0.0005 without rs9271069	125	GS	30	8708	-3.10	**4.22×10^{-3}**
		ORCADES	94	2120	-4.64	**1.03×10^{-5}**
		VIKING	16	2158	-2.87	0.01
rs9271069 only	1	GS	30	8708	-2.11	0.04
		ORCADES	94	2120	-3.94	**1.53×10^{-4}**
		VIKING	16	2158	-1.32	0.21

Table 13: Two-sided t-test results comparing MS cases and controls
Polygenic risk scores (PGRS) of Multiple Sclerosis cases and controls are compared within three datasets (Generation Scotland (GS); ORCADES and VIKING) using three groups of SNP sets (pT 0.0005, pT 0.0005 without rs9271069 and rs9271069 only). Significant results, corrected to account for 9 tests, are highlighted in bold (corrected p-value significance level of <0.005).

5.3.3 How much variance in MS risk do common risk variants explain?

A logistic regression model was developed to analyse the quality of the PGRS by assessing how much of MS risk variance could be explained by common risk variants. The logistic regression model (described in section 5.2.6: Equation 22) predicted Multiple Sclerosis status using PGRS, with age, sex and the first two principal components included as covariates. Full results for each model can be found in Supplementary Table 4.

Nagelkerke's pseudo R^2 was calculated to determine how much variance in MS risk common risk variants could explain (Figure 22). Without covariates, PGRS alone generally explained less than 5% of variance in MS risk: the full pT 0.0005 SNP set explained 3%, followed by pT 0.0005 without rs9271069 with 2% and rs9271069 alone with 1% variance explained. The addition of covariates in the model resulted in a higher explanation of variance, with the weighted means of each group increasing by an

average of 4% (R^2 values for pT 0.0005, pT 0.0005 without rs9271069 and rs9271069 alone at 8%, 6% and 5% for each SNP set, respectively).

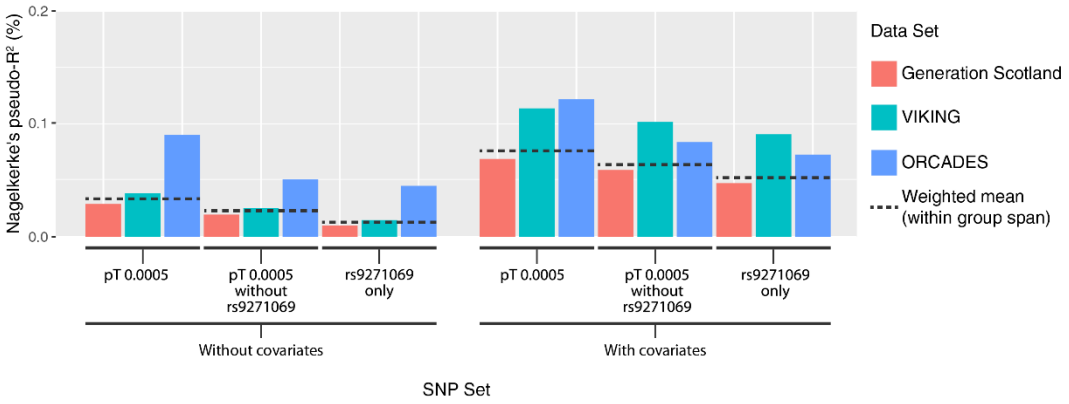


Figure 22: Nagelkerke's pseudo-R^2 results for the prediction of MS risk by PGRS
Nagelkerke's pseudo-R^2 results are shown for the prediction of Multiple Sclerosis risk in Generation Scotland (n cases = 30, n controls = 8708), ORCADES (n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158) using the model MS~PGRS and the model MS~PGRS+age+sex+PC1+PC2. Nagelkerke's pseudo R^2 is shown for scores derived using three SNP sets - pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1).

5.3.4 Can common risk variants predict MS status?

To determine if common risk variants could predict an individual's MS status, the results from the logistic regression model (described in section 5.2.6: Equation 22) were used to calculate a receiver operator curve (ROC) and the area under this curve (AUC).

The model with covariates was more successful in predicting MS status than just PGRS alone (Figure 23). This was reflected in the AUC values: for using PGRS alone to predict MS status, the meta-analysed AUC scores for each SNP group were 0.69 (95% CI 0.65, 0.74), 0.65 (95% CI 0.60, 0.69) and 0.62 (95% CI 0.57, 0.66) for the full pT 0.0005 group, pT 0.0005 without rs9271069 and rs9271069 alone (Supplementary Figure 2). When covariates were included, these scores increased to 0.74 (95% CI 0.70, 0.79), 0.73 (95% CI 0.68, 0.77) and 0.70 (95% CI 0.66, 0.75) respectively. The full pT 0.0005 SNP group had the most success in predicting MS status.

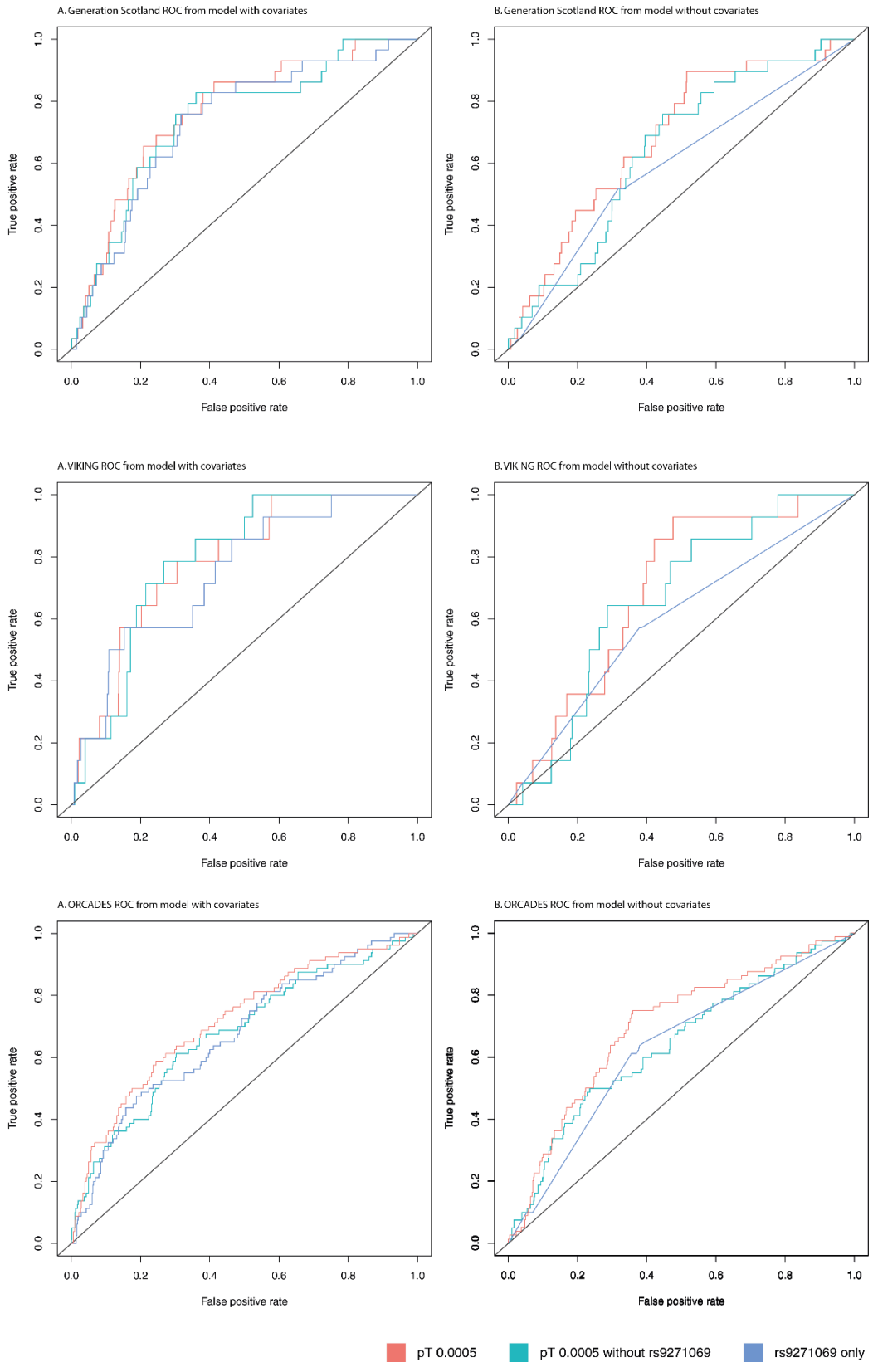

A. Generation Scotland ROC from model with covariates
B. Generation Scotland ROC from model without covariates
A. VIKING ROC from model with covariates
B. VIKING ROC from model without covariates
A. ORCADES ROC from model with covariates
B. ORCADES ROC from model without covariates
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
pT 0.0005
pT 0.0005 without rs9271069
rs9271069 only

Figure 23: ROC curves for predicting MS status

ROC curves showing the average predictive performance (the true positive rate and false positive rate) of predicting Multiple Sclerosis status using the model MS~PGRS+age+sex+PC1+PC2 and the model MS~PGRS for three SNP sets: pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1), using three datasets: Generation Scotland (n cases = 30, n controls = 8708), ORCADES (n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158).

5.3.5 How much do common risk variants contribute to excess MS risk in the Northern Isles?

The principal aim of this thesis chapter was to determine how much contribution common risk variants made to the excess risk of MS found in the Northern Isles of Scotland, particularly Orkney. Common risk variants were first compared in controls between the three population groups to determine if there was a significant difference in the overall effect of common risk variants. Consequently, that difference was quantified and the expected effect of common risk variants to excess MS risk in each population was calculated. This was then compared to the observed excess MS risk, with the aim of determining how much of the increased prevalence in the Northern Isles could be explained by common risk variants.

Do common risk variants for MS differ between mainland Scotland and the Northern Isles?

To determine the difference in common risk alleles between the three populations, the controls from Generation Scotland, ORCADES and VIKING were compared (only controls were compared to not bias the results with a high proportion of individuals with MS).

A statistical difference between Generation Scotland and both ORCADES and VIKING was seen in the full pT 0.0005 SNP set (p-values 1.97×10^{-5} and 3.16×10^{-5} respectively; Table 14). There was no statistical difference detected between ORCADES and VIKING in this SNP set. When rs9271069 was removed from the SNP set, there was no statistical difference detectable between any of the control populations. When rs9271069 was looked at alone, a statistical difference was seen when comparing all

populations (p-values for GS / ORCADES: 1.01×10^{-20}; GS / VIKING: 7.15×10^{-8}; ORCADES / VIKING 1.93×10^{-3}). However, when observing the probability density overlap of PGRS in controls, there is no obvious distinction between populations (

Figure 24).

SNP set	n (SNP)	Population (P1)	n (P1)	Population (P2)	n (P2)	t test statistic	p-value
pT 0.0005	126	GS	8708	ORCADES	2120	-4.27	**1.97×10^{-5}**
		GS	8708	VIKING	2158	-4.17	**3.16×10^{-5}**
		ORCADES	2120	VIKING	2158	0.05	0.96
pT 0.0005 without rs9271069	125	GS	8708	ORCADES	2120	0.26	0.79
		GS	8708	VIKING	2158	-1.87	0.06
		ORCADES	2120	VIKING	2158	-1.69	0.09
rs9271069 only	1	GS	8708	ORCADES	2120	-9.39	**1.01×10^{-20}**
		GS	8708	VIKING	2158	-5.40	**7.15×10^{-8}**
		ORCADES	2120	VIKING	2158	3.10	**1.93×10^{-3}**

Table 14: Comparison of PGRS of MS controls between populations
Two-sided t-test results comparing polygenic risk scores (PGRS) of Multiple Sclerosis controls between Generation Scotland, ORCADES and VIKING, using three groups of SNP sets pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1). Significant results (with significance corrected to account for 9 tests) are highlighted in bold.

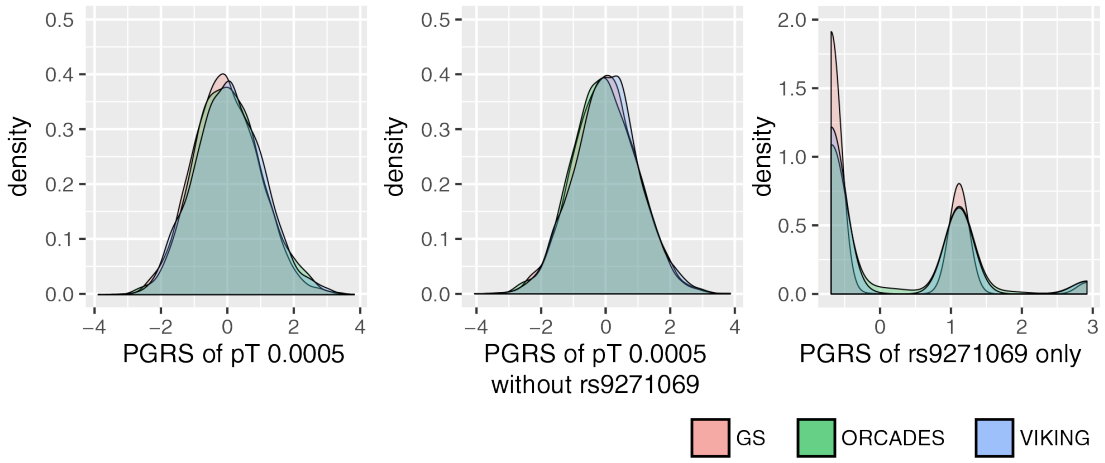


Figure 24: Population comparison of probability density plots of z-scored polygenic risk scores for MS controls

The probability density plots use z-scored polygenic risk scores (PGRS) for Multiple Sclerosis controls in Generation Scotland (n controls = 8708), ORCADES (n controls = 2120) and VIKING (n controls = 2158). Three SNP sets are used for comparison: pT 0.0005 (n=126); pT 0.0005 without rs9271069 (n=125); rs9271069 alone (n=1).

How much excess risk in MS in the Northern Isles is caused by common risk variants?

To determine the contribution of common risk variants to excess MS risk in the Northern Isles, a comparison was made between the calculated expected risk and the observed risk seen from MS prevalence data (Table 15).

The calculated difference between the mean PGRS using the full pT 0.0005 SNP set was 0.06 (95% CI 0.03, 0.09) between Generation Scotland controls and VIKING controls, and 0.05 (95% CI 0.02, 0.08) between Generation Scotland controls and ORCADES controls. Removing rs9271069 from the risk score calculation caused the mean PGRS in ORCADES controls to drop 0.01 (95% CI -0.04, 0.01) below Generation Scotland controls, however the mean for VIKING controls remained 0.02 (95% CI 0.00, 0.05) above Generation Scotland. Both these values did not significantly differ from 0. When rs9271069 was looked at alone, the calculated difference between the mean PGRS in controls of Generation Scotland and VIKING was 0.04 (95% CI 0.02, 0.05) and Generation Scotland and ORCADES was 0.06 (95% CI 0.05, 0.07).

To calculate the expected MS risk caused by these common risk variants, the mean values were multiplied by the logistic regression model (with covariates) beta (see section 5.2.6 for full methodology).

Therefore, in Shetland the expected log of odds ratios due to common risk variants is 0.06 (95% CI 0.03, 0.09), which can be compared to the observed log of odds ratio of 0.71 (95% CI 0.51, 0.91). This accounts for 9 cases (95% CI 5, 14) out of 150 observed excess cases per 100,000 individuals in Shetland (Figure 25). The majority of this expected risk is from the *HLA* SNP rs9271069, which contributes a log of odds ratio of 0.04 (95% CI 0.02, 0.05): equivalent to 6 cases (95% CI 3, 8) per 100,000 individuals.

In Orkney, all the expected excess risk is due to *HLA* SNP rs9271069, which contributes 0.06 (95% CI 0.05, 0.07) out of the observed 1.02 (95% CI 0.83, 1.21) log of odds ratios. This accounts for 9 cases (95% CI 8, 11) of the observed 257 excess cases per 100,000 individuals in Orkney.

	Generation Scotland	VIKING (95% CI)	ORCADES (95% CI)
Expected excess MS risk due to all common risk variants	0	0.06 (0.03, 0.09)	0.05 (0.02, 0.08)
Expected excess MS risk due to common risk variants without HLA SNP rs9271069	0	0.02 (0.00, 0.05)	-0.01 (-0.04, 0.01)
Expected excess MS risk due to HLA SNP rs9271069	0	0.04 (0.02,0.05)	0.06 (0.05,0.07)
Observed excess MS risk in populations	0	0.71 (0.51, 0.91)	1.02 (0.83, 1.21)

Table 15: The contribution of common risk variants to excess MS prevalence in the Northern Isles

Expected and observed excess MS risk (log of odds ratios) in both VIKING (Shetland; n controls = 2158) and ORCADES (Orkney; n controls = 2120) when compared to Generation Scotland, (n controls = 8708). Expected log(OR) values were calculated from the logistic regression results for the model MS ~ PGRS + age + sex + array + PC1 + PC2 by multiplying it with the mean PGRS calculated from the pT 0.0005 SNP set (n = 126) and rs9271069 alone (n=1). Significant differences between either ORCADES or VIKING and Generation Scotland for expected MS risk differences (taken from

Table **14**) are highlighted in bold. Observed log of odds values were calculated from the prevalence data found in the paper by Visser et al (Visser *et al.*, 2012).

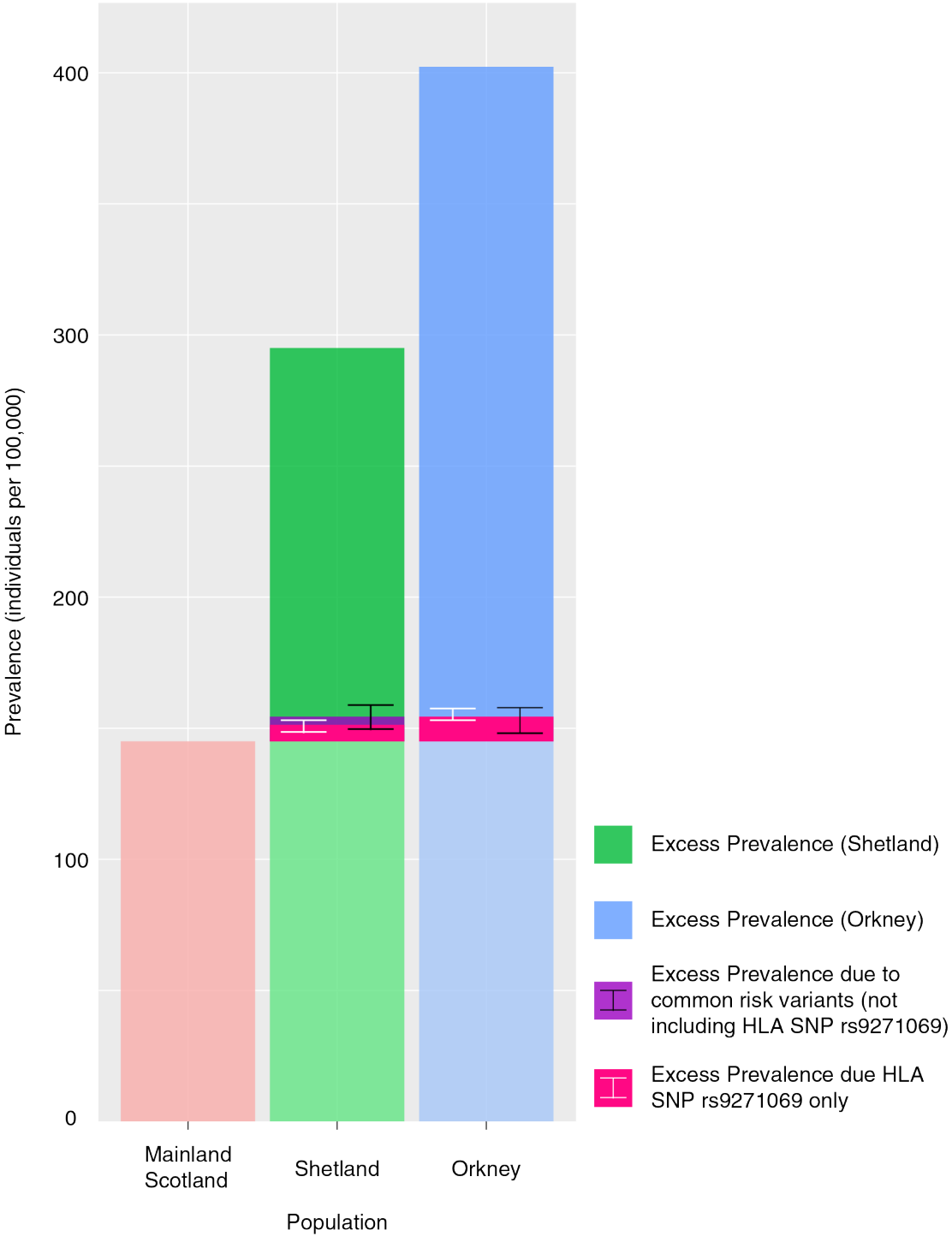


Figure 25: Excess prevalence of Multiple Sclerosis in the Northern Isles
Multiple Sclerosis prevalence (per 100,000 individuals) in Mainland Scotland, Shetland and Orkney, with data taken from Visser et al (Visser *et al.*, 2012). Excess prevalence is described as any additional prevalence in Orkney and Shetland that is over the baseline mainland Scotland prevalence. Excess prevalence due to common risk variants (not including HLA SNP rs9271069) and to HLA SNP rs9271069 alone is highlighted. These values are calculated as described in Methods section 5.2.6.

5.4 Discussion

Summary of Findings

Calculating polygenic risk scores for individuals in Orkney, Shetland and mainland Scotland allowed for the contribution of common risk variants to excess MS in the Northern Isles to be assessed. Orkney and to a lesser extent Shetland have an excessively high prevalence of MS, so it was expected that this could be explained by these islands having a higher frequency of common risk variants. These findings suggest that the cumulative risk from common risk variants is not a major contributing factor towards the excessively high rates of MS in Orkney or Shetland. However, a small proportion of risk can be attributed to a tag SNP for the *HLA-DRB1* haplotype *HLA-DRB1*1501*, the major genetic risk factor for MS.

As a collective, common risk variants do not make a significant or meaningful contribution to the excess of MS in the Northern Isles; any difference found in the risk scores when comparing between population groups is due to *HLA-DRB1*. The mean risk score in Orkney is higher than the risk score in mainland Scotland, conditional on the presence of the *HLA-DRB1* tag SNP; when this SNP is removed from the risk score calculations, no statistical difference between the controls in Orkney, Shetland and mainland Scotland is detected. Upon inclusion of this SNP in the risk score calculations, a statistical difference between risk scores in mainland Scotland and the Northern Isles populations is detected. If the Northern Isles had a higher burden of common risk variants than mainland Scotland, it would be expected that the statistical difference of PGRS scores would be maintained with the removal of the *HLA-DRB1* tag SNP, which is not the case. This indicates that without *HLA-DRB1*, the genetic effect of common risk variants as a contributor to excess MS prevalence is very weak and has no

meaningful impact on increased prevalence in the Northern Isles. In 1981, Compston found that there were higher frequencies of four common risk variants in Orkney (specifically *HLA*, *B7*, *Dw2* and *DR2*), and he was the first to suggest that Orkney may have generally higher frequencies of common risk variants than mainland Scotland (Compston, 1981). The results here suggest this is not the case, although they do not refute the findings made in 1981; several common risk variants such as the large-effect *HLA-DRB1* variant (which encodes the DR2 serotype studied in 1982) do have higher frequencies in Orkney, however as a collective the majority of variants do not have a significant impact on differential MS prevalence in the Northern Isles.

The contribution of HLA-DRB1

HLA-DRB1 tag SNP caused a significant difference between risk scores in mainland Scotland compared to the Northern Isles for two reasons: firstly, it has a large effect size, with a previously reported odds ratio of 2.77. Secondly, the frequency of the risk allele is significantly higher in Orkney (RAF = 0.23, p-value = 8×10^{-13}) and to a lesser extent Shetland (RAF = 0.21, p-value = 2.3×10^{-6}) than mainland Scotland (RAF = 0.17). This is a novel finding, as the frequency of this risk variant had not been previously measured in the Northern Isles. This frequency is higher than that previously reported in European heritage populations: Bahlo et al listed the frequency of this risk allele as 0.15, based on a combined sample of 4014 control individuals with reported European ancestry (Bahlo *et al.*, 2009). It is possible that the frequency of this allele is higher in the Northern Isles populations due to the founders of Orkney and Shetland having higher frequencies of the risk allele. Differing frequencies of single or a small number of variants in isolated populations compared to nearby non-isolated populations has been noted before in MS studies. For example, MS prevalence in Sardinia was compared to that of mainland Italy, and the findings indicated that higher frequencies of two TNF alpha gene variants in Sardinia contributed towards differences in MS prevalence (Wirz *et al.*, 2004). Conversely, the isolated population of the Sami was compared to the main Norwegian population, and lower frequencies of IL-10 risk variants between the Sami and Norwegians was suggested to contribute to the lower prevalence of MS seen in the Sami (Grytten Torkildsen *et al.*, 2008)

The contribution of the *HLA-DRB1* tag SNP, both by itself and with the other common risk variants, was quantified to determine the proportion of observed excess MS

prevalence that could be explained within the Northern Isles. A small proportion of excess MS in the Northern Isles is explained by all the common risk variants – however the majority of this is due to *HLA-DRB1*. The observed log of odds risk of MS for a Shetland individual in comparison to an individual in mainland Scotland was 0.71 – the proportion of log of odds risk that was explained by common risk variants was 0.06 (0.04 of which was explained by the *HLA-DRB1* tag SNP). In Orkney the observed log of odds risk of MS for an Orcadian in comparison to an individual in mainland Scotland was 1.03 – the proportion of log of odds risk that was explained by common risk variants was 0.05, slightly lower than the 0.06 explained by the *HLA-DRB1* tag SNP, as the expected risk from common risk variants not including this SNP was actually higher in mainland Scotland than Orkney. Although this proportion of explained excess risk appears small, the observed excess of MS risk captures both genetic and environmental contributors: considering that the majority of this expected excess in risk is comprised of the genetic effect of *HLA-DRB1*, this is a relatively important finding. Additionally, this is explaining more than is expected based on 3% of variance that is estimated to be explained by common risk variants, as indicated by the calculated Nagelkerke's pseudo R^2 value.

Variance explained by common risk variants

The amount of variance explained by common risk variants is an affirmation of a previous study finding from the IMSGC: in 2010, an IMSGC study reported that a polygenic risk score calculated from 12,627 SNPs explained roughly 3% of variance in MS risk (International Multiple Sclerosis Genetics Consortium, 2010). However, a study using 475,806 SNPs estimated around ~30% of MS risk variance to be attributable to common genetic variants (Watson *et al.*, 2012). This study uses 126 SNPs, which limits the amount of variance can be explained, although adding an additional ~12,500 SNPs does not appear to explain any additional variance. Three to five percent of total MS risk variance is small, however it should not be discounted as unimportant, particularly due to the relatively small number of SNPs that contribute to this variance percentage.

Predicting MS using common risk variants

Although common risk variants explained a small percentage of the variance of Multiple Sclerosis, they could possibly be used to predict MS status. To explore the prediction

capabilities of the common risk variants used in this study, the AUC was calculated: when using common risk variants alone to predict MS, the full common risk variant set had the highest success rate of 0.69 (95% CI 0.65, 0.74). When the covariates age, sex, PC1 and PC2 were included in this, the scores increased to 0.74 (95% CI 0.70, 0.79). Similar scores to this have been determined in the past; De Jager et al. in 2009 used 16 SNPs with a resulting AUC of 0.697, but this decreased to an AUC of 0.635 when replicated using smaller sample sizes (De Jager, Chibnik, *et al.*, 2009). A similar score of 0.69 was obtained by Jafari et al. in 2011 using 53 SNPs (Jafari *et al.*, 2011). Increasing the number of common risk variants has the potential to increase the predictive ability of common risk variants for MS.

Finding Implications

The majority of common risk variants, previously suggested to be a large influence on the excessively high prevalence in the Northern Isles, have been shown to have a similar collective effect size as those on mainland Scotland. Only the *HLA-DRB1* variant has been shown to differ between populations. The implications of this are significant for the people of Orkney and Shetland, as MS has a large bearing within the island groups, both clinically and socially. Here, it is not a burden of common risk variants that are causing the excess of MS prevalence. Research on MS in the Northern Isles specifically is not commonplace: the findings from this study update the knowledge of MS in the Northern Isles, which is important for individuals understanding why MS affects Orkney and Shetland so strongly. Furthermore, communicating the findings of this research to individuals in Orkney and Shetland has the potential to stimulate discussion generally on what the major risk factors of MS are, and could lead to promotion of better lifestyle choices (such as reducing smoking and taking vitamin D supplements) that could help in disease prevention.

Additionally, the findings here suggest research on MS in the Northern Isles should focus on other possible risk factors such as exploring the *HLA-DRB1* locus in further detail in Orkney and Shetland and identifying any possible rare variants which may be present. This is supported by the result that generally, only a small proportion of variation in MS is explained by known common risk factors.

The implications of these findings in a clinical setting are limited, as the common risk variants used here would not be an effective choice for targeting for treatment therapies,

and they can also not be used to predict MS with high enough accuracy. Although the use of common risk variants for indicating disease status have been used effectively for other diseases such as prostate cancer (as discussed in the introduction), a significantly larger collection of common risk variants would be needed to produce the same practical results.

Study Limitations

The number of common risk variants included here could be seen as a limitation. The inclusion of more common risk variants in the risk score calculation may not only increase the effectiveness of risk scores to predict MS, but it could also increase the proportion of observed log of odds risk that can be explained by common risk variants in the Northern Isles. If more common risk variants were included and the frequencies of these happened to differ between mainland Scotland and the Northern Isles, this could influence the impact of common risk variants between populations. The number of common variants used was therefore a limitation within this study. The common risk variants included here are informative as they capture a considerable proportion of the theoretical polygenic risk score that could be constructed using all MS risk variants, as the 126 SNPs used here had relatively high effect sizes. However, to capture the full effect of common risk scores between populations, a larger discovery GWAS providing a full list of variants would be needed. Repeating this analysis with a different discovery GWAS would therefore provide a useful addition to the findings in this study.

Another limitation with this study is the sample size and number of cases within each population, particularly within the Northern Isles groups. However, this remains the only cohort that has gathered genetic data from individuals in Orkney and Shetland, and as much of this is used as possible. Further data collection from the Northern Isles populations would therefore provide a great aid for clarifying the genetic basis of Multiple Sclerosis (as well as providing more information generally about complex disease within the islands). Finally, the risk score in this study is also partially based on imputed genotypes; the risk prediction algorithm could be improved if real genotypic data could be used.

5.5 Conclusion

This study sought to investigate the hypothesis that Orkney, and to a lesser extent, Shetland, have an excess of Multiple Sclerosis prevalence due to a higher frequency of common risk variants. Common risk variants as a collective do not make a significant or meaningful contribution to the excess of MS in the Northern Isles. Any significant difference seen in PGRS is due to the tag SNP for the *HLA-DRB1* locus. This risk variant has a significantly higher frequency in Orkney and, to a lesser extent, Shetland than mainland Scotland. This variant alone contributes 0.04 out of 0.06 of the log of odds risk of MS explained by common risk variants in Shetland, and all of the log of odds risk of MS explained by common risk variants in Orkney, explaining 0.06 log of odds ratios alone. Although this is a small proportion of the 0.71 observed log of odds risk of MS seen in Shetland and 1.03 observed log of odds risk of MS seen in Orkney, the common risk variants explain more than what was expected of them, as it was estimated that only 3% of variance in MS risk could be explained by common risk variants.

There is potential for improving both the quality of the risk score in predicting MS and the amount of excess risk in the Northern Isles explained by increasing the number of variants included in the risk score calculation. That, along with other improvements (such as using genotyped instead of imputed data and increasing the sample size through future additional data collection) would allow for an additional follow up study. Here, 9 out of 150 excess cases per 100,000 were explained in Shetland and 9 out of 257 excess cases per 100,000 were explained in Orkney, or about the equivalent of 2 patients in each archipelago. Therefore, although these findings explain some of the excess MS in the Northern Isles, additional studies focusing on other causes (both genetic and environmental) are needed.

CHAPTER 6: DISCUSSION AND CONCLUSION

6.1 Thesis findings

Chapter 3: Heritability of Multiple Sclerosis in the Northern Isles of Scotland

The Northern Isles of Scotland has one of the highest rates of Multiple Sclerosis in the world. We sought to understand the existence and extent of any genetic causes. My first finding estimated the SNP heritability in Orkney for MS at 0.31 (95% CI 0.13, 0.49). An estimate of SNP heritability for MS in Shetland could not be obtained. The current largest published estimate of SNP heritability for MS, using IMSGC data of 14,802 MS cases and 26,703 controls estimates SNP heritability as 0.19 (95% CI 0.18, 0.20) (International Multiple Sclerosis Genetics Consortium *et al.*, 2017). While heritability is a population-specific measure (IMSGC measured it within broadly European individuals) and it is possible that Orkney has a significantly higher heritability than that calculated by the IMSGC, it is likely that the true estimate of heritability in Orkney is much closer to that calculated by IMSGC. The IMSGC have a far larger number of cases resulting in a pinpointed estimate of heritability, while the estimate calculated in this thesis completely overlaps their estimate. Therefore, it is most likely that the heritability estimate calculated here is higher than the true heritability in Orkney.

Chapter 4: Genome Wide Association Study of Multiple Sclerosis in the Northern Isles of Scotland

I then considered trying to discover whether there were unique common variants in Orkney and Shetland using a GWAS of 112 MS cases and 4223 controls from the Northern Isles. I found 89 SNPs of suggestive significance, largely within six key regions of the genome. The six regions of suggestive significance were defined as having two or more SNPs within 1000 kb of one another that were below the suggestive significance threshold: these were found on chromosomes 2, 6, 12 and three regions on chromosome 18. As these SNPs did not reach the genome-wide significance level, it is important to determine if these results are real (either causative or in LD with a

causative SNP) or caused by chance. Within the literature, only one SNP (chromosome 6 SNP rs9268154) was previously associated with Multiple Sclerosis: in the 2018 IMSGC results (which used 32,367 MS cases and 36,012 controls) it has a listed p-value of 0 and it is also in moderate LD with *HLA-DRB1*1501* (r^2 = 0.65) (International Multiple Sclerosis Genetics Consortium *et al.*, 2018). The other five regions identified do not contain SNPs that are associated with MS, and the top SNP of each region has reported p-values in the 2018 IMSGC study ranging from 0.06 to 0.68. Based on this, it is likely that these regions appeared as suggestively significant due to chance. However, as four of the five regions lay in regions implicated in immune system functioning or have some previous link to an MS-related pathway, they are promising candidates for further studies in the Northern Isles of Scotland.

Chapter 5: Polygenic risk score study of Multiple Sclerosis in the Northern Isles of Scotland

Thirdly, I wanted to determine if the Northern Isles, by chance, have a higher frequency of common risk variants for MS. I found that a tag SNP for *HLA-DRB1*1501* (OR=2.77), rs9271069, had a significantly higher frequency in Orkney (RAF=0.23) and Shetland (RAF=0.21) than mainland Scotland (RAF=0.17), with respective p-values of 8×10^{-13} and 2.3×10^{-6}. Frequencies for this SNP reported in other cohorts appears to be similar to that in mainland Scotland or lower. TOPMed (n=125,568) and TwinsUK (n=3708) both report the frequency as 0.14, while 1000Genomes (n=5008) reports the frequency at 0.15 (Moayyeri *et al.*, 2013; The 1000 Genomes Project Consortium, 2015; National Center for Biotechnology Information, 2017). The significantly higher frequency of this SNP in the Northern Isles in comparison to both mainland Scotland and the general European population suggests that it plays an important role in the prevalence of MS in the Northern Isles.

When the polygenic risk scores from common MS risk variants were compared between controls from mainland Scotland and controls from Orkney and Shetland, a significant difference was found between populations (p-values 1.97×10^{-5} and 3.16×10^{-5} respectively). However, upon removal of the *HLA-DRB1*1501* tag SNP, rs9271069, from the polygenic risk score calculation, the statistical difference between populations disappeared. Any significant difference between mainland Scotland and Orkney / Shetland was caused by rs9271069. This was confirmed when rs9271069 was examined

alone, with a significant difference in the risk score from rs9271069 alone found between all population comparisons (Generation Scotland and ORCADES: 1.01×10^{-20}; Generation Scotland and VIKING: 7.15×10^{-8}; ORCADES and VIKING 1.93×10^{-3}).

When quantified, the contribution of rs9271069 in Shetland equates to causing 6 cases (95% CI 3, 8) out of 150 observed excess cases per 100,000 individuals. In Orkney, rs9271069 accounts for 9 cases (95% CI 8, 11) of the observed 257 excess cases per 100,000 individuals. This is approximately 4% of excess cases in both populations attributable to one SNP.

A study comparing polygenic MS risk scores of two control groups has not currently been published. However, in a 2016 study of polygenic risk scores, 103 common MS risk variants were compared between 452 multiplex family cases and sporadic cases (Mescheriakova *et al.*, 2016). Between the two groups of MS cases, a significant difference (p-value <0.0001) in risk scores was only found when the *HLA-DRB1*1501* locus was included in the risk score calculation. This study concluded that familial cases had a higher *HLA-DRB1*1501* allele frequency, leading to a greater burden of risk. it is possible that Orkney and Shetland have higher burdens of the *HLA-DRB1*1501* allele due to the strong family structures present within these population isolates, that is in turn causing an increase in the burden of MS.

In general, I found that the polygenic risk scores calculated with the 127 most strongly associated MS risk variants explained 3% of variance. This is approximately in line with previous research: 12,627 SNPs were able to account for approximately 3% of variance in an early IMSGC study (International Multiple Sclerosis Genetics Consortium, 2010). It would be expected that having 127 SNPs as opposed to 12,627 would result in a lower percentage of variance explained. However, the estimate of variance calculated in my results is increased by the higher frequency of the *HLA-DRB1*1501* SNP in the Northern Isles populations, which accounts for 1% of the 3% of explained variance.

Finally, I determined the predictive capacity of the polygenic risk scores for MS status as having an AUC score of 0.69 (95% CI 0.65, 0.74). This is generally in line with the scores previously published in literature. The 2016 study using 452 MS cases with 103 common risk variants estimated the AUC to be 0.72 (95% CI 0.69, 0.75) which overlaps with the results produced here (Mescheriakova *et al.*, 2016). Another 2016 study calculated the

predictive AUC using 78 cases, 121 unaffected siblings and 103 controls with 110 previously associated SNPs to produce an AUC of 0.82 (95% CI 0.75, 0.88) (Dobson *et al.*, 2016). This risk score appears relatively high, and the study notes that it is close to approaching the 0.85 AUC cut-off which is acceptable for clinically predicting a disease (Wilson *et al.*, 1998). However, that cut-off is based on more common, treatable diseases such as cardiovascular disease. Multiple Sclerosis is a (relatively) rare, heterogenous disease, and while it is useful to see to what degree risk scores could aid in prediction, a higher number of additional common variants or the incorporation of environmental factors into prediction models will be required to improve risk prediction.

6.2 What is causing the excessive prevalence of MS in the Northern Isles?

The Northern Isles have a high prevalence of MS and determining if predominantly environmental or genetic factors, or a mixture of both, are causing this is essential to disease understanding and formulating strategies for prevention in the region. Many environmental factors have been implicated generally in MS risk and development, including geographic region (specifically latitude), vitamin D levels, Epstein-Barr virus infection, smoking, obesity and gut microbiota. Most of these have been investigated within Orkney. The conventional environmental factors (such as vitamin D, latitude, vitamin D and smoking) do not appear to be explaining the excess prevalence, even if they contribute to individual risk.

Both the conventional genetic and environmental risk factors appear to be within normal bounds in the Northern Isles. This leads to ask the question: what is causing the excessive prevalence within this region? Logically, some factor, either genetic, environmental or both, differs in a way that is unique to the Northern Isles.

Latitude

MS rates increase as you move further from the equator, and we know that the Northern Isles have a high MS rate for their given latitude. Other regions on the same latitude (59°- 60°N) have substantially lower rates of MS: Rogaland, Norway – 176 per 100,000, South Estonia – 51 per 100,000, the majority of Russian territories - range from 30 to

70 per 100,000, across Canada – 260 per 100,000 (Gross, Kokk and Kaasik, 1993; Grytten, Torkildsen and Myhr, 2015; Widdifield *et al.*, 2015; Boyko *et al.*, 2016). In my opinion, the cause of the high rates of MS in the Northern Isles would not be due to an environmental factor that varies by latitude. If an environmental factor - either known to researchers or as yet unmeasured – that varied with latitude contributed to the excess of MS in Orkney and Shetland, the rates of MS would be more comparable to geographic locations of a similar latitude. As it stands, the prevalence of MS in the Northern Isles is much higher than locations of similar latitude, and so the cause of the excess is likely another factor. This is supported by the finding that vitamin D was not found to contribute to the excess of MS in the Northern Isles.

Population demographics

Women are more likely than men to develop MS. If the Northern Isles had a higher population of women than men, this may explain some of the excess. However, the ratio of men to women in Orkney (49.7:50.3) and Shetland (50.9:49.1) is comparable to the ratio found across Scotland (49:51) (National Records of Scotland, 2018), with Shetland being the only location in Scotland to have more men than women. Therefore, it is unlikely that the excess rate of MS is linked to gender.

In terms of population demographics, Orkney and Shetland do have higher numbers of the population aged 30 and above (70% and 66%) compared to mainland Scotland (65%). Orkney also has more people above the age of 80 (6%) than mainland Scotland (5%), although Shetland has less (4%) (National Records of Scotland, 2019). A greater percentage of the population above the age of onset of MS, combined with a population that appears to live longer, would lead to a higher prevalence of MS if the incidence rates were the same between Orkney and mainland Scotland. However, I do not think that the 5% higher increase in over-30s would equate to such a large difference in MS prevalence, although it may contribute a small number of cases per 100,000 individuals.

Viral exposure

Epstein-Barr virus is known to associate with MS, primarily due to links between infectious mononucleosis and evidence from seroepidemiological studies. Other viruses have also been implicated in MS (for example, human herpesvirus 6) and it may be possible that an unknown virus or other infectious agent has contributed (or continues

to contribute) to MS risk (Virtanen and Jacobson, 2012). Evidence for viral agents causing an increase in disease prevalence have previously been suggested in the Northern Isles. During the Second World War, a cluster of leukaemia occurred in children in Orkney and Shetland; the death rate for this disease was three times higher during a period where 60,000 British troops vastly outnumbered the local populace. It was suggested that this disease cluster was caused by an infectious, probably viral, agent brought from the British troops (Kinlen and Balkwill, 2001). Clusters of disease can often occur when isolated rural communities are exposed to a significant influx of outsiders. Is it possible that this WW2 occupation began a similar epidemic of MS in the Northern Isles? If Northern Isles children were exposed to an unfamiliar virus, this may have increased the risk of developing MS at a later age. Exposure to environmental factors during childhood are key for influencing risk of developing MS, as has been shown with migration studies: the incidence of developing MS has been shown to correlate with the risk of developing MS within the region of childhood residence (Compston and Coles, 2008). An individual who resides in a low-risk region but migrates to a high-risk region during adulthood will retain their low risk of developing MS (Gale and Martyn, no date). However, an individual who migrates from a low to high risk region during childhood (under 15 years of age) will acquire the high risk of their new host country (Gale and Martyn, no date). Therefore, individuals may have exposed children in the Northern Isles to a viral burden which, 30-40 years later, manifested in a higher prevalence of MS. This theory would be difficult to prove, given the extended time frame between childhood and the onset of development. It also appears that later cohorts of children continue to develop MS. Nevertheless, an infectious agent may contribute somewhat to the excess of MS cases in the islands.

Obesity

Obesity has been shown to influence MS and is a risk factor for the disease: a Mendelian Randomisation study showed that an increase of 1 standard deviation in BMI increased the odds of developing MS by 41% (95% CI 20, 66) (Mokry *et al.*, 2016). The exact mechanisms for how obesity increases the risk of MS is unclear, but it is possibly because individuals who are classed as overweight have a reduction in the amount of testosterone produced. Low testosterone has previously been associated with cognitive decline in MS, and in animal models testosterone seems to be protective against synaptic preservation by crossing the blood brain barrier to affect neuronal cells (Bove

et al., 2015). Additionally, individuals who are overweight generally have an increase in their immune response and potentially neuroinflammatory activity, as the body produces more proinflammatory mediators within fat tissues (Lumeng, Bodzin and Saltiel, 2007).

Orkney has the highest percentage of individuals who are classed as overweight and obese for any location in Scotland at 73% (95% CI 68, 78), while Shetland comes close with 71% (95% CI 64, 77). For comparison, the Scottish average is 65% (95% CI 64, 66) (Scottish Government, 2018). The higher levels of individuals classed as overweight and obese in Orkney and Shetland is likely to be reflected in a higher disease burden of MS, and I believe a reduction in the number of cases in the Northern Isles would occur if many islanders retained a BMI below 25 (not including BMI increases due to muscle mass). Given the difference in percentage between obesity levels in mainland Scotland and Orkney in particular, I think it is likely this contributes a small amount to the excess burden of MS.

Gut microbiota

Much attention has recently been given to gut microbiota and its role in Multiple Sclerosis (Cekanaviciute *et al.*, 2017; Chu *et al.*, 2018; Kirby and Ochoa-Repáraz, 2018; Mowry and Glenn, 2018). The gut microbiome is incredibly complex in both the way it impacts our cellular and immune system functioning, and the way it is influenced by external factors.

It is possible to theorise that the people of the Northern Isles may have a different gut microbiome composition to individuals in mainland Scotland, and it may be influencing the risk of MS. Gut microbiota in patients with MS differs from healthy individuals, with some microbiota found in MS cases associating with promoting inflammatory cytokines and general inflammation (Kirby and Ochoa-Repáraz, 2018). Loss of function of T regulatory cells and modulation of immune-mediated demyelination have also been reported as a result of changes to gut microbiome composition (Umeton *et al.*, 2018).

The gut microbiome is unique to each individual, but individuals from the same geographic region will generally have a more similar microbiome than individuals from different geographic regions. The geographic scale to which this extends has yet to be clarified, but it has been noted when comparing individuals from different countries

(Yatsunenko *et al.*, 2012). The gut microbiome is certainly influenced by the place of residence, along with diet and lifestyle factors (Chu *et al.*, 2018). If individuals in the Northern Isles have a specific factor which influences their gut microbiome enough for it to be significantly different from that in mainland Scotland, this may be enough to influence disease risk. However, two gut bacteria significantly associated with MS, *Akkermansia muciniphila* and *Acinetobacter calcoaceticus* (Cekanaviciute *et al.*, 2017), are unlikely to play a role in causing an increase in MS in the Northern Isles. *Akkermansia muciniphila* has been found to decrease with obesity; as individuals in Orkney are significantly more obese than individuals on mainland Scotland, it is unlikely that they have higher levels of this bacteria. Additionally, I do not believe the evidence for *Acinetobacter calcoaceticus* is strong enough to conclude it is a causal factor in the development of MS; it is an opportunistic bacterium found in patients with multiple underlying conditions, and so its presence is quite likely a consequence of MS. Additionally, much of the research carried out on gut bacteria is with a small sample size from a specific population group (for example, (Umeton *et al.*, 2018) which had samples from 42 cases). It would be interesting to investigate the composition of gut microbiota in the Northern Isles in comparison to that of mainland Scotland, as it has the potential to reveal interesting information about the gut microbiome and its relation to MS risk. However, a study of this nature would have more success on focusing on population differences (Northern Isles vs mainland Scotland) rather than trying to determine if specific bacteria were causing MS in the Northern Isles. With a disease such as Multiple Sclerosis, it is difficult to determine if gut bacteria are the cause rather than the consequence of developing MS. Other studies in the literature, for example, Kirby & Ochoa-Repáraz (2018), have found a difference between cases and controls for microbiome content within a population group but were unable to specify if the difference was cause or effect.

However, based on the current rather slim evidence, I do not think it is likely that gut microbiota play a significant role in determining MS risk in Orkney and Shetland. Although this remains to be investigated in future studies, I do not believe this should be given high priority in comparison to other potential areas of investigation such as rare genetic variants.

Hygiene hypothesis

To determine why the Northern Isles has such a high prevalence of MS, it can often be useful to look at other population isolates with similarly high frequencies. One of these is the island of Sardinia, which has an overall prevalence of 247.6 per 100,000 individuals (Pugliatti, Sotgiu and Rosati, 2002; Sotgiu et al., 2004), much higher than the surrounding low latitude and high vitamin D Mediterranean populations. The hygiene hypothesis was a theory proposed to help explain the high MS risk in this region. Within this population, the incidence rate of MS has increased substantially over the past forty years (Sotgiu et al., 2003) and the hygiene hypothesis was proposed to explain this. Post-World War II, a distinct lifestyle change (which included a malaria eradication campaign) left a population which had evolved to combat a multitude of parasites and pathogens with no need for such immunogenic mutations, leaving the Sardinian population prone to autoimmune diseases, where the immune system attacks the self (Sotgiu et al., 2003). This is unlikely to have happened in the Northern Isles. While it may be possible that the Northern Isles have undergone a lifestyle change in regards to increased use of antibiotics, this is i) not comparable to Sardinia given the lesser range of dangerous pathogens such as mosquito-borne diseases and ii) would likely not be so different from the rest of mainland Scotland and so would not explain why the prevalence differs between the two populations. Therefore, although this theory may help explain MS prevalence in Sardinia, I do not think it applies to the Northern Isles.

Unique or unmeasured environmental factor

Over 400 environmental risk factors have been examined in Multiple Sclerosis. These include exposure to environmental agents, traumatic events and accidents, vaccinations, surgeries, comorbidities and infectious, musculoskeletal and biochemical biomarkers (Belbasis *et al.*, 2015). It has been noted that many environmental risk factors appear to show differential association dependent on geographical region (Ramagopalan *et al.*, 2010). The Northern Isles may have a region-specific environmental factor that has not currently been measured. A similar study looking at geo-environmental exposures was conducted in South-West Sardinia (Monti *et al.*, 2016), an area which has a high prevalence of MS at 210 individuals per 100,000 (Cocco *et al.*, 2011). The study used geochemical data for heavy metals, along with UV exposure and urbanization data for all municipalities in the region.

They found an association between MS distribution and copper (Cu) levels; where for a 50 ppm increase in Cu concentrations, the adjusted odds of MS were 2.2 times higher (95% CI 1.3, 4.0, p = 0.006). Additionally, one village had a particularly high prevalence of MS (431 per 100,000) and one village had a particularly low prevalence of MS (46 per 100,000). The high prevalence village was associated with a high value of Cu (64.12±18.44 ppm) while the low prevalence village was associated with a low value of Cu (10.24±18.26 ppm). The study group hypothesized that environmental factors, such as Cu levels, along with the expression of genetic risk factors based on a specific environmental trigger may cause the geographical differences seen between regions in South West Sardinia. A similar study would be able to be carried out in the Northern Isles using data, if available, for environmental factors which have not previously been looked at, such as heavy metal data on the islands, and compare the results to mainland Scotland. Heavy metal data has been collected in Orkney and it has been noted that it has a higher concentration of heavy metals (Curtis and Simpson, 2001), as well as high concentrations of Uranium in the red sandstone (Smith and Great Britain. Environment Agency., 2003), as well as possibly radon gas, however the extent of this and their contribution, if any, to MS, is unknown.

Additionally, it may be possible that it is a completely unknown and unmeasured environmental factor that is unique to the Northern Isles. If this is the case, patterns in case data as time progresses may present more evidence to support this.

X chromosome

There has been some discussion over the involvement of the X chromosome in MS risk, given that the ratio of MS is three times higher in women than men (a ratio which is maintained within the Northern Isles). To date, only one variant has been identified as genome-wide significant (rs2807267) on the X chromosome, and with an odds ratio of 1.07 this would not explain differences by sex. It is very unlikely that the X chromosome of people in the Northern Isles has some significant and unique involvement in the excess of MS in the region, given the lack of evidence for genetic involvement of the sex chromosomes in general. Additionally, significant involvement of the X chromosomes has the potential to skew the ratio of men to women on the islands, which as it stands, does not differ from the ratio found for MS worldwide.

Gene-gene interactions

Gene-gene interactions have been shown to contribute a small amount of risk for MS. Principally, interactions between haplotypes *HLA-DQA1*01:01-HLA-DRB1*15:01* and *HLA-DQB1*03:01-HLA-DQB1*03:02* have been shown in a large-scale study with 17,465 cases and 30,385 controls (Moutsianas *et al.*, 2015). The increased frequency of *HLA-DRB1*1501* in the Northern Isles may lead to a slight increase in such interactive effects. Outside the HLA region, a study recently found an MS-associated *IL7R* variant interacted with an unknown, non-MS related variant in *DDX39B* (Galarza-Muñoz *et al.*, 2017). This interaction increased the risk of MS by almost three times due to the over production of the protein sIL7R which increased MS risk. It may be possible that a unique gene-gene interaction similar to this exists in the Northern Isles. If a variant that is rare elsewhere but common in the Northern Isles region interacts with another variant to produce an effect on an immunological or pro-inflammatory pathway, it may increase the risk of MS in a way that would not be seen in mainland Scotland if the interactive variant was of low frequency. The network of immune cells implicated in Multiple Sclerosis involves a vast array of molecules and cell types, including cytokines, signaling molecules and antigen receptors, whose interactions can lead to an almost infinite variation of functional outcomes, including redundancy which is often the case with MHC interactive effects. This complex system gives rise for many opportunities for gene-gene interaction to occur and this in turn has the potential to be unique to a population that differs in allele frequencies. Additionally, gene interactions may occur even in the absence of statically significant individual main effects (Motsinger *et al.*, 2007), and so an undetected common variant that is unique to the Northern Isles may have an interactive effect that contributes to a higher rate of MS in the region. Disruption in immunological and pro-inflammatory pathways due to gene-gene interactions has previously been linked to some cases of familial MS, and so it is not difficult to imagine a similar scenario occurring in the Northern Isles (Vilariño-Güell *et al.*, 2019).

However, there is no known MS hotspot which has, to date, been explained by gene-gene interactive effects. Although it's understood that interactive effects do contribute a small proportion to MS risk, this contribution is limited. Thus, while it is possible that unique gene-gene interactions exist on Orkney, these would likely not be enough to explain such a high proportion of excess prevalence.

Detecting gene-gene interactions within the Northern Isles would be very difficult, given the limited sample size. To detect interactions between genes, a much larger sample size is needed than simply to find individual effects. For example, in the very basic scenario of examining the interaction between two SNPs, with three genotype combinations each, these gives 9 genotype combinations: increasing the number of SNPs in each interaction term leads to exponentially bigger results. If any interactive effects do exist in the Northern Isles, these would therefore be very difficult to detect.

Gene-environment interactions

Much evidence has been presented to support the effect of interactions between genetics and environmental factors in MS. For vitamin D, the vitamin D response element (VDRE) is found in the promotor region for *HLA-DRB1*; thus it is very likely that the expression of *HLA-DRB1* is controlled by vitamin D (Ramagopalan, Maugeri, *et al.*, 2009). In EBV, an interaction between increased Epstein–Barr nuclear antigen 1 titres and HLA variants has been found to increase MS susceptibility (De Jager, Chibnik, *et al.*, 2009). Tobacco smoke exposure has been associated with greater MS risk, with 42 *NAT1* variants showing evidence for interaction (Napier *et al.*, 2016). Heavy metal exposure has been identified as interacting with MS variants in 5 genes examined in a study of 217 cases, including *VDR*, *TNF-α*, *TNF-β*, *MbP* and *APOE* (Napier *et al.*, 2016). Overweight individuals with a BMI of >27 kg/m^2 who had 1 or 2 *HLA-DRB1*15* risk variants were reported to have a seven-fold increased risk of MS in comparison with individuals who had no *HLA-DRB1*15* risk variants and a BMI less than 27 kg/m^2 (Gianfrancesco and Barcellos, 2016). These studies were conducted at least 200 individuals and suggest the presence of interaction between MS risk variants and the most influential environmental factors. It is therefore very likely that similar interactions exist in the Northern Isles, but to what degree these are affecting the MS prevalence is unknown. It is my suggestion that the increased frequency of the *HLA-DRB1*1501* variant combined with a high rate of overweight individuals will lead to a greater contribution of interactive effects to MS prevalence than in mainland Scotland. Investigating this is challenging due to the small sample size of cases, but prioritisation of specific environmental factors such as obesity would allow for a more streamlined approach. If gene-environment interactions were found to be significant MS risk factors in the Northern Isles, this would help effective strategies for MS prevention to be designed.

It may be the case that gene-environment interactions are not contributing to the excess prevalence in the Northern Isles. While gene-environment interactions may explain hot spots in other diseases, such as chronic kidney disease (Friedman, 2018), interactions have not previously explained any MS hotspots. However, Orkney is unique in terms of its combination of genetic and environmental factors, and I think it is likely that some, even if it is a small amount, of the excess prevalence could be explained gene-environment interactions (particularly involving obesity).

Comorbidities

Often, individuals who develop MS acquire other immune-mediated inflammatory diseases or have close family members that do (Nielsen *et al.*, 2008). This can be due to diseases sharing similar pathways, with overlapping genetics or environmental risk factors (Sawcer *et al.*, 2011). A high prevalence of MS may mean that there is a high prevalence of an associated disease; if so, this may give an indication as to the reason why MS is so high in the islands.

Multiple diseases have been associated with MS, however not every disease has a measured prevalence in Orkney and/or Shetland. Of diseases previously associated with MS, rheumatoid arthritis, Crohn's disease and diabetes each have an estimated prevalence for the Northern Isles. For rheumatoid arthritis, the Northern Isles do have a slightly higher prevalence (747 per 100,000 individuals compared to mainland Scotland's 708 per 100,000 individuals), however it does not show the same excess in case numbers that found in Multiple Sclerosis (Scottish Public Health Record, 2012). Crohn's disease had a slightly lower prevalence in the Northern Isles (136 per 100,000) than North-Eastern Scotland (147 per 100,000), although this was measured during the 1980s and prevalence rates may have changed since this time (Kyle, 1992). Diabetes was also at a slightly lower prevalence in the Northern Isles (4700 and 4800 per 100,00 in Orkney and Shetland) compared to mainland Scotland (5400 per 100,000) (NHS Scotland, 2016).

Although this only gives a picture of three diseases, they are some of the most strongly associated with MS. Additionally, both Orkney and Shetland are very community-based populations, and as such it is more obvious when a disease is at a particularly high prevalence. Multiple Sclerosis remains the principal disease of focus on the islands as so many residents are aware of its high prevalence due to local knowledge alone. There

has not been a similar disease noted by residents for its high prevalence. Although it would be useful to record the prevalence of other diseases specifically as island-specific prevalence rates, local knowledge should not be discounted as an indicator of disease spread.

As it does not appear that any other disease has such a high prevalence to the Northern Isles when compared to nearby regions (such as mainland Scotland), it suggests that the disease mechanism that is causing the excess prevalence in the Northern Isles is specific to Multiple Sclerosis. This insinuates that the cause of the excess prevalence on the islands are population-specific, but also disease-specific. If an environmental factor was solely to cause excess prevalence, it would likely have some effect on other, related diseases. For example: an excess of heavy metals would not just cause an increase in Multiple Sclerosis on the islands, but also other diseases, for which the risk is also increased by heavy metal exposure. Additionally, it is likely not one factor that is shared between multiple diseases that is causing MS. It is more likely that it is a combination of factors – some may be shared with other diseases, but not all – or an entirely MS-specific factor, such as rare genetic susceptibility variants.

Epigenetics

Epigenetics is the study of phenotypic changes caused by the modification of gene expression by alterations to the genome that do not change the nucleotide code, for example histone acetylation and DNA methylation (Petronis, 2010). A recent study found significant variation of DNA methylation within human groups in a single continent (Giuliani *et al.*, 2016). The variation in each group was shaped by environmental factors such as nutrients, UVA exposure and pathogen load, with ecological niches influencing an individual's DNA methylation profile. The geographical scale over which this variation extends (for example, mainland Scotland verses the Northern Isles) has yet to be seen. Several MS-associated environmental factors, such as vitamin D, dietary factors, smoking, EBV infection and UV radiation have been shown to alter DNA at an epigenetic level (Rito *et al.*, 2018). If these factors can change DNA methylation and cause other epigenetic changes, and they differ at a local level, it may be possible that there is a different epigenetic profile for individuals living in the Northern Isles as opposed to mainland Scotland. This would be further affected by the potential inheritance of epigenetic factors (Perez and Lehner, 2019); as the Northern

Isles was more isolated in the past, it may lead to a greater difference in environmental factors influencing epigenetics when compared to mainland Scotland.

However, there are two key questions when considering epigenetics as a cause of MS. Firstly, are environmental factors different enough in the Northern Isles to cause a change in epigenetic profiles between two population groups? Factors such as EBV infection, smoking and vitamin D have been shown not to differ between these populations. It may be possible that a difference in diet in general (given the higher rate of obesity in the Northern Isles) or an unknown or unmeasured environmental factor may cause a significant difference, but this has not been quantified.

Secondly, if a significant different in environmental factors is seen between populations, how much will this contribute to the variation of Multiple Sclerosis? The potential difference in epigenetics between mainland Scotland and the Northern Isles relies on not only there being a significant difference between key environmental factors, but for that difference to be great enough to have an influential effect on epigenetic changes that is enough to influence MS risk. Although it is known that environmental factors for MS can influence epigenetics, there has not been any specific quantification of how much variation in risk this results in.

I think that although diet or some unmeasured environment factor has the potential to cause a slight difference in epigenetics, I do not think that this will significantly impact the prevalence of MS in the Northern Isles. I certainly doubt that it is the reason for excess prevalence in the region. However, on a broader scale, it may be interesting to study the difference in epigenetic patterns between these and other populations, and how, if at all, this contributes to disease variation.

Rare variants

Possibly the most likely contributor towards the excess risk of MS in the Northern Isles are rare genetic variants: variants with a frequency of less than 0.01 that are usually not captured by SNP-array genotyping technology. Much evidence has been presented in recent years to support the influence of rare variants in MS risk, particularly amongst family groups. For example, whole exome sequencing on 127 individuals from 26 multiplex families identified 28 novel genes with rare variants that had segregated completely with the disease in at least two families (Haines et al., 2013). Another recent

study identified 9 rare variants within 6 genes that were all present within one multi-incident family (Traboulsee *et al.*, 2017). Even within a single family there may be more than one variant that is increasing MS risk. Another study identified a rare three-variant haplotype in purinergic receptor genes *P2RX4* and *P2RX7* within a multi-incident family (Sadovnick *et al.*, 2017); the rare variants were found to segregate in six family members diagnosed with MS. These three studies are examples of rare variants identified which likely cause MS, and there are other studies that have identified other rare variants associating with MS (Harding and Robertson, 2019). Within the Northern Isles, it is possible that there are one or more rare variants, possibly segregating within families, that are associated with Multiple Sclerosis. Rare variants of interest will have large effect sizes and are likely to be causal in MS, with obvious functional consequences. Therefore, these have the potential to explain a larger proportion of the excess cases of MS found in the islands.

Isolated populations such as the Northern Isles have generally experienced bottlenecks and genetic drift, and so by chance rare variants can increase in frequency. A 2017 study (which included Orkney) empirically showed that isolated populations were enriched for rare functional variants (Xue *et al.*, 2017). Enrichment of rare variants has been the source of high disease rates in other population isolates, for example high schizophrenia risk within a Finnish population isolate was attributed to rare variants (Pietiläinen *et al.*, no date). It is possible that the Northern Isles may have multiple rare variants and in combination they may explain a considerable proportion of variance for MS. Common variants contribute only a relatively small proportion to the overall heritability of MS, and so other types of variants, particularly rare variants, should be prioritised for future research. However, by their very nature rare variants are going to be extremely difficult to pin down in a population where there is a very low limit to the total number of patients that can ever be collected without waiting impractically long for cases to accrue, and hence limits on statistical power.

6.3 Strengths, limitations, implications and future work

6.3.1 Strengths of using ORCADES and VIKING

This thesis used the ORCADES and VIKING cohorts, which sample population isolates; with this comes both positive and negative characteristics for data use. Population isolates have the potential to reveal key insights into causative alleles. Alleles which may be at a rare or low frequency in the general population may have drifted to higher frequencies due to, for example, the founders of the populations having higher frequencies of these alleles by chance. This can allow their detection, which otherwise may have been missed. Population isolates can also have a unique profile of rare variants, which has the potential to reveal insights into disease pathways which would otherwise be undiscovered.

ORCADES and VIKING were created with the aim of discovering genes and variants that contribute to common, complex diseases. Although both datasets are relatively small, they sample a substantial proportion of the island populations and are the only genetic resources available for the unique Northern Isles populations, and so should be utilised fully. Both cohorts have rich phenotyping and genotyping and are an excellent resource for the population isolates of Orkney and Shetland in the Northern Isles of Scotland. Additionally, all participants have ancestry from either Orkney or Shetland and therefore have enhanced kinship information with recorded pedigrees. This can be exceedingly useful for data analysis as pairwise relationships can be informative within genetic analyses.

6.3.2 Limitations of using ORCADES and VIKING

Sample Size

Sample size is fundamental for every statistical study. An adequate sample size is required to detect an effect within a sampled population group; if a sample size is too low, it will lead to Type II errors. Conversely, if a sample size is too high, it is adding unnecessary additional costs to data collection. Generally, an effective sample size is one that is defined as the minimum number of samples needed to achieve adequate statistical power. Statistical power can be defined as the probability that a null hypothesis is rejected when the alternative hypothesis is true. An adequate statistical power of 80% is typically used; using power to this level avoids false negative associations while retaining cost effectiveness in data collection (Hong and Park, 2012).

Power can be affected by multiple factors, with different study designs resulting in different statistical power estimates. For example, dominant models require a lower sample size than additive models, common diseases and common SNPs with stronger effect sizes also require lower sample sizes, particularly when considering variants with high levels of LD. Increasing the number of controls per case until a ratio of 1:4 has been reached will also allow an increase in power. Hong and Park (2012) calculated that a GWAS testing a single SNP will require 248 cases, while testing 500,000 SNPs requires 1206 cases, under an assumption of 1:1 case/control ratio, complete LD, 5% MAF, 5% disease prevalence and an OR of 2 (Hong and Park, 2012).

Within this study, the principal limiting factor was sample size. Multiple sclerosis in particular has many variants of small to moderate effect (OR 1.02 - 1.2). Therefore, larger sample sizes will inherently be needed to achieve the power needed to detect these common variants of smaller effect. The inadequate sample size was particularly evident within this thesis in both the heritability analysis and GWAS. Case numbers were not a limiting factor within the key analysis in Chapter 5, as the principal comparison of polygenic risk scores to determine if there were a difference in common allele frequencies in mainland Scotland and the Northern Isles was only conducted on control individuals.

In Chapter 3, an accurate heritability estimate was unable to be obtained for VIKING (15 cases and 2090 controls), and the heritability estimate for ORCADES (97 cases and 2118 controls) had had a confidence interval spanning 0.1 to 0.5. With estimating SNP heritability, increasing sample size allows the estimate to approach the true SNP-heritability of the population. Compared to the IMSGC estimate (14,802 cases and 26,703 controls), whose confidence interval spanned 0.18-0.20, the effect of large sample size is apparent. Under the assumption of a trait heritability of 0.2 (on the liability scale), an additional 560 ORCADES cases and 595 VIKING cases would be needed to achieve an adequate statistical power in each SNP heritability analysis (Visscher *et al.*, 2014).

Within the GWAS, this study was unable to find any statistically significant SNPs with 112 cases and 4,208 controls for the Northern Isles. GWAS in particular requires a stringent p-value threshold. Within one association, the observed signal is statistically significant if the p-value is lower than the predetermined 0.05 required to reject the null

hypothesis. However, when testing a large number of SNPs, this p-value is corrected for multiple comparison to reduce false positive results. Therefore, a significantly larger sample size is needed to achieve enough power for discovery (Klein, 2007). Specifically for this study, 698 additional samples would be required to detect variants to genome-wide significance level, with the assumption of the disease allele frequency equal to 0.5 and a genotype relative risk of 1.5 (Menashe, Rosenberg and Chen, 2008).

The ORCADES and VIKING dataset are limited in what they can achieve in terms of variant discovery when using traditional methods such as GWAS. A relatively small population (approximately 22,000 individuals in Orkney and 23,000 individuals in Shetland) will yield small numbers of MS cases, despite the high rates of MS in the islands. This in turn results in reduced power and therefore the likelihood of discovery of disease contributors is low, particularly when comparing these studies to the huge cohort studies conducted by IMSGC.

Unfortunately, it is not possible to increase the case numbers within ORCADES and VIKING, and so these cohorts must be maximised with novel, and often relatively expensive, methods of analysis (such as WGS) to gain new knowledge of variants. Although the analyses conducted in this study were unlikely to prove fruitful given the low power, it was still important to conduct this research. Heritability, GWAS and PGRS analyses had not previously been performed for the Northern Isles, and so it was important to conduct these even given the small possibility that significant results would be found.

The use of common reference panels for imputation

Within this study, a GWAS was performed on an imputed dataset. In the absence of whole genome sequencing, imputation provides a cost-effective solution for enriching genotyped datasets. It is based on haplotype sharing between individuals and can be carried out using a framework of common variants from a SNP array and a collection of haplotypes in a reference panel. The reference panels most often used for imputation, common reference panels, are composed of a large number of individuals of European ancestry. However, several factors will influence the quality and success of imputation and the genomic coverage of variants, including haplotype structure, the presence of population-specific variants, and differing allele frequencies (Marchini and Howie,

2010). Thus, a one-size-fits-all approach to common reference panels may not be appropriate for unique populations, such as population isolates.

Within population isolates, common risk variants can generally be imputed with good accuracy. However, any rare or unique variants which are specific to the population would be difficult, if not impossible, to impute. To resolve this, a population-specific reference panel can be designed to produce higher quality imputation results that are more accurate and that provide more genomic coverage and give a better enrichment of variants. By capturing rare variation more effectively, downstream analyses would be more likely to identify these rare variants with success. Had a population-specific reference panel been available for the Northern Isles dataset, it is possible that the GWAS would identify variants to statistical significance or bring more truly causal variants to suggestive significance. Population isolates are an excellent resource for identifying unique and rare variants which contribute to a disease, and so a population-specific reference panel is a powerful strategy for achieving higher genomic coverage, higher imputation accuracy and a better estimation of allele frequencies.

Linkage disequilibrium

It is important to consider how much linkage disequilibrium exists within a population, and how this may influence genetic analyses. Longer stretches of LD means longer haplotypes, which helps imputation and facilitates GWAS in identifying an associated region of the genome with a trait (Hatzikotoulas, Gilly and Zeggini, 2014). However, having longer LD blocks adds extraneous markers to a possible causal variant, which may hinder efforts to isolate the causative variant from other variants that are in complete LD with them (Kristiansson, Naukkarinen and Peltonen, 2008). Therefore, longer LD blocks that may help to initially identify a locus also lead to a lesser resolution in defining causal variants within a wider associated area. This is important to discuss in relation to population isolates, as younger isolates tend to have LD that exists over slightly longer genomic regions. LD is influenced by many factors, including recombination, population size and population structure, and this leads to regions of LD in older populations tending to be shorter than LD regions in younger populations. Although this was not a particular issue within this study (as no statistically significant variants were identified through GWAS), it is an important point to note within population isolates that can cause issues with analyses further on from GWAS.

6.3.3 Implications of this research

The research in this thesis has several implications for future MS research. Firstly, I showed that common risk variants do not appear to cause the geographic hotspot. This may help researchers looking to find the cause of high rates of MS in other regions, with the suggestion that it may not be due to common variants (particularly because known common risk variants also appear to contribute less than 5% to disease variance, and also tend to be found at similar frequencies across populations). Additionally, several potential common variants which may contribute to MS risk have been highlighted. If these are later proven to be real associations, they have the potential to be informative of novel pathways of disease, which in turn could lead to new drug treatments.

However, one of the primary reasons of conducting the research in this study, aside from adding to the current knowledge on MS, was to pass the findings on to the Orkney and Shetland communities. There is much speculation within the public sphere in the Northern Isles regarding the cause of the excess of MS cases. It is therefore important for these individuals to know that they do not carry an unusual burden of common risk variants. These variants, although important in determining MS risk, are not placing an additional burden on the people of the Northern Isles in comparison to those in mainland Scotland. The strongest known variant, *HLA-DRB1*1501*, appears to be at a slightly higher frequency in Orkney and Shetland; individuals on the islands could potentially take an at-home genetic test to determine if they have the risk alleles for this variant, although this advice may be excessive given the small proportion it contributes to developing MS.

Public health information

Communicating the findings of past, current and future studies of MS research in an effective way to the public is essential to avoid misconceptions about MS cause and prevention. For example, in a previous study by Weiss et al, it was found that low blood plasma vitamin D levels were not causing an excess of MS cases in the region. This does not mean that vitamin D deficit is not important in influencing MS risk, but only that it is not causing additional cases of MS in the region than would be expected. It is

important for individuals of the public to know that it is still important to keep vitamin D intake at adequate levels to prevent increased risk of disease.

It is particularly important to convey to individuals who live in the Northern Isles that although previous studies indicate that environmental factors do not appear to contribute to the excess risk of MS on the islands, the influence on the environment on disease risk and development is still important. Many residents are aware of the high rates of MS, and caution should be made before publicising that environmental factors such as vitamin D are not contributory to the excess risk, as they still contribute to the general level of risk. Any healthcare on the island should not take emphasis away from the importance of diet and healthy lifestyle choices.

In general, it is hoped that the findings of this research are communicated to individuals in the Northern Isles with the intent of starting discussion on the risk factors of MS in general, with the goal of encouraging better lifestyle choices. Although the environmental risk factors are no more significant in causing MS than mainland Scotland, they should still be targeted for improvement. Emphasis should be made to reduce the rate of overweight and obese individuals on the islands, given the high percentage of individuals who are classed as overweight. This would not only lower the risk of MS for individuals in the island, but many other overweight-associated health conditions.

6.3.4 Future research

This study has found that common risk variants as a collective, do not make a meaningful contribution to the excess of MS in the Northern Isles. However, a small proportion of the excess risk is due to the strongest known common variant, *HLA-DRB1*1501*. Yet, the majority of the excess risk in the Northern Isles remains unexplained.

The findings of this research are limited in answering this question, however it is possible that the excess rates are caused by an unlucky combination of multiple environmental and genetic factors: a higher frequency of the *HLA-DRB1*1501* variant, strong rare variants which are confined to several family lines within the Northern Isles,

higher obesity levels and gene-environment interactions. There may also be unknown environmental factors which contribute to an increased risk, such as heavy metal toxicity or an infectious agent.

Given that there are most likely multiple contributors to the excess rate of MS, there are numerous avenues to which future research could be directed.

The most obvious of these appears to be a targeted search for rare variants. Rare variants are often enriched in population isolates due to founder effect and genetic drift and can have strong effects on disease risk. Recent MS studies have also provided further evidence that multiplex families often have one or more rare variants. Multiple Sclerosis is a heterogenous disease, and as such it is likely that individuals have causal variants of different frequencies contributing towards the disease. Identifying rare variants, even if they are unique to the Northern Isles, will contribute to the full spectrum of known variants on MS and provide a greater understanding to disease mechanisms. Additionally, if the discovered variants have a high effect on MS risk, measuring them in a clinical setting could lead to personalised medicine for individuals with the highest genetic risk.

It is suggested that rare variant research should use whole genome sequencing to scan multiplex families to potentially identify causal rare and low frequency variants. It is possible that there are private variants found only within specific families – if a variant is not observed in large collections of genomic data (e.g. gnomAD) and only exists within a family, the only way to detect it is through WGS. There have been substantial decreases in the cost of sequencing, as well as increases in the throughput, which have made it more favourable to carry out WGS at a larger scale. Additionally, an alternative to whole genome sequencing is whole exome sequencing, which only sequences coding regions of the genome. Although this method does not capture variation in non-coding regions, it is more cost effective. Using whole genome or exome sequencing within the Shetland and Orkney populations would be particularly useful as the pedigree information held for these populations would allow information gained to be extrapolated across the related individuals in the cohorts.

Specific recommendations

A primary focus of future research on multiple sclerosis in the Northern Isles should be

on rare variants. It is recommended that the most cost-effective solution for effectively detecting these variants is to whole-genome sequence individuals in ORCADES (primarily) and VIKING to a low depth. This would capture variants in non-coding regions but would have more power than sequencing less individuals at a higher depth (Li et al., 2011). This strategy has proven successful for improving power to detect rare variants (Holm et al., 2011). In the samples that have been sequenced, variants can be phased with long-range haplotype phasing and then imputed back into the sample set. This is similar to using the sequenced group as a population-specific reference panel for imputation. In terms of specific analyses to detect rare variants, it is suggested that a gene-based burden testing approach should be used (Guo et al., 2018). This takes sequence data to look at the number of individuals carrying rare variants in each gene and comparing them between MS cases and controls. By combining multiple rare variants across a candidate gene, it overcomes the power issue faced when considering rare variants alone. WGS would (in theory) be able to detect Orkney and Shetland specific rare variants, although power will always be limited by the low absolute numbers of cases, overall and in any particular kindred.

6.3.5 Summary

Multiple Sclerosis is a multifactorial disease of autoimmune origin which is increasingly common at higher latitudes, including Scotland. It has previously been shown that the Northern Isles of Scotland have the highest prevalence of MS in the world. Various risk factors, both genetic and environmental, are implicated in MS, but the reasons for the peak in Orkney and Shetland are not well understood. This thesis sought to better understand these very high rates through several approaches using the data of the Northern Isles Multiple Sclerosis study, Orkney Complex Disease Study and the Viking Health Study - Shetland.

As well as shining a light on the very high rates this thesis set out to contribute the findings to the general understanding of MS as well as in the communities of Orkney and Shetland. Firstly, the SNP heritability in Orkney for MS was estimated at 0.31 (95% CI 0.13, 0.49), while an estimate for Shetland could not be obtained due to low case numbers. Secondly, a GWAS identified 89 SNPs of suggestive significance, largely

within six key regions of the genome; four of the six regions were implicated in immune system functioning or have some previous link to an MS-related pathway and so may be useful for further study, while one was in the known HLA region.

Thirdly, a tag SNP for the most strongly associated MS risk variant, *HLA-DRB1*1501*, was found to have a significantly higher frequency in Orkney (0.23) and Shetland (0.21) than mainland Scotland (0.17). This SNP equates to causing 6 cases (95% CI 3, 8) out of 150 observed excess cases per 100,000 individuals in Shetland and 9 cases (95% CI 8, 11) of the observed 257 excess cases per 100,000 individuals in Orkney. This SNP therefore explains approximately 4% of the excess cases found in both populations.

Finally, common risk variants in general (not including the *HLA-DRB1*1501* tag SNP) were not found to differ between mainland Scotland and Orkney and Shetland. Overall, they explained 3% of the variance of MS risk in the Northern Isles; the *HLA-DRB1*1501* SNP accounted for 1% of this.

The results in this study have shown that common variants do not dominate the excess MS risk in Orkney and Shetland and have provided several suggestive candidate variants for follow up studies. A small part of the picture of MS in the Northern Isles has been painted here, but much remains to be elucidated about the disease. It is likely that variants of multiple frequencies likely contribute to the genetic architecture of MS, rather than one sole contributor.

It is known that environmental factors have a strong influence of MS, and here I have shown that genetics make at least a small contribution to the excess rates of MS in this region.

However, the cause of the excess MS prevalence in the Northern Isles remains mostly unexplained and is still a case for investigation. Based on previous findings as well as the findings in this thesis, I think that the excess prevalence of MS can likely be attributed to several factors. The biggest proportion of excess cases is likely caused by rare variants with strong effects. Smaller contributions will also be made from the high rates of overweight and obesity found in the islands: being obese raises risk of MS, but also has interactive effects with genetic variants, further increasing disease risk. A higher frequency of *HLA-DRB1*1501* is causing a small number of cases, but I think that there may be further gene-environment interactive effects from this variant that

contribute to MS risk. I also think it would be worth investigating heavy metal levels within the Northern Isles, as heavy metals have been associated with other MS hotspots.

6.3.6 Conclusion

Genetic studies of Multiple Sclerosis have progressed immensely over the past few years. MS has changed from an unknown and untreatable disease to a more manageable condition which we now understand with greater clarity. Genetic studies of MS indicate that the immune system is the primary source of MS pathology, with variants of numerous frequencies influencing disease risk along with multiple environmental factors. Identifying additional variants and determining their function and interactions with the environment will give us new information to leverage as novel targets for therapy, prevention and personalised healthcare.

A major goal for MS research is developing precision medicine for the disease with personalised treatments tailored towards sub-populations with unique causative profiles. This could prove particularly useful for communities such as the Northern Isles, who likely have a unique risk profile and would benefit greatly from personalised healthcare strategies.

Researching MS has the potential to bring to light new surprises, leading to a changing understanding of the disease with each new discovery made. The future of MS research in the Northern Isles and beyond will lie in new methods to discover and explore the function of susceptibility variants and their interactions with the world around us.

BIBLIOGRAPHY

Acheson, E. D., Bachrach, C. A. and Wright, F. M. (1960) 'Some comments on the relationship of the distribution of multiple sclerosis to latitude, solar radiation, and other variables', *Acta Psychiatrica Scandinavica*. Wiley/Blackwell (10.1111), 35(S147), pp. 132–147. doi: 10.1111/j.1600-0447.1960.tb08674.x.

Ahn, J. *et al.* (2010) 'Genome-wide association study of circulating vitamin D levels', *Human Molecular Genetics*, 19(13), pp. 2739–2745. doi: 10.1093/hmg/ddq155.

Alla, S. *et al.* (2014) 'The Increasing Prevalence of Multiple Sclerosis in New Zealand', *Neuroepidemiology*, 42(3), pp. 154–160. doi: 10.1159/000358174.

Alonso, A. and Hernan, M. A. (2008) 'Temporal trends in the incidence of multiple sclerosis: A systematic review', *Neurology*. Lippincott Williams & Wilkins, 71(2), pp. 129–135. doi: 10.1212/01.wnl.0000316802.35974.34.

Altmüller, J. *et al.* (2001) 'Genomewide scans of complex human diseases: true linkage is hard to find.', *American Journal of Human Genetics*, 69(5), pp. 936–50. doi: 10.1086/324069.

Altshuler, D., Daly, M. J. and Lander, E. S. (2008) 'Genetic mapping in human disease.', *Science*. NIH Public Access, 322(5903), pp. 881–8. doi: 10.1126/science.1156409.

Aly, M. *et al.* (2011) 'Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study.', *European Urology*. NIH Public Access, 60(1), pp. 21–8. doi: 10.1016/j.eururo.2011.01.017.

Amankwah, N. *et al.* (2017) 'Multiple sclerosis in Canada 2011 to 2031: results of a microsimulation modelling study of epidemiological and economic impacts.', *Health Promotion and Chronic Disease Prevention in Canada : Research, Policy and Practice*, 37(2), pp. 37–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28273039 (Accessed: 15 May 2018).

Antel, J. *et al.* (2016) 'NR1H3 p.Arg415Gln Is Not Associated to Multiple Sclerosis Risk', *Neuron*, 92(2), pp. 333–335. doi: 10.1016/j.neuron.2016.09.052.

Anttila, V. *et al.* (2017) 'Analysis of shared heritability in common disorders of the brain', *bioRxiv*. Cold Spring Harbor Laboratory, p. 048991. doi: 10.1101/048991.

Ascherio, A. and Munger, K. L. (2007) 'Environmental risk factors for multiple sclerosis. Part I: The role of infection', *Annals of Neurology*, 61(4), pp. 288–299. doi: 10.1002/ana.21117.

Aulchenko, Y. S. *et al.* (2007) 'GenABEL: an R library for genome-wide association analysis.', *Bioinformatics*. Oxford University Press, 23(10), pp. 1294–6. doi: 10.1093/bioinformatics/btm108.

Azzopardi, L., Coles, A. and Sklerozda Alemtuzumab, M. (2011) 'Alemtuzumab in Multiple Sclerosis', *Noro Psikiyatri Arsivi*, 48(2), pp. 79–82. doi: 10.4274/Npa.Y6426.

Bäärnhielm, M., Olsson, T. and Alfredsson, L. (2014) 'Fatty fish intake is associated with decreased occurrence of multiple sclerosis', *Multiple Sclerosis Journal*, 20(6), pp. 726–732. doi: 10.1177/1352458513509508.

Bahlo, M. *et al.* (2009) 'Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20', *Nature Genetics*, 41(7), pp. 824–828. doi: 10.1038/ng.396.

Band, G. and Marchini, J. (2018) *qctool*. Available at: http://www.well.ox.ac.uk/~gav/qctool/#overview (Accessed: 30 September 2017).

Baranzini, S. E. *et al.* (2010) 'Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis', *Nature*, 464(7293), pp. 1351–1356. doi: 10.1038/nature08990.

Baranzini, S. E. and Oksenberg, J. R. (2017) 'The Genetics of Multiple Sclerosis: From 0

to 200 in 50 Years', *Trends in Genetics*. Elsevier Current Trends, 33(12), pp. 960–970. doi: 10.1016/J.TIG.2017.09.004.

Barcellos, L. F. *et al.* (2002) 'Genetic basis for clinical expression in multiple sclerosis.', *Brain : A Journal of Neurology*, 125(Pt 1), pp. 150–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11834600 (Accessed: 30 April 2018).

Barnett, M. H. *et al.* (2016) 'Migration and multiple sclerosis in immigrants from United Kingdom and Ireland to Australia: a reassessment. III: risk of multiple sclerosis in UKI immigrants and Australian-born in Hobart, Tasmania', *Journal of Neurology*. Springer Berlin Heidelberg, 263(4), pp. 792–798. doi: 10.1007/s00415-016-8059-6.

Barrett, J. C. *et al.* (2008) 'Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.', *Nature Genetics*, 40(8), pp. 955–62. doi: 10.1038/ng.175.

Bazerman, M. H. and Samuelson, W. F. (1983) 'I Won the Auction but Don't Want the Prize', *The Journal of Conflict Resolution*. Sage Publications, Inc., pp. 618–634. doi: 10.2307/173888.

Beecham, A. H. *et al.* (2013) 'Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis.', *Nature Genetics*, 45(11), pp. 1353–60. doi: 10.1038/ng.2770.

Belbasis, L. *et al.* (2015) 'Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses', *Lancet Neurology*. Elsevier, 14(3), pp. 263–273. doi: 10.1016/S1474-4422(14)70267-4.

Benjaminsen, E. *et al.* (2014) 'Multiple sclerosis in the far north--incidence and prevalence in Nordland County, Norway, 1970-2010.', *BMC Neurology*. BioMed Central, 14, p. 226. doi: 10.1186/s12883-014-0226-8.

Bethesda (MD): National Center for Biotechnology Information, N. L. of M. (2005) *Database of Single Nucleotide Polymorphisms (dbSNP)*. National Center for Biotechnology Information (US). Available at: http://www.ncbi.nlm.nih.gov/SNP/ (Accessed: 11 March 2019).

Bielekova, B. *et al.* (2004) 'Humanized anti-CD25 (daclizumab) inhibits disease activity

in multiple sclerosis patients failing to respond to interferon', *Proceedings of the National Academy of Sciences*, 101(23), pp. 8705–8708. doi: 10.1073/pnas.0402653101.

Biran, I. and Steiner, I. (2004) 'Smoking is a risk factor for multiple sclerosis.', *Neurology*, 63(4), pp. 763; author reply 763. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15326276 (Accessed: 28 July 2017).

Björkegren, J. L. M. *et al.* (2015) 'Genome-Wide Significant Loci: How Important Are They?: Systems Genetics to Understand Heritability of Coronary Artery Disease and Other Common Complex Disorders', *Journal of the American College of Cardiology*. Elsevier, 65(8), pp. 830–845. doi: 10.1016/J.JACC.2014.12.033.

Bodmer, W. and Bonilla, C. (2008) 'Common and rare variants in multifactorial susceptibility to common diseases.', *Nature Genetics*. Europe PMC Funders, 40(6), pp. 695–701. doi: 10.1038/ng.f.136.

Booth, D. R. *et al.* (2009) 'The expanding genetic overlap between multiple sclerosis and type I diabetes', *Genes and Immunity*, 10(1), pp. 11–14. doi: 10.1038/gene.2008.83.

Bornancin, F. *et al.* (2015) 'Deficiency of MALT1 paracaspase activity results in unbalanced regulatory and effector T and B cell responses leading to multiorgan inflammation.', *Journal of Immunology*. American Association of Immunologists, 194(8), pp. 3723–34. doi: 10.4049/jimmunol.1402254.

Bove, R. *et al.* (2015) 'The 2D:4D ratio, a proxy for prenatal androgen levels, differs in men with and without MS.', *Neurology*. American Academy of Neurology, 85(14), pp. 1209–13. doi: 10.1212/WNL.0000000000001990.

Boyko, A. *et al.* (2016) 'Epidemiology of MS in Russia, a historical review', *Multiple Sclerosis and Demyelinating Disorders*. BioMed Central, 1(1), p. 13. doi: 10.1186/s40893-016-0016-9.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) 'An Expanded View of Complex Traits: From Polygenic to Omnigenic.', *Cell*. Elsevier, 169(7), pp. 1177–1186. doi: 10.1016/j.cell.2017.05.038.

Brey, R. L. (2003) 'Patient page. Cigarette smoking and multiple sclerosis (MS): yet

another reason to quit.', *Neurology*, 61(8), pp. E11-2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14581703 (Accessed: 27 July 2017).

Briggs, F. B. S. *et al.* (2014) 'Smoking and Risk of Multiple Sclerosis', *Epidemiology*, 25(4), pp. 605–614. doi: 10.1097/EDE.0000000000000089.

Bronner, C. E. *et al.* (1994) 'Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer', *Nature*, 368(6468), pp. 258–261. doi: 10.1038/368258a0.

Brønnum-Hansen, H., Koch-Henriksen, N. and Stenager, E. (2004) 'Trends in survival and cause of death in Danish patients with multiple sclerosis.', *Brain : A Journal of Neurology*, 127(Pt 4), pp. 844–50. doi: 10.1093/brain/awh104.

Browne, P. *et al.* (2014) 'Atlas of Multiple Sclerosis 2013: A growing global problem with widespread inequity.', *Neurology*. American Academy of Neurology, 83(11), pp. 1022–4. doi: 10.1212/WNL.0000000000000768.

Brüstle, A. *et al.* (2012) 'The NF-κB regulator MALT1 determines the encephalitogenic potential of Th17 cells.', *The Journal of Clinical Investigation*. American Society for Clinical Investigation, 122(12), pp. 4698–709. doi: 10.1172/JCI63528.

Bulik-Sullivan, B. *et al.* (2015) 'An atlas of genetic correlations across human diseases and traits', *Nature Genetics*, 47(11), pp. 1236–1241. doi: 10.1038/ng.3406.

Bulik-Sullivan, B. K. *et al.* (2015) 'LD Score regression distinguishes confounding from polygenicity in genome-wide association studies', *Nature Genetics*. NIH Public Access, 47(3), pp. 291–295. doi: 10.1038/ng.3211.

Caballero, A. *et al.* (1999) 'DQB1*0602 confers genetic susceptibility to multiple sclerosis in Afro-Brazilians.', *Tissue Antigens*, 54(5), pp. 524–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10599893 (Accessed: 28 July 2017).

Cabre, P. (2007) 'Migration and multiple sclerosis: The French West Indies experience', *Journal of the Neurological Sciences*, 262(1–2), pp. 117–121. doi: 10.1016/j.jns.2007.06.044.

Calabresi, P. A. (2004) 'Diagnosis and management of multiple sclerosis.', *American Family Physician*, 70(10), pp. 1935–44. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/15571060 (Accessed: 26 April 2018).

Cantorna, M. T. (2006) 'Vitamin D and its role in immunology: Multiple sclerosis, and inflammatory bowel disease', *Progress in Biophysics and Molecular Biology*, 92(1), pp. 60–64. doi: 10.1016/j.pbiomolbio.2006.02.020.

Capelli, C. *et al.* (2003) 'A Y chromosome census of the British Isles.', *Current biology : CB*, 13(11), pp. 979–84. doi: 10.1016/s0960-9822(03)00373-7.

Cardon, L. R. and Palmer, L. J. (2003) 'Population stratification and spurious allelic association.', *Lancet*, 361(9357), pp. 598–604. doi: 10.1016/S0140-6736(03)12520-2.

Carlborg, Ö. and Haley, C. S. (2004) 'Epistasis: too often neglected in complex trait studies?', *Nature Reviews Genetics*. Nature Publishing Group, 5(8), pp. 618–625. doi: 10.1038/nrg1407.

Cekanaviciute, E. *et al.* (2017) 'Gut bacteria from multiple sclerosis patients modulate human T cells and exacerbate symptoms in mouse models.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(40), pp. 10713–10718. doi: 10.1073/pnas.1711235114.

Chan, W. W. (1977) 'Eskimos and multiple sclerosis.', *Lancet*, 1(8026), p. 1370. Available at: http://www.ncbi.nlm.nih.gov/pubmed/69091 (Accessed: 15 May 2018).

Chanock, S. J. *et al.* (2007) 'Replicating genotype-phenotype associations.', *Nature*, 447(7145), pp. 655–60. doi: 10.1038/447655a.

Chao, M. *et al.* (2011) 'MHC transmission: Insights into gender bias in MS susceptibility', *Neurology*, 76(3), pp. 242–246. doi: 10.1212/WNL.0b013e318207b060.

Chu, F. *et al.* (2018) 'Gut Microbiota in Multiple Sclerosis and Experimental Autoimmune Encephalomyelitis: Current Applications and Future Perspectives', *Mediators of Inflammation*. Hindawi, 2018, pp. 1–17. doi: 10.1155/2018/8168717.

Cipolla, G. *et al.* (2018) 'Long Non-Coding RNAs in Multifactorial Diseases: Another Layer of Complexity', *Non-Coding RNA*, 4(2), p. 13. doi: 10.3390/ncrna4020013.

Clarke, L. *et al.* (2012) 'The 1000 Genomes Project: data management and community access', *Nature Methods*. Nature Research, 9(5), pp. 459–462. doi:

10.1038/nmeth.1974.

Claussnitzer, M. *et al.* (2015) 'FTO Obesity Variant Circuitry and Adipocyte Browning in Humans', *New England Journal of Medicine*. Massachusetts Medical Society, 373(10), pp. 895–907. doi: 10.1056/NEJMoa1502214.

Cocco, E. *et al.* (2011) 'Epidemiology of multiple sclerosis in south-western Sardinia', *Multiple Sclerosis Journal*. SAGE PublicationsSage UK: London, England, 17(11), pp. 1282–1289. doi: 10.1177/1352458511408754.

Coleman, C. *et al.* (2016) 'Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci', *European Journal of Human Genetics*. Nature Publishing Group, 24(2), pp. 291–297. doi: 10.1038/ejhg.2015.87.

Coles, A. J. *et al.* (1999) 'Monoclonal antibody treatment exposes three mechanisms underlying the clinical course of multiple sclerosis.', *Annals of Neurology*, 46(3), pp. 296–304. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10482259 (Accessed: 15 May 2018).

Comi, G. *et al.* (2012) 'Placebo-Controlled Trial of Oral Laquinimod for Multiple Sclerosis', *New England Journal of Medicine*. Massachusetts Medical Society , 366(11), pp. 1000–1009. doi: 10.1056/NEJMoa1104318.

Compston, A. (1981) 'Multiple Sclerosis in the Orkneys', *Lancet*, 318(8237), p. 98. doi: 10.1016/S0140-6736(81)90454-2.

Compston, A. and Coles, A. (2008) 'Multiple sclerosis', *Lancet*, 372(9648), pp. 1502–1517. doi: 10.1016/S0140-6736(08)61620-7.

Confavreux, C., Vukusic, S. and Adeleine, P. (2003) 'Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process.', *Brain : A Journal of Neurology*, 126(Pt 4), pp. 770–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12615637 (Accessed: 2 February 2016).

Corboy, J. R. and Miravalle, A. A. (2010) 'Emerging therapies for treatment of multiple sclerosis.', *Journal of Inflammation Research*. Dove Press, 3, pp. 53–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22096357 (Accessed: 16 May 2018).

Cordell, H. J. (2002) 'Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans', *Human Molecular Genetics*. Oxford University Press, 11(20), pp. 2463–2468. doi: 10.1093/hmg/11.20.2463.

Costelloe, L. *et al.* (2008) 'Long-term clinical relevance of criteria for designating multiple sclerosis as benign after 10 years of disease.', *Journal of Neurology, Neurosurgery and Psychiatry*, 79(11), pp. 1245–8. doi: 10.1136/jnnp.2008.143586.

Craddock, N. *et al.* (2010) 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', *Nature*, 464(7289), pp. 713–720. doi: 10.1038/nature08979.

Cruz-Orengo, L. *et al.* (2014) 'Enhanced sphingosine-1-phosphate receptor 2 expression underlies female CNS autoimmunity susceptibility', *The Journal of Clinical Investigation*. American Society for Clinical Investigation, 124(6), pp. 2571–2584. doi: 10.1172/JCI73408.

Curtis, C. and Simpson, G. (2001) *Summary of Research under DETR Contract 'Acidification of freshwaters: the role of nitrogen and the prospects for recovery', EPG1/3/117, ECRC Research Reports 79*. ECRC, University College London. Available at: http://discovery.ucl.ac.uk/34344/ (Accessed: 26 June 2019).

Davey Smith, G. and Ebrahim, S. (2005) 'What can mendelian randomisation tell us about modifiable behavioural and environmental exposures?', *British Medical Journal*. British Medical Journal Publishing Group, 330(7499), pp. 1076–9. doi: 10.1136/bmj.330.7499.1076.

Dean, G. (1967) 'Annual incidence, prevalence, and mortality of multiple sclerosis in white South-African-born and in white immigrants to South Africa.', *British Medical Journal*. BMJ Publishing Group, 2(5554), pp. 724–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6025978 (Accessed: 15 May 2018).

Debouverie, M. *et al.* (2008) 'Natural history of multiple sclerosis in a population-based cohort.', *European Journal of Neurology*, 15(9), pp. 916–21. doi: 10.1111/j.1468-1331.2008.02241.x.

Dendrou, C. A., Fugger, L. and Friese, M. A. (2015) 'Immunopathology of multiple sclerosis', *Nature Reviews Immunology*. Nature Research, 15(9), pp. 545–558. doi:

10.1038/nri3871.

Dilokthornsakul, P. *et al.* (2016) 'Multiple sclerosis prevalence in the United States commercially insured population', *Neurology*, 86(11), pp. 1014–1021. doi: 10.1212/WNL.0000000000002469.

Dobson, R. *et al.* (2016) 'A Risk Score for Predicting Multiple Sclerosis', *PLOS ONE*. Edited by O. Aktas, 11(11), p. e0164992. doi: 10.1371/journal.pone.0164992.

Dudbridge, F. (2013) 'Power and Predictive Accuracy of Polygenic Risk Scores', *PLOS Genetics*. Edited by N. R. Wray. Public Library of Science, 9(3), p. e1003348. doi: 10.1371/journal.pgen.1003348.

Dutta, R. and Trapp, B. D. (2011) 'Mechanisms of neuronal dysfunction and degeneration in multiple sclerosis.', *Progress in Neurobiology*. NIH Public Access, 93(1), pp. 1–12. doi: 10.1016/j.pneurobio.2010.09.005.

Dyment, D. A. *et al.* (2012) 'Exome sequencing identifies a novel multiple sclerosis susceptibility variant in the TYK2 gene', *Neurology*, 79(5), pp. 406–411. doi: 10.1212/WNL.0b013e3182616fc4.

Ebers, G. *et al.* (2004) 'Parent-of-origin effect in multiple sclerosis: observations in half-siblings', *Lancet*, 363(9423), pp. 1773–1774. doi: 10.1016/S0140-6736(04)16304-6.

Ebers, G. C. (2004) 'Natural history of primary progressive multiple sclerosis', *Multiple Sclerosis Journal*. SAGE PublicationsSage CA: Thousand Oaks, CA, 10(3_suppl), pp. S8–S15. doi: 10.1191/1352458504ms1025oa.

Edwards, S. L. *et al.* (2013) 'Beyond GWASs: illuminating the dark road from association to function.', *American Journal of Human Genetics*. Elsevier, 93(5), pp. 779–97. doi: 10.1016/j.ajhg.2013.10.012.

Efendi, H. (2015) 'Clinically Isolated Syndromes: Clinical Characteristics, Differential Diagnosis, and Management.', *Noro Psikiyatri Arsivi*. AVES, 52(Suppl 1), pp. S1–S11. doi: 10.5152/npa.2015.12608.

Ellinghaus, D. *et al.* (2016) 'Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci', *Nature Genetics*, 48(5), pp. 510–518. doi: 10.1038/ng.3528.

Ersoy, E. *et al.* (2005) 'The effects of interferon-beta on interleukin-10 in multiple sclerosis patients', *European Journal of Neurology*, 12(3), pp. 208–211. doi: 10.1111/j.1468-1331.2004.00986.x.

Evangelou, N. *et al.* (2000) 'Quantitative pathological evidence for axonal loss in normal appearing white matter in multiple sclerosis', *Annals of Neurology*. John Wiley & Sons, Inc., 47(3), pp. 391–395. doi: 10.1002/1531-8249(200003)47:3<391::AID-ANA20>3.0.CO;2-J.

Falconer, D. and Mackay, T. (1996) *Introduction to Quantitative Genetics*. Fourth Edi. Essex: Pearson.

Feng, S. *et al.* (2015) 'Methods for association analysis and meta-analysis of rare variants in families.', *Genetic Epidemiology*. NIH Public Access, 39(4), pp. 227–38. doi: 10.1002/gepi.21892.

Filippi, M. *et al.* (2003) 'Evidence for widespread axonal damage at the earliest clinical stage of multiple sclerosis.', *Brain : A Journal of Neurology*. Oxford University Press, 126(Pt 2), pp. 433–7. doi: 10.1093/brain/awg038.

Fletcher, J. M. *et al.* (2010) 'T cells in multiple sclerosis and experimental autoimmune encephalomyelitis', *Clinical & Experimental Immunology*, 162(1), pp. 1–11. doi: 10.1111/j.1365-2249.2010.04143.x.

Fogdell-Hahn, A. *et al.* (2000) 'Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease.', *Tissue Antigens*, 55(2), pp. 140–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10746785 (Accessed: 5 August 2016).

Frazer, K. A. *et al.* (2007) 'A second generation human haplotype map of over 3.1 million SNPs.', *Nature*, 449(7164), pp. 851–61. doi: 10.1038/nature06258.

Friedman, D. J. (2018) 'Genes and environment in chronic kidney disease hotspots', *Current Opinion in Nephrology and Hypertension*, p. 1. doi: 10.1097/MNH.0000000000000470.

Friese, M. A. and Fugger, L. (2005) 'Autoreactive CD8+ T cells in multiple sclerosis: a new target for therapy?', *Brain : A Journal of Neurology*. Oxford University Press,

128(Pt 8), pp. 1747–63. doi: 10.1093/brain/awh578.

Galarza-Muñoz, G. *et al.* (2017) 'Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk', *Cell*, 169(1), pp. 72-84.e13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28340352 (Accessed: 6 May 2018).

Gale, C. R. and Martyn, C. N. (1995) 'Migrant studies in multiple sclerosis.', *Progress in Neurobiology*, 47(4–5), pp. 425–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8966212 (Accessed: 3 August 2017).

Gao, X. *et al.* (2015) 'DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies', *Clinical Epigenetics*. BioMed Central, 7(1), p. 113. doi: 10.1186/s13148-015-0148-3.

Gholipour, T. *et al.* (2011) 'Demographic and clinical characteristics of malignant multiple sclerosis', *Neurology*, 76(23), pp. 1996–2001. doi: 10.1212/WNL.0b013e31821e559d.

Gianfrancesco, M. A. and Barcellos, L. F. (2016) 'Obesity and Multiple Sclerosis Susceptibility: A Review.', *Journal of Neurology & Neuromedicine*. NIH Public Access, 1(7), pp. 1–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27990499 (Accessed: 8 July 2019).

Gibson, G. (2009) 'Decanalization and the origin of complex disease.', *Nature Reviews Genetics*. Nature Publishing Group, 10(2), pp. 134–40. doi: 10.1038/nrg2502.

Gibson, G. (2010) 'Hints of hidden heritability in GWAS', *Nature Genetics*. Nature Publishing Group, 42(7), pp. 558–560. doi: 10.1038/ng0710-558.

Gibson, G. (2012) 'Rare and common variants: twenty arguments', *Nature Reviews Genetics*. Nature Publishing Group, 13(2), pp. 135–145. doi: 10.1038/nrg3118.

Gibson, G. (2018) 'Population genetics and GWAS: A primer', *PLOS Biology*. Public Library of Science, 16(3), p. e2005485. doi: 10.1371/journal.pbio.2005485.

Gibson, G. and Wagner, G. (2000) 'Canalization in evolutionary genetics: a stabilizing theory?', *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 22(4), pp. 372–80. doi: 10.1002/(SICI)1521-1878(200004)22:4<372::AID-BIES7>3.0.CO;2-J.

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995) 'Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models', *Biometrics*. International Biometric Society, 51(4), p. 1440. doi: 10.2307/2533274.

Giuliani, C. *et al.* (2016) 'Epigenetic Variability across Human Populations: A Focus on DNA Methylation Profiles of the KRTCAP3, MAD1L1 and BRSK2 Genes.', *Genome Biology and Evolution*. Oxford University Press, 8(9), pp. 2760–73. doi: 10.1093/gbe/evw186.

Goldman, D. (2014) 'The missing heritability of behavior: the search continues.', *Psychophysiology*. NIH Public Access, 51(12), pp. 1327–8. doi: 10.1111/psyp.12362.

Goodacre, S. *et al.* (2005) 'Genetic evidence for a family-based Scandinavian settlement of Shetland and Orkney during the Viking periods', *Heredity*. Nature Publishing Group, 95(2), pp. 129–135. doi: 10.1038/sj.hdy.6800661.

Goodin, D. S. (2009) 'The Causal Cascade to Multiple Sclerosis: A Model for MS Pathogenesis', *PLOS ONE*. Edited by E. Scalas. Public Library of Science, 4(2), p. e4565. doi: 10.1371/journal.pone.0004565.

Goris, A. *et al.* (2014) 'No evidence for shared genetic basis of common variants in multiple sclerosis and amyotrophic lateral sclerosis', *Human Molecular Genetics*, 23(7), pp. 1916–1922. doi: 10.1093/hmg/ddt574.

Gottesman, I. I. and Shields, J. (1967) 'A Polygenic Theory of Schizophrenia', 48(1), pp. 199–205. Available at: https://pdfs.semanticscholar.org/b881/3ecb7947cea17047df80d36ba04d35ec0dd3.pdf (Accessed: 29 August 2018).

Gross, K., Kokk, A. and Kaasik, A. E. (1993) 'Prevalence of MS in south Estonia. Evidence of a new border of the Fennoscandian focus.', *Acta Neurologica Scandinavica*, 88(4), pp. 241–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8256565 (Accessed: 24 June 2019).

Grossniklaus, U. *et al.* (2013) 'Transgenerational epigenetic inheritance: how important is it?', *Nature Reviews Genetics*, 14(3), pp. 228–235. doi: 10.1038/nrg3435.

Grytten, N., Torkildsen, Ø. and Myhr, K.-M. (2015) 'Time trends in the incidence and

prevalence of multiple sclerosis in Norway during eight decades', *Acta Neurologica Scandinavica*. John Wiley & Sons, Ltd (10.1111), 132, pp. 29–36. doi: 10.1111/ane.12428.

Grytten Torkildsen, N. *et al.* (2008) 'Survival and cause of death in multiple sclerosis: results from a 50-year follow-up in Western Norway.', *Multiple Sclerosis*, 14(9), pp. 1191–8. doi: 10.1177/1352458508093890.

Guo, M. H. *et al.* (2018) 'Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data.', *American journal of human genetics*. Elsevier, 103(4), pp. 522–534. doi: 10.1016/j.ajhg.2018.08.016.

Gusev, S. (2015) *GitHub - sashagusev/SKK-REML-sim: Evaluating GREML consistency with simulated genetic data*. Available at: https://github.com/sashagusev/SKK-REML-sim (Accessed: 15 November 2018).

Hafler, D. A. *et al.* (2007) 'Risk alleles for multiple sclerosis identified by a genomewide study.', *New England Journal of Medicine*, 357(9), pp. 851–62. doi: 10.1056/NEJMoa073493.

Hajian-Tilaki, K. (2013) 'Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.', *Caspian Journal of Internal Medicine*. Babol University of Medical Sciences, 4(2), pp. 627–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24009950 (Accessed: 29 August 2017).

Handel, A. E. and Ramagopalan, S. V. (2011) 'Smoking and Multiple Sclerosis: A Matter of Global Importance', *Neuroepidemiology*, 37(3–4), pp. 243–244. doi: 10.1159/000333241.

Handel, A. E. and Ramagopalan, S. V (2012) 'Vitamin D and multiple sclerosis: an interaction between genes and environment', *Multiple Sclerosis Journal*, 18(1), pp. 2–4. doi: 10.1177/1352458511418353.

Harding, K. E. and Robertson, N. P. (2019) 'New rare genetic variants in multiple sclerosis', *Journal of Neurology*. Springer Berlin Heidelberg, 266(1), pp. 278–280. doi: 10.1007/s00415-018-9128-9.

Hatzikotoulas, K., Gilly, A. and Zeggini, E. (2014) 'Using population isolates in genetic association studies.', *Briefings in Functional Genomics*. Oxford University Press, 13(5),

pp. 371–7. doi: 10.1093/bfgp/elu022.

Hauser, S. L. and Oksenberg, J. R. (2006) 'The Neurobiology of Multiple Sclerosis: Genes, Inflammation, and Neurodegeneration', *Neuron*, 52(1), pp. 61–76. doi: 10.1016/j.neuron.2006.09.011.

Hawkes, C. and Macgregor, A. (2009) 'Twin studies and the heritability of MS: a conclusion', *Multiple Sclerosis Journal*, 15(6), pp. 661–667. doi: 10.1177/1352458509104592.

Hedström, A. K. *et al.* (2011) 'Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis', *Brain*, 134(3), pp. 653–664. doi: 10.1093/brain/awq371.

Hedström, A. K. *et al.* (2014) 'Interaction between passive smoking and two HLA genes with regard to multiple sclerosis risk.', *International Journal of Epidemiology*. Oxford University Press, 43(6), pp. 1791–8. doi: 10.1093/ije/dyu195.

Hermann Eichhorst in Ziirich, V. (1896) 'Ueber infantile und hereditäre multiple Sklerose', *Vierzehnte Folge*, 146(2), pp. 172–192. Available at: https://link.springer.com/content/pdf/10.1007/BF01882681.pdf (Accessed: 2 May 2018).

Hernán, M. A. *et al.* (2005) 'Cigarette smoking and the progression of multiple sclerosis', *Brain*, 128(6), pp. 1461–1465. doi: 10.1093/brain/awh471.

Herrera, B. M. *et al.* (2008) 'Parent-of-origin effects in MS: Observations from avuncular pairs', *Neurology*, 71(11), pp. 799–803. doi: 10.1212/01.wnl.0000312377.50395.00.

Herzig, A. F. *et al.* (2018) 'Strategies for phasing and imputation in a population isolate', *Genetic Epidemiology*. Wiley-Blackwell, 42(2), pp. 201–213. doi: 10.1002/gepi.22109.

Hewer, S. *et al.* (2013) 'Vitamin D and multiple sclerosis', *Journal of Clinical Neuroscience*, 20(5), pp. 634–641. doi: 10.1016/j.jocn.2012.10.005.

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*. Public Library of Science, 4(2), p. e1000008. doi: 10.1371/journal.pgen.1000008.

Hirschhorn, J. N. *et al.* (2002) 'A comprehensive review of genetic association studies', *Genetics in Medicine*, 4(2), pp. 45–61. doi: 10.1097/00125817-200203000-00002.

Hirschhorn, J. N. and Daly, M. J. (2005) 'Genome-wide association studies for common diseases and complex traits.', *Nature Reviews Genetics*, 6(2), pp. 95–108. doi: 10.1038/nrg1521.

Høglund, R. A. and Maghazachi, A. A. (2014) 'Multiple sclerosis and the role of immune cells.', *World journal of Experimental Medicine*. Baishideng Publishing Group Inc, 4(3), pp. 27–37. doi: 10.5493/wjem.v4.i3.27.

Holmøy, T. and Hestvik, A. L. K. (2008) 'Multiple sclerosis: immunopathogenesis and controversies in defining the cause', *Current Opinion in Infectious Diseases*, 21(3), pp. 271–278. doi: 10.1097/QCO.0b013e3282f88b48.

Hong, E. P. and Park, J. W. (2012) 'Sample size and statistical power calculation in genetic association studies.', *Genomics & informatics*. Korea Genome Organization, 10(2), pp. 117–22. doi: 10.5808/GI.2012.10.2.117.

Hoppenbrouwers, I. A. *et al.* (2008) 'Maternal Transmission of Multiple Sclerosis in a Dutch Population', *Archives of Neurology*. American Medical Association, 65(3), pp. 120–124. doi: 10.1001/archneurol.2007.63.

Huebner, E. A. and Strittmatter, S. M. (2009) 'Axon Regeneration in the Peripheral and Central Nervous Systems', in *Results and Problems in Cell Differentiation*, pp. 305–360. doi: 10.1007/400_2009_19.

Huffman, J. E. (2018) 'Examining the current standards for genetic discovery and replication in the era of mega-biobanks', *Nature Communications*. Nature Publishing Group, 9(1), p. 5054. doi: 10.1038/s41467-018-07348-x.

Human Genome Sequencing Consortium, I. (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931–945. doi: 10.1038/nature03001.

Hunt, K. A. *et al.* (2008) 'Newly identified genetic risk variants for celiac disease related to the immune response', *Nature Genetics*, 40(4), pp. 395–402. doi: 10.1038/ng.102.

ImmunoBase (2019) *ImmunoBase*.

International Human Genome Sequencing Consortium (2001) 'Initial sequencing and analysis of the human genome', *Nature*. Nature Publishing Group, 409(6822), pp. 860–921. doi: 10.1038/35057062.

International Multiple Sclerosis Genetics Consortium (2010) 'Evidence for Polygenic Susceptibility to Multiple Sclerosis-The Shape of Things to Come', *American Journal of Human Genetics*. Elsevier, 86(4), pp. 621–625. doi: 10.1016/j.ajhg.2010.02.027.

International Multiple Sclerosis Genetics Consortium (2011) 'Genome-wide association study of severity in multiple sclerosis', *Genes & Immunity*, 12(8), pp. 615–625. doi: 10.1038/gene.2011.34.

International Multiple Sclerosis Genetics Consortium *et al.* (2017) 'The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility', *bioRxiv*. Cold Spring Harbor Laboratory, p. 143933. doi: 10.1101/143933.

International Multiple Sclerosis Genetics Consortium *et al.* (2018) 'Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk.', *Cell*. Elsevier, 175(6), pp. 1679-1687.e7. doi: 10.1016/j.cell.2018.09.049.

International Multiple Sclerosis Genetics Consortium (2018) 'Low frequency and rare coding variation contributes to multiple sclerosis risk', *Cell*, 175(6), pp. 1679–1687.e7. doi: 10.1101/286617.

Isobe, N. *et al.* (2013) 'Genetic risk variants in African Americans with multiple sclerosis', *Neurology*, 81(3), pp. 219–227. doi: 10.1212/WNL.0b013e31829bfe2f.

Isobe, N. *et al.* (2015) 'An ImmunoChip study of multiple sclerosis risk in African Americans', *Brain*. doi: 10.1093/brain/awv078.

Ivashkiv, L. B. and Donlin, L. T. (2014) 'Regulation of type I interferon responses.', *Nature Reviews Immunology*. NIH Public Access, 14(1), pp. 36–49. doi: 10.1038/nri3581.

Jafari, N. *et al.* (2011) 'Perspectives on the use of multiple sclerosis risk genes for prediction.', *PLOS ONE*. Public Library of Science, 6(12), p. e26493. doi: 10.1371/journal.pone.0026493.

Jagannath, V. A. *et al.* (2010) 'Vitamin D for the management of multiple sclerosis',

Cochrane Database of Systematic Reviews, (12), p. CD008422. doi: 10.1002/14651858.CD008422.pub2.

De Jager, P. L., Chibnik, L. B., *et al.* (2009) 'Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score', *Lancet Neurology*, 8(12), pp. 1111–1119. doi: 10.1016/S1474-4422(09)70275-3.

De Jager, P. L., Jia, X., *et al.* (2009) 'Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci', *Nature Genetics*. Nature Publishing Group, 41(7), pp. 776–782. doi: 10.1038/ng.401.

Jakkula, E. *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*. Elsevier, 86(2), pp. 285–91. doi: 10.1016/j.ajhg.2010.01.017.

Jaworski, M. *et al.* (2014) 'Malt1 protease inactivation efficiently dampens immune responses but causes spontaneous autoimmunity', *The EMBO Journal*. EMBO Press, 33(23), pp. 2765–2781. doi: 10.15252/embj.201488987.

Jersild C, Svejgaard A, F. T. (1972) 'HL-A antigens and multiple sclerosis.', *Lancet*, 3(7762), pp. 1240–1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4113225.

Johnson, B. A. *et al.* (2010) 'Multiple sclerosis susceptibility alleles in African Americans.', *Genes and immunity*. NIH Public Access, 11(4), pp. 343–50. doi: 10.1038/gene.2009.81.

Jorde, L. B. *et al.* (2000) 'Gene mapping in isolated populations: new roles for old friends?', *Human Heredity*, 50(1), pp. 57–65. doi: 22891.

Juilland, M. and Thome, M. (2018) 'Holding All the CARDs: How MALT1 Controls CARMA/CARD-Dependent Signaling', *Frontiers in Immunology*. Frontiers, 9, p. 1927. doi: 10.3389/fimmu.2018.01927.

Kantarci, O. and Wingerchuk, D. (2006) 'Epidemiology and natural history of multiple sclerosis: new insights.', *Current Opinion in Neurology*, 19(3), pp. 248–54. doi: 10.1097/01.wco.0000227033.47458.82.

Kappos, L. *et al.* (2010) 'A Placebo-Controlled Trial of Oral Fingolimod in Relapsing Multiple Sclerosis', *New England Journal of Medicine*. Massachusetts Medical Society ,

362(5), pp. 387–401. doi: 10.1056/NEJMoa0909494.

Kappos, L. *et al.* (2015) 'Long-term effects of fingolimod in multiple sclerosis: The randomized FREEDOMS extension trial', *Neurology*, 84(15), pp. 1582–1591. doi: 10.1212/WNL.0000000000001462.

Kennedy, R. B. *et al.* (2012) 'Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients', *Human Genetics*, 131(9), pp. 1403–1421. doi: 10.1007/s00439-012-1174-2.

Khankhanian, P. *et al.* (2015) 'Genetic contribution to multiple sclerosis risk among Ashkenazi Jews', *BMC Medical Genetics*. BioMed Central, 16(1), p. 55. doi: 10.1186/s12881-015-0201-2.

Khoury, M. J. *et al.* (2006) 'On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies', *International Journal of Epidemiology*, 36(2), pp. 439–445. doi: 10.1093/ije/dyl253.

Kim, J. I. *et al.* (2018) 'The effects of GRIN2B and DRD4 gene variants on local functional connectivity in attention-deficit/hyperactivity disorder', *Brain Imaging and Behavior*, 12(1), pp. 247–257. doi: 10.1007/s11682-017-9690-2.

Kimiskidis, V. *et al.* (2008) 'Autologous stem-cell transplantation in malignant multiple sclerosis: a case with a favorable long-term outcome.', *Multiple Sclerosis*. SAGE Publications, 14(2), pp. 278–83. doi: 10.1177/1352458507082604.

Kingwell, E. *et al.* (2013) 'Incidence and prevalence of multiple sclerosis in Europe: a systematic review', *BMC Neurology*, 13(1), p. 128. doi: 10.1186/1471-2377-13-128.

Kinlen, L. J. and Balkwill, A. (2001) 'Infective cause of childhood leukaemia and wartime population mixing in Orkney and Shetland, UK.', *Lancet*. Elsevier, 357(9259), p. 858. doi: 10.1016/s0140-6736(00)04208-2.

Kipp, M. *et al.* (2017) 'Multiple sclerosis animal models: a clinical and histopathological perspective', *Brain Pathology*. Wiley/Blackwell (10.1111), 27(2), pp. 123–137. doi: 10.1111/bpa.12454.

Kirby, T. O. and Ochoa-Repáraz, J. (2018) 'The Gut Microbiome in Multiple Sclerosis: A Potential Therapeutic Avenue.', *Medical Sciences*. Multidisciplinary Digital Publishing

Institute (MDPI), 6(3). doi: 10.3390/medsci6030069.

Kittles, R. A. *et al.* (1998) 'Dual Origins of Finns Revealed by Y Chromosome Haplotype Variation', *American Journal of Human Genetics*, 62(5), pp. 1171–1179. doi: 10.1086/301831.

Klareskog, L., Catrina, A. I. and Paget, S. (2009) 'Rheumatoid arthritis', *Lancet*, 373(9664), pp. 659–672. doi: 10.1016/S0140-6736(09)60008-8.

Klein, R. J. (2007) 'Power analysis for genome-wide association studies', *BMC Genetics*. BioMed Central, 8(1), p. 58. doi: 10.1186/1471-2156-8-58.

Koch, J. *et al.* (2003) 'cTAGE: A Cutaneous T Cell Lymphoma Associated Antigen Family with Tumor-Specific Splicing', *Journal of Investigative Dermatology*, 121(1), pp. 198–206. doi: 10.1046/j.1523-1747.2003.12318.x.

Koch, M. *et al.* (2008) 'Factors associated with the risk of secondary progression in multiple sclerosis.', *Multiple Sclerosis*, 14(6), pp. 799–803. doi: 10.1177/1352458508089361.

Krapohl, E. *et al.* (2018) 'Multi-polygenic score approach to trait prediction', *Molecular Psychiatry*. Nature Publishing Group, 23(5), pp. 1368–1374. doi: 10.1038/mp.2017.163.

Krishna Kumar, S. *et al.* (2016) 'Limitations of GCTA as a solution to the missing heritability problem.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(1), pp. E61-70. doi: 10.1073/pnas.1520109113.

Kurtzke, J. F., Dean, G. and Botha, D. P. (1970) 'A method for estimating the age at immigration of white immigrants to South Africa, with an example of its importance.', *South African Medical Journal*, 44(23), pp. 663–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5427147 (Accessed: 3 August 2017).

Kurtzke, J. F., Delasnerie-Lauprêtre, N. and Wallin, M. T. (1998) 'Multiple sclerosis in North African migrants to France.', *Acta Neurologica Scandinavica*, 98(5), pp. 302–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9858098 (Accessed: 3 August 2017).

Kyle, J. (1992) 'Crohn's Disease in the Northeastern and Northern Isles of Scotland: An

Epidemiological Review', *Gastroenterology*. Available at: https://www.gastrojournal.org/article/0016-5085(92)90826-K/pdf (Accessed: 9 July 2019).

Lander, E. S. and Schork, N. J. (1994) 'Genetic dissection of complex traits.', *Science*, 265(5181), pp. 2037–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8091226 (Accessed: 25 May 2015).

Lee, S. H. *et al.* (2011) 'Estimating missing heritability for disease from genome-wide association studies.', *American Journal of Human Genetics*. Elsevier, 88(3), pp. 294–305. doi: 10.1016/j.ajhg.2011.02.002.

Levine, J. M., Reynolds, R. and Fawcett, J. W. (2001) 'The oligodendrocyte precursor cell in health and disease', *Trends in Neurosciences*, 24(1), pp. 39–47. doi: 10.1016/S0166-2236(00)01691-X.

Li, Y. *et al.* (2011) 'Low-coverage sequencing: Implications for design of complex trait association studies', *Genome Research*, 21(6), pp. 940–951. doi: 10.1101/gr.117259.110.

Liao, W., Lin, J.-X. and Leonard, W. J. (2011) 'IL-2 family cytokines: new insights into the complex roles of IL-2 as a broad regulator of T helper cell differentiation', *Current Opinion in Immunology*, 23(5), pp. 598–604. doi: 10.1016/j.coi.2011.08.003.

Lill, C. M. (2014) 'Recent advances and future challenges in the genetics of multiple sclerosis.', *Frontiers in Neurology*. Frontiers Media SA, 5, p. 130. doi: 10.3389/fneur.2014.00130.

Lincoln, M. R. *et al.* (2005) 'A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis', *Nature Genetics*. Nature Publishing Group, 37(10), pp. 1108–1112. doi: 10.1038/ng1647.

Lincoln, M. R. *et al.* (2009) 'Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility', *Proceedings of the National Academy of Sciences*, 106(18), pp. 7542–7547. doi: 10.1073/pnas.0812664106.

Long, A. D. and Langley, C. H. (1999) 'The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits.', *Genome Research*, 9(8), pp. 720–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10447507

(Accessed: 27 September 2018).

de los Campos, G. *et al.* (2015) 'Genomic Heritability: What Is It?', *PLOS Genetics*. Edited by G. S. Barsh. Public Library of Science, 11(5), p. e1005048. doi: 10.1371/journal.pgen.1005048.

Lublin, F. D. *et al.* (2014) 'Defining the clinical course of multiple sclerosis: the 2013 revisions.', *Neurology*, 83(3), pp. 278–86. doi: 10.1212/WNL.0000000000000560.

Lublin, F. D. and Reingold, S. C. (1996) 'Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis.', *Neurology*, 46(4), pp. 907–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8780061 (Accessed: 5 December 2014).

Lucas, R. M. *et al.* (2015) 'Ultraviolet radiation, vitamin D and multiple sclerosis', *Neurodegenerative Disease Management*, 5(5), pp. 413–424. doi: 10.2217/nmt.15.33.

Lumeng, C. N., Bodzin, J. L. and Saltiel, A. R. (2007) 'Obesity induces a phenotypic switch in adipose tissue macrophage polarization', *Journal of Clinical Investigation*, 117(1), pp. 175–184. doi: 10.1172/JCI29881.

Macgregor, S. *et al.* (2006) 'Bias, precision and heritability of self-reported and clinically measured height in Australian twins', *Human Genetics*, 120(4), pp. 571–580. doi: 10.1007/s00439-006-0240-z.

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(21), pp. 3555–7. doi: 10.1093/bioinformatics/btv402.

Magro Checa, C. *et al.* (2013) 'Demyelinating disease in SLE: Is it multiple sclerosis or lupus?', *Best Practice & Research Clinical Rheumatology*. Baillière Tindall, 27(3), pp. 405–424. doi: 10.1016/J.BERH.2013.07.010.

Mak, T. *et al.* (2017) 'Polygenic scores via penalized regression on summary statistics', *bioRxiv*. Available at: http://www.biorxiv.org/content/early/2017/03/22/058214 (Accessed: 29 August 2017).

Makhani, N. *et al.* (2016) 'Viral exposures and MS outcome in a prospective cohort of children with acquired demyelination', *Multiple Sclerosis Journal*, 22(3), pp. 385–388. doi: 10.1177/1352458515595876.

Mallucci, G. *et al.* (2015) 'The role of immune cells, glia and neurons in white and gray matter pathology in multiple sclerosis.', *Progress in Neurobiology*, 127–128, pp. 1–22. doi: 10.1016/j.pneurobio.2015.02.003.

Maltby, V. E. *et al.* (2015) 'Genome-wide DNA methylation profiling of CD8+ T cells shows a distinct epigenetic signature to CD4+ T cells in multiple sclerosis patients.', *Clinical Epigenetics*. BioMed Central, 7, p. 118. doi: 10.1186/s13148-015-0152-7.

Mangalam, A., Rodriguez, M. and David, C. (1994) 'Role of MHC class II expressing CD4 + T cells in proteolipid protein 91–110 -induced EAE in HLA-DR3 transgenic mice', *Journal of Pathology, Microbiology and Immunology*, 102(1–6), pp. 241–248. doi: 10.1002/eji.200636217.

Manichaikul, A. *et al.* (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*. Oxford University Press, 26(22). doi: 10.1093/bioinformatics/btq559.

Manolio, T. a *et al.* (2009) 'Finding the missing heritability of complex diseases.', *Nature*. Nature Publishing Group, 461(7265), pp. 747–753. doi: 10.1038/nature08494.

Manousaki, D. *et al.* (2017) 'Low-Frequency Synonymous Coding Variation in CYP2R1 Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis.', *American Journal of Human Genetics*. Elsevier, 101(2), pp. 227–238. doi: 10.1016/j.ajhg.2017.06.014.

Marigorta, U. M. and Navarro, A. (2013) 'High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants', *PLOS Genetics*. Edited by S. M. Williams. Public Library of Science, 9(6), p. e1003566. doi: 10.1371/journal.pgen.1003566.

Marrie, R. A., Hall, N. and Sadovnick, A. D. (2016) 'Multiple sclerosis in First Nations Canadians: A pilot comparison study.', *Multiple Sclerosis Journal - Experimental, Translational and Clinical*. SAGE Publications, 2, p. 2055217316666093. doi: 10.1177/2055217316666093.

Marrosu, M. G. *et al.* (1997) 'Multiple Sclerosis in Sardinia Is Associated and in Linkage Disequilibrium with HLA-DR3 and -DR4 Alleles', *American Journal of Human Genetics*, 61(2), pp. 454–457. doi: 10.1016/S0002-9297(07)64074-9.

Martin, R. (2008) 'HLA class I: friend and foe of multiple sclerosis', *Nature Medicine*. Nature Publishing Group, 14(11), pp. 1150–1151. doi: 10.1038/nm1108-1150.

Martinelli, V. *et al.* (2014) 'Vitamin D levels and risk of multiple sclerosis in patients with clinically isolated syndromes', *Multiple Sclerosis Journal*, 20(2), pp. 147–155. doi: 10.1177/1352458513494959.

Maver, A. *et al.* (2017) 'Identification of rare genetic variation of NLRP1 gene in familial multiple sclerosis.', *Scientific Reports*. Nature Publishing Group, 7(1), p. 3715. doi: 10.1038/s41598-017-03536-9.

Mayama, T., Marr, A. and Kino, T. (2016) 'Differential Expression of Glucocorticoid Receptor Noncoding RNA Repressor Gas5 in Autoimmune and Inflammatory Diseases', *Hormone and Metabolic Research*, 48(08), pp. 550–557. doi: 10.1055/s-0042-106898.

Mayhew, A. J. and Meyre, D. (2017) 'Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities.', *Current Genomics*. Bentham Science Publishers, 18(4), pp. 332–340. doi: 10.2174/1389202918666170307161450.

Mc Guire, C. *et al.* (2013) 'Paracaspase MALT1 deficiency protects mice from autoimmune-mediated demyelination.', *Journal of Immunology*. American Association of Immunologists, 190(6), pp. 2896–903. doi: 10.4049/jimmunol.1201351.

McCarthy, S. *et al.* (2016) 'A reference panel of 64,976 haplotypes for genotype imputation.', *Nature Genetics*, 48(10), pp. 1279–83. doi: 10.1038/ng.3643.

McDonald, W. I. *et al.* (2001) 'Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis.', *Annals of Neurology*, 50(1), pp. 121–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11456302 (Accessed: 5 August 2016).

McDonald, W. I., Miller, D. H. and Thompson, A. J. (1994) 'Are magnetic resonance findings predictive of clinical outcome in therapeutic trials in multiple sclerosis? The dilemma of interferon-?', *Annals of Neurology*. Wiley Subscription Services, Inc., A

Wiley Company, 36(1), pp. 14–18. doi: 10.1002/ana.410360106.

McLeod, J. G., Hammond, S. R. and Hallpike, J. F. (1994) 'Epidemiology of multiple sclerosis in Australia. With NSW and SA survey results.', *The Medical Journal of Australia*, 160(3), pp. 117–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8295576 (Accessed: 15 May 2018).

McQuillan, R. *et al.* (2008) 'Runs of Homozygosity in European Populations', *The American Journal of Human Genetics*, 83(3), pp. 359–372. doi: 10.1016/j.ajhg.2008.08.007.

McWhirter, R. E. *et al.* (2012) 'Genome-wide homozygosity and multiple sclerosis in Orkney and Shetland Islanders', *European Journal of Human Genetics*. Nature Publishing Group, 20(2), pp. 198–202. doi: 10.1038/ejhg.2011.170.

Menashe, I., Rosenberg, P. S. and Chen, B. E. (2008) 'PGA: power calculator for case-control genetic association analyses', *BMC Genetics*. BioMed Central, 9(1), p. 36. doi: 10.1186/1471-2156-9-36.

Mescheriakova, J. Y. *et al.* (2016) 'Burden of genetic risk variants in multiple sclerosis families in the Netherlands.', *Multiple Sclerosis Journal - Experimental, Translational and Clinical*. SAGE Publications, 2, p. 2055217316648721. doi: 10.1177/2055217316648721.

Meyer, K. and Tier, B. (2012) '"SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model.', *Genetics*. Genetics Society of America, 190(1), pp. 275–7. doi: 10.1534/genetics.111.134841.

Miller, D. H. *et al.* (2008) 'Differential diagnosis of suspected multiple sclerosis: a consensus approach.', *Multiple Sclerosis*. SAGE Publications, 14(9), pp. 1157–74. doi: 10.1177/1352458508096878.

Minikel, E. V. and MacArthur, D. G. (2016) 'Publicly Available Data Provide Evidence against NR1H3 R415Q Causing Multiple Sclerosis', *Neuron*, 92(2), pp. 336–338. doi: 10.1016/j.neuron.2016.09.054.

Mirzaei, F. *et al.* (2011) 'Gestational vitamin D and the risk of multiple sclerosis in offspring', *Annals of Neurology*, 70(1), pp. 30–40. doi: 10.1002/ana.22456.

Moayyeri, A. *et al.* (2013) 'The UK Adult Twin Registry (TwinsUK Resource).', *Twin Research and Human Genetics*. Europe PMC Funders, 16(1), pp. 144–9. doi: 10.1017/thg.2012.89.

Mokry, L. E. *et al.* (2016) 'Obesity and Multiple Sclerosis: A Mendelian Randomization Study', *PLOS Medicine*. Edited by P. A. Muraro. Public Library of Science, 13(6), p. e1002053. doi: 10.1371/journal.pmed.1002053.

Mokry, L. E. LE *et al.* (2015) 'Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study', *PLOS Medicine*. Edited by P. A. Muraro. Public Library of Science, 12(8), p. e1001866. doi: 10.1371/journal.pmed.1001866.

Monti, M. C. *et al.* (2016) 'Is Geo-Environmental Exposure a Risk Factor for Multiple Sclerosis? A Population-Based Cross-Sectional Study in South-Western Sardinia', *PLOS ONE*. Edited by S. V. Ramagopalan, 11(9), p. e0163313. doi: 10.1371/journal.pone.0163313.

Morgan, T. M. *et al.* (2007) 'Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study.', *JAMA: The Journal of the American Medical Association*. American Medical Association, 297(14), pp. 1551–61. doi: 10.1001/jama.297.14.1551.

Morris, A. P. (2011) 'Transethnic meta-analysis of genomewide association studies', *Genetic Epidemiology*, 35(8), pp. 809–822. doi: 10.1002/gepi.20630.

Mosmann, T. R. and Sad, S. (1996) 'The expanding universe of T-cell subsets: Th1, Th2 and more', *Immunology Today*. Elsevier Current Trends, 17(3), pp. 138–146. doi: 10.1016/0167-5699(96)80606-2.

Motsinger, A. A. *et al.* (2007) 'Complex gene–gene interactions in multiple sclerosis: a multifactorial approach reveals associations with inflammatory genes', *Neurogenetics*. Springer-Verlag, 8(1), pp. 11–20. doi: 10.1007/s10048-006-0058-9.

Moutsianas, L. *et al.* (2015) 'Class II HLA interactions modulate genetic risk for multiple sclerosis.', *Nature Genetics*. NIH Public Access, 47(10), pp. 1107–13. doi: 10.1038/ng.3395.

Mowry, E. M. and Glenn, J. D. (2018) 'The Dynamics of the Gut Microbiome in Multiple

Sclerosis in Relation to Disease', *Neurologic Clinics*, 36(1), pp. 185–196. doi: 10.1016/j.ncl.2017.08.008.

Mumford, C. J. *et al.* (1994) 'The British Isles survey of multiple sclerosis in twins.', *Neurology*, 44(1), pp. 11–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8290043 (Accessed: 16 May 2018).

Muraro, P. A. *et al.* (2017) 'Long-term Outcomes After Autologous Hematopoietic Stem Cell Transplantation for Multiple Sclerosis', *JAMA Neurology*, 74(4), p. 459. doi: 10.1001/jamaneurol.2016.5867.

Nagelkerke, N. J. D. (1991) 'A Note on a General Definition of the Coefficient of Determination', *Biometrika*, 78(3), pp. 691–692. Available at: http://links.jstor.org/sici?sici=0006-3444%28199109%2978%3A3%3C691%3AANOAGD%3E2.0.CO%3B2-V (Accessed: 13 July 2017).

Napier, M. D. *et al.* (2016) 'Heavy metals, organic solvents, and multiple sclerosis: An exploratory look at gene-environment interactions', *Archives of Environmental & Occupational Health*, 71(1), pp. 26–34. doi: 10.1080/19338244.2014.937381.

National Center for Biotechnology Information (2017) *NHLBI TOPMed: Phase III variation data*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/400167 (Accessed: 13 June 2019).

National Records of Scotland (2018) *Local Area Migration*, *National Records of Scotland*. National Records of Scotland. Available at: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/migration/migration-statistics/local-area-migration (Accessed: 4 August 2017).

National Records of Scotland (2019) *Population Estimates Time Series Data*, *National Records of Scotland*. National Records of Scotland. Available at: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/population-estimates-time-series-data (Accessed: 24 June 2019).

National Records of Scotland, W. T. (no date) 'National Records of Scotland', *National Records of Scotland*. National Records of Scotland. Available at:

https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/migration/migration-statistics/local-area-migration (Accessed: 4 August 2017).

Neale, B., Neale and Ben (2014) 'Liability Threshold Models', in *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9781118445112.stat06439.

Nelson, M. R. *et al.* (2015) 'The support of human genetic evidence for approved drug indications', *Nature Genetics*, 47(8), pp. 856–860. doi: 10.1038/ng.3314.

NHS Scotland (2016) *Scottish Diabetes Survey 2016: Scottish Diabetes Survey Monitoring Group*. Available at: https://www.diabetes.org.uk/resources-s3/2017-09/Scottish Diabetes Survey 2016.pdf?_ga=2.231129370.468428599.1505127410-1295258485.1505127410 (Accessed: 9 July 2019).

NICE (2014) *Alemtuzumab for treating relapsing-remitting multiple sclerosis*. NICE. Available at: https://www.nice.org.uk/guidance/ta312 (Accessed: 15 May 2018).

Nielsen, N. *et al.* (2008) 'Autoimmune diseases in patients with multiple sclerosis and their first-degree relatives: a nationwide cohort study in Denmark', *Multiple Sclerosis Journal*, 14(6), pp. 823–829. doi: 10.1177/1352458508088936.

O'Gorman, C. *et al.* (2013) 'Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data', *Neuroepidemiology*, 40(1), pp. 1–12. doi: 10.1159/000341902.

Odoardi, F. *et al.* (2012) 'T cells become licensed in the lung to enter the central nervous system', *Nature*, 488(7413), pp. 675–679. doi: 10.1038/nature11337.

Oksenberg, J. R. and Barcellos, L. F. (2000) 'The complex genetic aetiology of multiple sclerosis.', *Journal of NeuroVirology*, 6 Suppl 2, pp. S10-4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10871777 (Accessed: 27 July 2017).

Olsson, T., Barcellos, L. F. and Alfredsson, L. (2017) 'Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis', *Nature Reviews Neurology*. Nature Publishing Group, 13(1), pp. 25–36. doi: 10.1038/nrneurol.2016.187.

Orton, S.-M. *et al.* (2006) 'Sex ratio of multiple sclerosis in Canada: a longitudinal study', *Lancet Neurology*, 5(11), pp. 932–936. doi: 10.1016/S1474-4422(06)70581-6.

Palmer, A. J. *et al.* (2013) 'A novel method for calculating prevalence of multiple sclerosis in Australia', *Multiple Sclerosis Journal*, 19(13), pp. 1704–1711. doi: 10.1177/1352458513479841.

Palmer, C. and Pe'er, I. (2017) 'Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies', *PLOS Genetics*. Edited by J. Marchini. Public Library of Science, 13(7), p. e1006916. doi: 10.1371/journal.pgen.1006916.

Papathemeli, D. *et al.* (2016) 'Development of a primary cutaneous CD30(+) anaplastic large-cell T-cell lymphoma during treatment of multiple sclerosis with fingolimod', *Multiple Sclerosis Journal*, 22(14), pp. 1888–1890. doi: 10.1177/1352458516645868.

Pasaniuc, B. and Price, A. L. (2017) 'Dissecting the genetics of complex traits using summary association statistics.', *Nature Reviews Genetics*. NIH Public Access, 18(2), pp. 117–127. doi: 10.1038/nrg.2016.142.

Patrikios, P. *et al.* (2006) 'Remyelination is extensive in a subset of multiple sclerosis patients', *Brain*. Oxford University Press, 129(12), pp. 3165–3172. doi: 10.1093/brain/awl217.

Patsopoulos, N. A. *et al.* (2011) 'Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci', *Annals of Neurology*, 70(6), pp. 897–912. doi: 10.1002/ana.22609.

Patsopoulos, N. A. *et al.* (2013) 'Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects', *PLOS Genetics*. Edited by G. Gibson. Public Library of Science, 9(11), p. e1003926. doi: 10.1371/journal.pgen.1003926.

Patsopoulos, N. A. (2018) 'Genetics of Multiple Sclerosis: An Overview and New Directions', *Cold Spring Harbor Perspectives in Medicine*, p. a028951. doi: 10.1101/cshperspect.a028951.

Patsopoulos, N. A. and (IMSGS), I. M. S. G. C. (2016) '200 loci complete the genetic puzzle of multiple sclerosis.', in *ASHG Annual Meeting*. Vancouver, BC, Canada: NIH Public Access, p. Unit1.19. doi: 10.1002/0471142905.hg0119s68.

Patterson, H. D. and Thompson, R. (1971) 'Recovery of Inter-Block Information when Block Sizes are Unequal', *Biometrika*. Oxford University PressBiometrika Trust, 58(3), p. 545. doi: 10.2307/2334389.

Pearson, T. A. and Manolio, T. A. (2008) 'How to interpret a genome-wide association study.', *JAMA: The Journal of the American Medical Association*. American Medical Association, 299(11), pp. 1335–44. doi: 10.1001/jama.299.11.1335.

Peltonen, L., Palotie, A. and Lange, K. (2000) 'Use of population isolates for mapping complex traits', *Nature Reviews Genetics*. Nature Publishing Group, 1(3), pp. 182–190. doi: 10.1038/35042049.

Perez, M. F. and Lehner, B. (2019) 'Intergenerational and transgenerational epigenetic inheritance in animals', *Nature Cell Biology*. Nature Publishing Group, 21(2), pp. 143–151. doi: 10.1038/s41556-018-0242-9.

Petronis, A. (2010) 'Epigenetics as a unifying principle in the aetiology of complex traits and diseases', *Nature*, 465(7299), pp. 721–727. doi: 10.1038/nature09230.

Peyrot, W. J. *et al.* (2014) 'Effect of polygenic risk scores on depression in childhood trauma.', *The British Journal of Psychiatry : The Journal of Mental Science*. The Royal College of Psychiatrists, 205(2), pp. 113–9. doi: 10.1192/bjp.bp.113.143081.

Pickrell, J. K. *et al.* (2016) 'Detection and interpretation of shared genetic influences on 42 human traits', *Nature Genetics*. Nature Publishing Group, 48(7), pp. 709–717. doi: 10.1038/ng.3570.

Pietiläinen, O. P. H. *et al.* (no date) 'Recessively inherited deletion confers risk for schizophrenia and intellectual disability', *Genome Research*, p. 12159423. Available at: https://genome.cshlp.org/site/press/Pietilainen.pdf (Accessed: 12 July 2019).

Pittock, S. J. *et al.* (2004) 'Clinical implications of benign multiple sclerosis: a 20-year population-based follow-up study.', *Annals of Neurology*, 56(2), pp. 303–6. doi: 10.1002/ana.20197.

Platzer, K. and Lemke, J. R. (1993) 'GRIN2B-Related Neurodevelopmental Disorder', *GeneReviews®*. University of Washington, Seattle. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29851452 (Accessed: 12 March 2019).

Plenge, R. M., Scolnick, E. M. and Altshuler, D. (2013) 'Validating therapeutic targets through human genetics', *Nature Reviews Drug Discovery*. Nature Research, 12(8), pp. 581–594. doi: 10.1038/nrd4051.

Polman, C. H. *et al.* (2006) 'A Randomized, Placebo-Controlled Trial of Natalizumab for Relapsing Multiple Sclerosis', *New England Journal of Medicine*. Massachusetts Medical Society , 354(9), pp. 899–910. doi: 10.1056/NEJMoa044397.

Polman, C. H. *et al.* (2011) 'Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria', *Annals of Neurology*, 69(2), pp. 292–302. doi: 10.1002/ana.22366.

Poser, C. M. (1994) 'The epidemiology of multiple sclerosis: A general overview', *Annals of Neurology*. Wiley-Blackwell, 36(S2), pp. S180–S193. doi: 10.1002/ana.410360805.

Pritchard, J. K. (2002) 'The allelic architecture of human disease genes: common disease-common variant... or not?', *Human Molecular Genetics*, 11(20), pp. 2417–2423. doi: 10.1093/hmg/11.20.2417.

Pugliatti, M., Sotgiu, S. and Rosati, G. (2002) 'The worldwide prevalence of multiple sclerosis.', *Clinical Neurology and Neurosurgery*, 104(3), pp. 182–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12127652 (Accessed: 27 July 2017).

Pugliese, A. *et al.* (1997) 'The insulin gene is transcribed in the human thymus and transcription levels correlate with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes', *Nature Genetics*, 15(3), pp. 293–297. doi: 10.1038/ng0397-293.

Purcell, S. *et al.* (2007) *PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses*, *American Journal of Human Genetics*. doi: 10.1086/519795.

Purcell, S. M. *et al.* (2009) 'Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.', *Nature*, 460(7256), pp. 748–52. doi: 10.1038/nature08185.

Purcell, S. M. *et al.* (2014) 'A polygenic burden of rare disruptive mutations in schizophrenia.', *Nature*, 506(7487), pp. 185–90. doi: 10.1038/nature12975.

Ragonese, P. *et al.* (2008) 'Mortality in multiple sclerosis: a review.', *European Journal of Neurology*, 15(2), pp. 123–7. doi: 10.1111/j.1468-1331.2007.02019.x.

Ramagopalan, S. V. *et al.* (2011) 'Rare variants in the CYP27B1 gene are associated with multiple sclerosis', *Annals of Neurology*, 70(6), pp. 881–886. doi: 10.1002/ana.22678.

Ramagopalan, S. V, Giovannoni, G., *et al.* (2009) 'Can we predict multiple sclerosis?', *Lancet Neurology*. Elsevier, 8(12). doi: 10.1016/S1474-4422(09)70273-X.

Ramagopalan, S. V, Maugeri, N. J., *et al.* (2009) 'Expression of the Multiple Sclerosis-Associated MHC Class II Allele HLA-DRB1*1501 Is Regulated by Vitamin D', *PLOS Genetics*. Edited by D. C. Roopenian. Public Library of Science, 5(2), pp. 938–952. doi: 10.1371/JOURNAL.PGEN.1000369.

Ramagopalan, S. V *et al.* (2010) 'Multiple sclerosis: risk factors, prodromes, and potential causal pathways', *Lancet Neurology*, 9(7), pp. 727–739. doi: 10.1016/S1474-4422(10)70094-6.

Reich, D. E. and Lander, E. S. (2001) 'On the allelic spectrum of human disease', *Trends in Genetics*, 17(9), pp. 502–510. doi: 10.1016/S0168-9525(01)02410-6.

Ridaura, V. K. *et al.* (2013) 'Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice', *Science*, 341(6150), pp. 1241214–1241214. doi: 10.1126/science.1241214.

Riise, T., Nortvedt, M. W. and Ascherio, A. (2003) 'Smoking is a risk factor for multiple sclerosis.', *Neurology*, 61(8), pp. 1122–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14581676 (Accessed: 27 July 2017).

Ringnér, M. (2008) 'What is principal component analysis?', *Nature Biotechnology*. Nature Publishing Group, 26(3), pp. 303–304. doi: 10.1038/nbt0308-303.

Ripke, S. *et al.* (2014) 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*, 511(7510).

Ristori, G. *et al.* (2006) 'Multiple sclerosis in twins from continental Italy and Sardinia: A nationwide study', *Annals of Neurology*, 59(1), pp. 27–34. doi: 10.1002/ana.20683.

Rito, Y. *et al.* (2018) 'Epigenetics in Multiple Sclerosis: Molecular Mechanisms and

Dietary Intervention', *Central Nervous System Agents in Medicinal Chemistry*, 18(1). doi: 10.2174/1871524916666160226131842.

Rivera, V. M. and Cabrera, J. A. (2001) 'Aboriginals with multiple sclerosis: HLA types and predominance of neuromyelitis optica.', *Neurology*, 57(5), pp. 937–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11552042 (Accessed: 15 May 2018).

Rizvi, S. A. and Agius, M. A. (2004) 'Current approved options for treating patients with multiple sclerosis', *Neurology*. Lippincott Williams & Wilkins, 63(Issue 12, Supplement 6), pp. S8–S14. doi: 10.1212/WNL.63.12_suppl_6.S8.

Roberts, D. F., Roberts, M. J. and Poskanzer, D. C. (1983) 'Genetic analysis of multiple sclerosis in Shetland.', *Journal of epidemiology and community health*. BMJ Publishing Group, 37(4), pp. 281–5. doi: 10.1136/jech.37.4.281.

Rosati, G. (2001) 'The prevalence of multiple sclerosis in the world: an update.', *Neurological Sciences : Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 22(2), pp. 117–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11603614 (Accessed: 15 May 2018).

Røysamb, E. and Tambs, K. (2016) 'The beauty, logic and limitations of twin studies', *Norsk Epidemiologi*, 26(1–2). doi: 10.5324/nje.v26i1-2.2014.

Runmarker, B. and Andersen, O. (1993) 'Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up.', *Brain : A Journal of Neurology*, 116 (Pt 1, pp. 117–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8453453 (Accessed: 27 January 2016).

Sadovnick, A. D. *et al.* (1996) 'Evidence for genetic basis of multiple sclerosis. The Canadian Collaborative Study Group.', *Lancet*, 347(9017), pp. 1728–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8656905 (Accessed: 28 July 2017).

Sadovnick, A. D. *et al.* (2004) 'A population-based study of multiple sclerosis in twins: Update', *Annals of Neurology*, 33(3), pp. 281–285. doi: 10.1002/ana.410330309.

Sadovnick, A. D. *et al.* (2017) 'Purinergic receptors *P2RX4* and *P2RX7* in familial multiple sclerosis', *Human Mutation*, 38(6), pp. 736–744. doi: 10.1002/humu.23218.

Sadovnick, A. D., Bulman, D. and Ebers, G. C. (1991) 'Parent-child concordance in

multiple sclerosis', *Annals of Neurology*, 29(3), pp. 252–255. doi: 10.1002/ana.410290304.

Sandoval-Motta, S. *et al.* (2017) 'The Human Microbiome and the Missing Heritability Problem.', *Frontiers in Genetics*. Frontiers Media SA, 8, p. 80. doi: 10.3389/fgene.2017.00080.

Sawcer, S. *et al.* (2011) 'Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.', *Nature*, 476(7359), pp. 214–9. doi: 10.1038/nature10251.

Sayao, A.-L., Devonshire, V. and Tremlett, H. (2007) 'Longitudinal follow-up of "benign" multiple sclerosis at 20 years.', *Neurology*, 68(7), pp. 496–500. doi: 10.1212/01.wnl.0000253185.03943.66.

Scalfari, A. *et al.* (2010) 'The natural history of multiple sclerosis: a geographically based study 10: relapses and long-term disability.', *Brain : A Journal of Neurology*, 133(Pt 7), pp. 1914–29. doi: 10.1093/brain/awq118.

Scottish Government (2018) *Scottish health survey: results for local areas 2014 to 2017*. Available at: https://www.gov.scot/publications/scottish-health-survey-results-local-areas-2014-2015-2016-2017/pages/2/ (Accessed: 25 June 2019).

Scottish Public Health Record (2012) *Health Care Needs Assessment of Services for Adults with Rheumatoid Arthritis*. Available at: https://www.scotphn.net/wp-content/uploads/2016/03/2012_09_12-PART-B_RA-HCNA_EPIDEMIOLOGY-ES-V2-1.pdf (Accessed: 9 July 2019).

Selmi, C., Lu, Q. and Humble, M. C. (2012) 'Heritability versus the role of the environment in autoimmunity', *Journal of Autoimmunity*, 39(4), pp. 249–252. doi: 10.1016/j.jaut.2012.07.011.

Shahbazi, M. *et al.* (2010) 'High frequency of the IL-2 –330 T/HLA-DRB1*1501 haplotype in patients with multiple sclerosis', *Clinical Immunology*, 137(1), pp. 134–138. doi: 10.1016/j.clim.2010.05.010.

Shahbazi, M. *et al.* (2011) 'Interaction of HLA-DRB1*1501 allele and TNF-alpha – 308 G/A single nucleotide polymorphism in the susceptibility to multiple sclerosis', *Clinical*

Immunology. Academic Press, 139(3), pp. 277–281. doi: 10.1016/J.CLIM.2011.02.012.

Shan, M. *et al.* (2009) 'Lung Myeloid Dendritic Cells Coordinately Induce TH1 and TH17 Responses in Human Emphysema', *Science Translational Medicine*, 1(4), pp. 4ra10-4ra10. doi: 10.1126/scitranlsmed.3000154.

Silventoinen, K. *et al.* (2003) 'Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries', *Twin Research*, 6(5), pp. 399–408. doi: 10.1375/136905203770326402.

Simpson, S. *et al.* (2011) 'Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis.', *Journal of Neurology, Neurosurgery and Psychiatry*. BMJ Publishing Group Ltd, 82(10), pp. 1132–1141. doi: 10.1136/jnnp.2011.240432.

Sirota, M. *et al.* (2009) 'Autoimmune Disease Classification by Inverse Association with SNP Alleles', *PLOS Genetics*. Edited by D. B. Allison, 5(12), p. e1000792. doi: 10.1371/journal.pgen.1000792.

Sivakumaran, S. *et al.* (2011) 'Abundant pleiotropy in human complex diseases and traits.', *American Journal of Human Genetics*, 89(5), pp. 607–18. doi: 10.1016/j.ajhg.2011.10.004.

Smith, B. and Great Britain. Environment Agency. (2003) *Information on land quality in Scotland : sources of information (including background contaminants)*. Environment Agency. Available at: https://www.gov.uk/government/publications/information-on-land-quality-in-scotland-sources-of-information-including-background-contaminants (Accessed: 26 June 2019).

Smith, B. H. *et al.* (2013) 'Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness', *International Journal of Epidemiology*, 42(3), pp. 689–700. doi: 10.1093/ije/dys084.

So, H.-C. and Sham, P. C. (2017) 'Improving polygenic risk prediction from summary statistics by an empirical Bayes approach.', *Scientific Reports*. Nature Publishing Group, 7, p. 41262. doi: 10.1038/srep41262.

Soilu-Hänninen, M. *et al.* (2005) '25-Hydroxyvitamin D levels in serum at the onset of multiple sclerosis', *Multiple Sclerosis Journal*, 11(3), pp. 266–271. doi: 10.1191/1352458505ms1157oa.

Sotgiu, S. *et al.* (2003) 'Does the "hygiene hypothesis" provide an explanation for the high prevalence of multiple sclerosis in Sardinia?', *Autoimmunity*, 36(5), pp. 257–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14567554 (Accessed: 27 July 2017).

Sotgiu, S. *et al.* (2004) 'Genes, environment, and susceptibility to multiple sclerosis', *Neurobiology of Disease*, 17(2), pp. 131–143. doi: 10.1016/j.nbd.2004.07.015.

Spain, S. L. and Barrett, J. C. (2015) 'Strategies for fine-mapping complex traits.', *Human Molecular Genetics*. Oxford University Press, 24(R1), pp. R111-9. doi: 10.1093/hmg/ddv260.

Speed, D. *et al.* (2012) 'Improved heritability estimation from genome-wide SNPs.', *American Journal of Human Genetics*. Elsevier, 91(6), pp. 1011–21. doi: 10.1016/j.ajhg.2012.10.010.

Stahl, E. A. *et al.* (2010) 'Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci', *Nature Genetics*, 42(6), pp. 508–514. doi: 10.1038/ng.582.

Staley, J. R. *et al.* (2016) 'PhenoScanner: A database of human genotype-phenotype associations', *Bioinformatics*, 32(20), pp. 3207–3209. doi: 10.1093/bioinformatics/btw373.

Stanton-Geddes, J. *et al.* (2013) 'Estimating heritability using genomic data', *Methods in Ecology and Evolution*. Edited by J. Hadfield. Wiley/Blackwell (10.1111), 4(12), pp. 1151–1158. doi: 10.1111/2041-210X.12129.

Sudlow, C. *et al.* (2015) 'UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.', *PLOS Medicine*. Public Library of Science, 12(3), p. e1001779. doi: 10.1371/journal.pmed.1001779.

Sundqvist, E. *et al.* (2012) 'Epstein-Barr virus and multiple sclerosis: interaction with HLA', *Genes and Immunity*, 13(1), pp. 14–20. doi: 10.1038/gene.2011.42.

Taylor, B. V *et al.* (2010) 'MS prevalence in New Zealand, an ethnically and latitudinally diverse country', *Multiple Sclerosis Journal*, 16(12), pp. 1422–1431. doi: 10.1177/1352458510379614.

Tesli, M. *et al.* (2014) 'Polygenic risk score and the psychosis continuum model', *Acta Psychiatrica Scandinavica*, 130(4), pp. 311–317. doi: 10.1111/acps.12307.

The 1000 Genomes Project Consortium (2015) 'A global reference for human genetic variation', *Nature*. Nature Publishing Group, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

The International HapMap Consortium, Altshuler, D. and Donnelly, P. (2005) 'A haplotype map of the human genome', *Nature*. Nature Publishing Group, 437(7063), pp. 1299–1320. doi: 10.1038/nature04226.

Tienari, P. J. *et al.* (2004) 'Multiple sclerosis in western Finland: evidence for a founder effect', *Clinical Neurology and Neurosurgery*, 106(3), pp. 175–179. doi: 10.1016/j.clineuro.2004.02.009.

Traboulsee, A. L. *et al.* (2017) 'Common genetic etiology between "multiple sclerosis-like" single-gene disorders and familial multiple sclerosis', *Human Genetics*, 136(6), pp. 705–714. doi: 10.1007/s00439-017-1784-9.

Tremlett, H. and Devonshire, V. (2008) 'Natural history of secondary-progressive multiple sclerosis', *Multiple Sclerosis*, 14(3), pp. 314–324. doi: 10.1177/1352458507084264.

Tremlett, H., Paty, D. and Devonshire, V. (2006) 'Disability progression in multiple sclerosis is slower than previously reported.', *Neurology*, 66(2), pp. 172–7. doi: 10.1212/01.wnl.0000194259.90286.fe.

Tremlett, H. and Rieckmann, P. (2010) 'New perspectives in the natural history of multiple sclerosis', *Neurology*.

Tsang, B. K.-T. and Macdonell, R. (2011) 'Multiple sclerosis- diagnosis, management and prognosis.', *Australian Family Physician*, 40(12), pp. 948–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22146321 (Accessed: 15 May 2018).

Turner, S. *et al.* (2011) 'Quality control procedures for genome-wide association

studies.', *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*. NIH Public Access, Chapter 1, p. Unit1.19. doi: 10.1002/0471142905.hg0119s68.

Umeton, R. *et al.* (2018) 'The Gut Microbiome in Relapsing Multiple Sclerosis Patients Compared to Controls. (P2.355)', *Neurology*. Advanstar Communications, 90(15 Supplement), p. P2.355. Available at: https://n.neurology.org/content/90/15_Supplement/P2.355 (Accessed: 25 June 2019).

Vilariño-Güell, C. *et al.* (2019) 'Exome sequencing in multiple sclerosis families identifies 12 candidate genes and nominates biological pathways for the genesis of disease', *PLOS Genetics*. Edited by G. Sirugo. Public Library of Science, 15(6), p. e1008180. doi: 10.1371/journal.pgen.1008180.

Virtanen, J. O. and Jacobson, S. (2012) 'Viruses and multiple sclerosis.', *CNS & Neurological Disorders Drug Targets*. NIH Public Access, 11(5), pp. 528–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22583435 (Accessed: 24 June 2019).

Visscher, P. M. *et al.* (2014) 'Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples.', *PLoS genetics*. Public Library of Science, 10(4), p. e1004269. doi: 10.1371/journal.pgen.1004269.

Visscher, P. M. *et al.* (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation.', *American Journal of Human Genetics*. Elsevier, 101(1), pp. 5–22. doi: 10.1016/j.ajhg.2017.06.005.

Visscher, P. M., Hill, W. G. and Wray, N. R. (2008) 'Heritability in the genomics era — concepts and misconceptions', *Nature Reviews Genetics*. Nature Publishing Group, 9(4), pp. 255–266. doi: 10.1038/nrg2322.

Visser, E. M. *et al.* (2012) 'A new prevalence study of multiple sclerosis in Orkney, Shetland and Aberdeen city.', *Journal of Neurology, Neurosurgery and Psychiatry*. BMJ Publishing Group Ltd, 83(7), pp. 719–24. doi: 10.1136/jnnp-2011-301546.

Vitart, V. *et al.* (2005) *Increased Level of Linkage Disequilibrium in Rural Compared with Urban Communities: A Factor to Consider in Association-Study Design*, *Am. J. Hum. Genet.* Available at: https://www.cell.com/ajhg/pdf/S0002-9297(07)60723-X.pdf (Accessed: 9 August 2019).

Vizier, C. *et al.* (1999) 'Role of autoreactive CD8+ T cells in organ-specific autoimmune diseases: insight from transgenic mouse models', *Immunological Reviews*. Blackwell Publishing Ltd, 169(1), pp. 81–92. doi: 10.1111/j.1600-065X.1999.tb01308.x.

Wadia, N. H. and Bhatia, K. (1990) 'Multiple sclerosis is prevalent in the zoroastrians (Parsis) of India', *Annals of Neurology*, 28(2), pp. 177–179. doi: 10.1002/ana.410280211.

Wagner, G. P. and Zhang, J. (2011) 'The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms', *Nature Reviews Genetics*. Nature Publishing Group, 12(3), pp. 204–213. doi: 10.1038/nrg2949.

Wang, Z. *et al.* (2016) 'Nuclear Receptor NR1H3 in Familial Multiple Sclerosis', *Neuron*, 90(5), pp. 948–954. doi: 10.1016/j.neuron.2016.04.039.

Ware, E. B. *et al.* (no date) 'Heterogeneity in polygenic scores for common human traits', *bioRxiv*. doi: 10.1101/106062.

Watson, C. T. *et al.* (2012) 'Estimating the proportion of variation in susceptibility to multiple sclerosis captured by common SNPs', *Scientific Reports*. Nature Publishing Group, 2, pp. 938–952. doi: 10.1038/srep00770.

Weinshenker, B. G. *et al.* (1989) 'The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability.', *Brain : A Journal of Neurology*, 112 (Pt 1, pp. 133–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2917275 (Accessed: 13 December 2015).

Weinshenker, B. G. *et al.* (1991) 'The natural history of multiple sclerosis: a geographically based study. 3. Multivariate analysis of predictive factors and models of outcome.', *Brain : A Journal of Neurology*, 114 (Pt 2, pp. 1045–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2043940 (Accessed: 2 February 2016).

Weiss, E. *et al.* (2016) 'Farming, Foreign Holidays, and Vitamin D in Orkney', *PLOS ONE*. Edited by A. T. Slominski. Public Library of Science, 11(5), p. e0155633. doi: 10.1371/journal.pone.0155633.

Welter, D. *et al.* (2014) 'The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.', *Nucleic Acids Research*. Oxford University Press, 42(Database issue), pp.

D1001-6. doi: 10.1093/nar/gkt1229.

Westerlind, H., Ramanujam, R., *et al.* (2014) 'Modest familial risks for multiple sclerosis: a registry-based study of the population of Sweden', *Brain*, 137(3), pp. 770-. doi: 10.1093/brain/awt356.

Westerlind, H., Kuja-Halkola, R., *et al.* (2014) 'Reply: Shared environmental effects on multiple sclerosis susceptibility: conflicting evidence from twin studies', *Brain*. Oxford University Press, 137(7), pp. e288–e288. doi: 10.1093/brain/awu099.

Widdifield, J. *et al.* (2015) 'Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario, Canada', *Multiple Sclerosis Journal*, 21(8), pp. 1045–1054. doi: 10.1177/1352458514556303.

Wilson, J. F. *et al.* (2001) 'Genetic evidence for different male and female roles during cultural transitions in the British Isles.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 98(9), pp. 5078–83. doi: 10.1073/pnas.071036898.

Wilson, P. W. F. *et al.* (1998) 'Prediction of Coronary Heart Disease Using Risk Factor Categories', *Circulation*. Lippincott Williams & Wilkins, 97(18), pp. 1837–1847. doi: 10.1161/01.CIR.97.18.1837.

Wingerchuk, D. M. and Carter, J. L. (2014) 'Multiple sclerosis: current and emerging disease-modifying therapies and treatment strategies.', *Mayo Clinic Proceedings*. Elsevier, 89(2), pp. 225–40. doi: 10.1016/j.mayocp.2013.11.002.

Wirz, S. *et al.* (2004) 'High frequency of TNF alleles 238A and 376A in individuals from northern Sardinia', *Cytokine*, 26(4), pp. 149–154. doi: 10.1016/j.cyto.2004.02.006.

World Health Organization (2008) 'Atlas: Multiple Sclerosis Resources in the World 2008', *World Health Organization*. WHO Press, Geneva.

Wortsman, J. *et al.* (2000) 'Decreased bioavailability of vitamin D in obesity', *The American Journal of Clinical Nutrition*, 72(3), pp. 690–693. doi: 10.1093/ajcn/72.3.690.

Wray, N. R., Goddard, M. E. and Visscher, P. M. (2007) 'Prediction of individual genetic

risk to disease from genome-wide association studies.', *Genome Research*. Cold Spring Harbor Laboratory Press, 17(10), pp. 1520–8. doi: 10.1101/gr.6665407.

Wu, C. *et al.* (2011) 'A comparison of association methods correcting for population stratification in case-control studies.', *Annals of Human Genetics*. NIH Public Access, 75(3), pp. 418–27. doi: 10.1111/j.1469-1809.2010.00639.x.

Xue, Y. *et al.* (2017) 'Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations', *Nature Communications*. Nature Publishing Group, 8(1), p. 15927. doi: 10.1038/ncomms15927.

Yang, J. *et al.* (2010) 'Common SNPs explain a large proportion of the heritability for human height.', *Nature Genetics*. NIH Public Access, 42(7), pp. 565–9. doi: 10.1038/ng.608.

Yang, J. *et al.* (2011) 'GCTA: A Tool for Genome-wide Complex Trait Analysis', *American Journal of Human Genetics*. Cell Press, 88(1), pp. 76–82. doi: 10.1016/J.AJHG.2010.11.011.

Yatsunenko, T. *et al.* (2012) 'Human gut microbiome viewed across age and geography.', *Nature*. NIH Public Access, 486(7402), pp. 222–7. doi: 10.1038/nature11053.

Yeo, T. W. *et al.* (2007) 'A second major histocompatibility complex susceptibility locus for multiple sclerosis.', *Annals of Neurology*. Wiley-Blackwell, 61(3), pp. 228–36. doi: 10.1002/ana.21063.

Youl, B. D. *et al.* (1991) 'Destructive lesions in demyelinating disease.', *Journal of Neurology, Neurosurgery and Psychiatry*, 54(4), pp. 288–292. doi: 10.1136/jnnp.54.4.288.

Zaitlen, N. *et al.* (2013) 'Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits', *PLOS Genetics*. Edited by P. M. Visscher. Public Library of Science, 9(5), p. e1003520. doi: 10.1371/journal.pgen.1003520.

Zarate, C. A. *et al.* (2006) 'A Randomized Trial of an N-methyl-D-aspartate Antagonist in Treatment-Resistant Major Depression', *Archives of General Psychiatry*. American Medical Association, 63(8), p. 856. doi: 10.1001/archpsyc.63.8.856.

Zheng, J. *et al.* (2017) 'LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis', *Bioinformatics*. Oxford University Press, 33(2), pp. 272–279. doi: 10.1093/bioinformatics/btw613.

Zheng, S. L. *et al.* (2008) 'Cumulative Association of Five Genetic Variants with Prostate Cancer', *New England Journal of Medicine*, 358(9), pp. 910–919. doi: 10.1056/NEJMoa075819.

Zhu, Z. *et al.* (2015) 'Dominance genetic variation contributes little to the missing heritability for human complex traits.', *American Journal of Human Genetics*. Elsevier, 96(3), pp. 377–85. doi: 10.1016/j.ajhg.2015.01.001.

Zhu, Z. *et al.* (2016) 'Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets', *Nature Genetics*, 48(5), pp. 481–487. doi: 10.1038/ng.3538.

Zuk, O. O. *et al.* (2012) 'The mystery of missing heritability: Genetic interactions create phantom heritability.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 109(4), pp. 1193–1198. doi: 10.1073/pnas.1119675109.

GLOSSARY

Epidemiological vocabulary

Aetiology	The cause(s) of a disease or condition
Epidemiology	The study of health and disease determinants and distributions
Incidence rate	The number of new cases that develop in a population within a specific time-period
Inflammatory disease	A group of diseases which result from an individual's immune system attacking the body's own cells or tissues
Prevalence rate	The number of cases present in a population at a specific point in time

Genetic vocabulary

Additive genetic effects	When the combined effects of alleles are equal to the sum of their individual effects
Allele	One form of a given gene
Allelic Dosage	The estimated number of copies of each allele
Batch effects	Variation introduced to samples due to laboratory handling procedures
Call rate	The number of SNPs on a chip receiving a genotype call divided by the total number of SNPs

Candidate gene study	A study that evaluates genetic variation in a specific region of genes, chosen using prior knowledge
Causal variant	A variant that has a causal effect on a trait (as opposed to a variant that is only associated with a trait due to LD)
Common variant	A variant with a population frequency greater than 5%
Complex disease	A disease influenced by a combination of environmental factors and multiple genes
Copy number variation	Where sections of the genome are repeated, with the number of repeats varying between individuals in a population
Deletion	A type of DNA mutation where part of a DNA sequence is lost during DNA replication
Disease risk factor	Any attribute or exposure that increases an individual's likelihood of developing a disease
DNA strand	DNA is double-stranded; for a reference chromosome, one of these strands is deemed the forward or plus strand, while the complimentary strand is deemed the reverse or minus strand
Dominance	Where an allele at one gene masks the contribution of another allele at the same gene
Effect size	The affect size of an allele is the magnitude of the effect on the phenotype
Enhancer region	A short stretch of DNA which increases the likelihood of transcription of a nearby gene
Epigenetics	The study of heritable changes within the phenotype that are not caused by DNA sequence alterations
Epistasis	Interactions between genes where the effect of one gene is dependent on the presence of one or more modifier genes

Exon	A region of DNA or RNA that codes for a protein or peptide sequence
Exon splicing	A process where introns are removed, and exons are joined together
Fine-mapping	A method used to determine the causal variant, by taking evidence of association from a GWAS to identify and subsequently explore regions of interest for independent effects on the trait
Gene	A region of DNA which can influence one or more phenotypes and can be transferred from parent to offspring
Gene expression	The process where information from a gene is used to synthesize a gene product
Genetic bottleneck	A sharp decline in the size of a population due to an extreme event(s) such as famine or genocide
Genetic drift	The change in frequency of an allele in a population due to random sampling
Genetic variant	An alteration in a common DNA sequence; includes SNPs and insertions/deletions
Genome	The full set of genetic material present in an organism or cell
Genome-wide association study (GWAS)	A method used to detect associations between a trait of interest and genetic variants in a population sample
Genotype	The genetic composition of an organism, or more specifically the set of alleles carried by an organism
Genotyping platforms	Equipment designed to carry out genotyping, for example array technologies
Haplotype	A group of alleles that tend to always occur together and are likely inherited together

Haplotype reference consortium	A large reference panel of human haplotypes
Hardy-Weinberg equilibrium	An equation that can be used to calculate genetic variation of a population at equilibrium. The principal of the HWE states that the genetic variation (allele frequencies) within a population will remain constant between generations in the absence of disrupting factors
Heritability	The proportion of variance of a trait within a population at a specific time that is due to genetic variation between individuals
Heterozygosity	The possession of two different alleles of a gene by an individual
High throughput technologies	Refers to sequencing techniques which allows large amounts of DNA to be sequenced at once
Homozygosity	The possession of two identical alleles of a gene by an individual
Human leukocyte antigen complex	A gene complex that encodes the major histocompatibility complex in humans
Identity by descent	Identical segments of DNA which are shared by two or more people that have been inherited from a common ancestor
Identity by state	Identical segments of DNA which is shared by two or more people that do not have a common ancestor
Imputation	The statistical inference to determine unobserved genotypes
Inbreeding	The production of offspring from mating closely genetically related individuals
Insertion	A type of DNA mutation where one or more nucleotide base pairs are inserted into a DNA sequence

Intragenic	A stretch of DNA sequence that is positioned between genes
Intron	A region of DNA or RNA within a gene that does not code for a protein
Inversion	A type of DNA mutation where a segment of a chromosome is reversed end to end
Jackpot effect	Where the frequency of a rare variant is increased in a population or group by chance due to population events
Linkage disequilibrium	The non-random association between alleles at separate loci
Linkage study	A family-based study used to map a trait to a specific location in the genome through the identification of segments of DNA that co-segregate with the trait through families
Locus	A location on a chromosome
Low frequency variant	A variant with a minor allele frequency between 0.5% and 5%
Major histocompatibility complex	A gene complex that codes for cell surface proteins, necessary for the acquired immune system to recognise foreign molecules
Marker	A DNA sequence whose locus on the genome is known
Mendelian disease	Diseases caused by mutations found in one gene
Mendelian randomisation	Where genetic variants are used as instrumental variables to determine if a modifiable exposure is causally related to a specific trait.
Monogenic disease	A disease caused by a single defective gene

Mutation	A permanent alteration of the DNA sequence of a gene, caused by altering a single nucleotide or the deletion, insertion or rearrangement of DNA segments
Non-synonymous mutation	A change in the DNA sequence that changes an amino acid within a protein
Oligogenic disease	A disease which is caused by a small number of genes
Parent-of-origin effect	When the phenotypic effect of an allele depends on whether it is inherited from an individual's father or mother
Phasing	Assigning alleles to maternal and paternal chromosomes
Phenotype	An observable physical property in an organism resulting from both its genotype and environment
Pleiotropic	When one gene influences more than one phenotypic trait
Polygenic risk score	A genetic risk score based on genetic loci and their associated weights
Polygenic trait	A trait controlled by more than one gene
Population stratification	The presence of systematic differences in allele frequencies between groups within a population, for example due to different ancestries
Population structure	The presence of subpopulations within a population which can have differences to the main population, e.g. in allele frequency
Promotor region	A region of DNA that leads to the start of transcription of a specific gene
Protein isoforms	Very similar proteins that originate from the same gene but have differences in structure and function due to genetic differences

Proxy variant	A variant which represents another, usually due to very high LD between the two variants
Rare variant	A variant with a minor allele frequency below 0.5%
Recessive	Relating to a trait controlled by genes which are expressed only when the variant is inherited from both parents
Recombination rate	The rate at which recombination happens; the process where DNA is broken and recombined to create new allele combinations
Recurrence rate	The risk that offspring will be affected by a disease or trait given a specific set of relatives
Reference allele	The allele that is found in the reference genome; not necessarily the major allele
Relatedness coefficient	A measure of the degree of consanguinity between two individuals; double the kinship coefficient
Reproductive fitness	The reproductive success of an individual; a quantitative representation of selection
Risk allele	The allele which confers a risk of developing a trait, usually a disease
Single nucleotide polymorphism	A single substitution of a nucleotide within a DNA sequence
SNP array	A group of microscopic DNA spots attached to a surface, designed to detect single nucleotide polymorphisms within individuals
Splicing	The process of removing introns and joining exons together in messenger RNA
Strand flip	Changing the strand (plus or minus) of a SNP
Synonymous mutation	A change in the DNA sequence that codes for an amino acid, but the amino acid is not changed

Tag SNP	A SNP that exists in a region of high LD that is used to represent a specific haplotype
Transcription	The process by which DNA is copied into RNA by RNA polymerase; it is the first stage of gene expression
Translation	The process of translating a sequence of mRNA to amino acids to synthesise proteins
Translocation	The movement of a segment of chromosome from one position to another (either on the same or different chromosome)
Whole-genome sequencing	The process of determining an organism's complete DNA sequence
Winner's curse	The tendency for initial studies of a trait to overestimate the effect size of some variants, the very fact it was found in that study generates an upward bias in the estimate.

Mathematical vocabulary

Area under the curve (AUC)	The area under a curve; when used in relation to a ROC curve, it can be used as a method of evaluating a model's performance, with a high AUC value corresponding to a better predictive ability of the model
Association test	A study used to determine if a genetic variant is associated with a specific trait
Beta	The resulting coefficient from a model fit; gives an indication of effect size
Chi-squared statistic	A measurement for how expectation of data compares to observed data

Confounding	Occurs when a variable influences both the independent and dependent variable to cause a spurious association
Correlation coefficient	A number between -1 and 1 that describes a linear dependence between two variables
Covariate	A variable that may be predictive of the outcome of a study, or may be a confounding or interacting variable
Fixed effects model	A model where the independent variables are constant, with only the dependent variable changing in response to differing levels of independent variables
Genetic relationship matrix (GRM)	A matrix of genetic pairwise relationships for a group of individuals, where genetic relationships are estimated using SNP data
Genome-based restricted maximum likelihood (GREML)	A statistical method used to estimate SNP heritability
GRAMMAR+ Residuals	Residuals (the difference between an observed value and its estimated value) produced using the genome wide rapid association using mixed model and regression
Incidence matrix	A matrix containing relationship-values between two classes of objects
Kinship coefficient	The probability that a pair of randomly sampled alleles from two individuals are identical and from the same ancestor; a measure of relatedness
Kinship matrix	A matrix containing kinship coefficient values for a group of individuals
Liability	In the context of genetics, liability describes the collective contribution of genetic and environmental factors in developing a trait (usually a disease)
Linear fixed effects model	A linear regression model that only uses fixed effects

Linear mixed effects model	A linear regression model that uses both fixed and random effects
Linear regression	A predictive analysis that models the relationship between a dependent variable and one or more independent variables using a linear approach
Linkage disequilibrium score regression (LDSR)	A statistical approach to quantify the contribution of polygenic effects and confounding factors using GWAS summary statistics
Log likelihood	The natural logarithm of the likelihood; used to determine optimal values for the coefficients estimated by a model
Logistic regression	A predictive analysis that models the relationship between a dependent variable and one or more independent variables where the dependent variable is binary
Logit transformation	A method to transform sigmoid distributions into a linear distribution
Meta-analysis	Pooling data from multiple independent studies on a trait to determine an overall effect
Nagelkerke's pseudo r^2	A metric used to determine the goodness-of-fit of a model by comparing the improvement of the full model compared to the intercept model
Normal cumulative distribution	A function that shows the area under the probability density function from minus infinity to x, the distribution of random variable X; it is a logistic distribution
Normal probability density function	A function that describes a probability distribution for a continuous random variable, where the AUC for a given interval on the x-axis is the probability of the random variable occurring
Null model	A model generated with random samples under a specific distribution where some elements are constant, and some can vary stochastically; for

	example, the null model for Y = $\alpha + \beta 1X + \epsilon$ would be Y = $\alpha + \epsilon$, where $\alpha + \epsilon$ would be equal to the mean of Y
Odds ratio	A measurement of the strength of association between two events A and B; it is the ratio of the odds of B occurring in the presence of A and the odds of B occurring in the absence of A; a value above 1 indicates A and B are positively correlated, a value below 1 indicates A and B are negatively correlated, and a value of 1 indicates A and B are independent
Power	The probability of a test to reject a false null hypothesis
Principal component analysis	A method that converts a group of observations which may be correlated into linearly uncorrelated variables (called principal components); the first component accounts for as much variability in the data as possible, and every subsequent component has the highest variance under the constraints of the preceding components; it can be used to visualise genetic distance and relatedness
p-value	The probability of finding observed results when the null hypothesis of a study is true
Quantile-Quantile (Q-Q) plot	A probability plot that compares the quantiles from two distributions (with the x and y axis corresponding to the two distributions); if the points lie approximately along $y=x$, the distributions are deemed similar
Random effects model	A model where the model parameters are random variables
Rare variant burden test	A test that combines rare variant information within a single region into a summary dose variable
Receiver Operator Characteristic (ROC) curve	A curve which plots the true positive rate (sensitivity) against the false positive rate (specificity) of a binary classifier model to determine its diagnostic ability

Significance threshold	A threshold chosen for a specific test; corresponds to an appropriate probability of committing a type I error (or registering a false positive)
Standard deviation	A measure used to quantify the dispersion of a set of data values; a high standard deviation indicates that data points are spread out over a wide range of values
Two-sample t-test	A statistical test that is used to test the difference between two population means
Type I error / False Positive	When the null hypothesis is true, but is rejected
Type II error / False Negative	When the null hypothesis is false, but is not rejected
Vector	A quantity that has both direction and magnitude; they remain invariant to the coordinate system
Z-score	The number of standard deviations above or below a population mean a data point is

Cellular vocabulary

Adaptive immunity	A subsystem of the immune system made up of specialized cells and processes that destroy pathogens or prevent pathogen growth; it creates an immunological memory after an initial pathogen response
Antibody	A protein produced in response to the presence of an antigen; antibodies attach to antigens to help remove pathogens from the body
Antibody effector function	The action of an antibody to destroy or neutralise a pathogen, for example antigens can block pathogen action or activate other parts of the immune system

Antigen	A foreign or toxic substance which can produce an immune response, particularly via increasing antibody production
Astrocyte	Star-shaped non-neuronal cells found in the brain and spinal cord; a large variety exist and help in CNS functioning
Autoreactive	Acting against the organism that produced it
Axon	The nerve fibre for a neuron that conducts electrical impulses away from the neuron body
B cell	A white blood cell which is part of the adaptive immune system; can bind to specific antigens to initiate an antibody response, as well as present antigens and secrete cytokines
Blood brain barrier	A semi-permeable barrier that separates circulating blood from the brain and CNS fluid
Bystander signal	A bystander signal can occur in cytokines and cause bystander T cell activation, which enables T cells to bypass some control checkpoints
Cation	A positively charged ion
Cell-mediated immunity	Part of the immune system which does not involve antibodies; it can include the activation of phagocytes, release of cytokines and activation of antigen-specific cytotoxic T cells
Chemokine	A small signalling molecule secreted by cells that function to attract immune cells to an infection site
Chromatin	A combination of protein, RNA and DNA which together make up the chromosomes of an organism
Complement system	A biochemical cascade that helps antibodies to remove or mark pathogens
Cortical tissue	The outer layer of the cerebrum in the brain

Cytokines	A large group of substances which are secreted by immune cells to induce an effect in another cell; includes interleukin, interferons and growth factors
Cytotoxic T cells	Also known as CD8+ T cells, these cells function to kill damaged or infected cells, as well as cancer cells; they express T cell receptors which can recognise antigens and produce an immune response
Dendritic cells	Antigen presenting cells – they process antigen material and present it on their surface to attract T cells
Differentiation	The process by which a cell develops to become more specialised and perform a specific function
DNA methylation	The addition of methyl groups to a DNA molecule to change the activity of that DNA, for example to repress gene transcription
Enzyme	A protein that acts as a biological catalyst
Experimental autoimmune encephalomyelitis	An animal model for brain inflammation that is very similar to Multiple Sclerosis; it is an inflammatory demyelinating disease of the CNS
Helper T cell	Also known as CD4+ T cells, they help activate B cells, macrophages and cytotoxic T cells
Hematopoietic stem cells	Stem cells that give rise to blood cells; occurs in the bone marrow
Humoral immunity	Part of the immune system that is mediated by macromolecules found in extracellular fluids, for example secreted antibodies
IgG antibodies	The most common type of antibody in blood circulation, it is produced and released by plasma B cells
Immunoglobulin	An antibody that attaches to the B cell membrane

Innate immune system	Non-specific defence mechanisms within the immune system to remove pathogens and toxic substances, including physical barriers (for example, the skin), inflammation and white blood cell activity
Interferons	Signalling proteins that are produced in response to viral infection
Interleukin-1 family (IL-1)	A group of cytokines that help regulate immune and inflammatory responses
Leukocyte	Also known as a white blood cell, leukocytes circulate primarily in the blood and help counteract toxic substances, pathogens and disease
Lymphocytes	A type of white blood cell that is essential for immune responses; the two main types of lymphocytes are B cells and T cells
Lymphoid organ	Organs or parts of tissues in which lymphocytes can differentiate and proliferate
Macrophages	A large white blood cell which can phagocytose or 'eat' particles, for example bacteria and viruses
Metabolites	A molecule essential to or formed in the metabolic process, for example amino acids
MHC class I molecules	Found on the surface of all nucleated cells; act to display peptide fragments to cytotoxic T cells
MHC class II molecules	Found on the surface of professional antigen presenting cells such as dendritic cells; act to display antigens to initiate an immune response
Microglia	Non-neuronal cells that function as macrophages in the CNS
Monoclonal	A monoclonal cell is produced from a single ancestral cell – it is a clone of the original cell

Monocytes	A type of white blood cell that can help destroy pathogens and can differentiate into macrophages and dendritic cells
Myelin sheath	An insulating cover that wraps around an axon, it helps increase the speed of nerve impulses
Naive T cells	A mature T cell that has not yet encountered an antigen
Natural killer (NK) cells	A type of white blood cell and part of the innate immune system, NK cells reject both virally infected cells and tumour cells; similar to cytotoxic T cells, except NK cells cannot recognise antigens
Oligoclonal bands	Bands of immunoglobulins that can be seen when an individual's cerebrospinal fluid or blood serum is analysed; can be used to diagnose disease such as MS
Oligodendrocytes	A non-neuronal cell that myelinates CNS axons
Peptide	A molecule containing two or more amino acids, they are shorter in amino acid chain length than proteins
Phagocytes	A type of cell that can engulf and absorb pathogens and other particles
Plasma B cells	Also called effector B cells, these B cells secrete antibodies in response to antigen presentation
Post-translational histone modification	A method of regulating gene expression; the genome is organised into active euchromatin (where DNA can be transcribed) or inactive heterochromatin (where DNA is less accessible for transcription)
Proinflammatory	The opposite of anti-inflammatory; the act of increasing inflammation and making a disease worse
Proliferation	The rapid increase in number of a substance, such as a cell

Regulatory T cells	A subgroup of T cells that help regulate the immune system, prevent autoimmune disease, and maintain tolerance to self-antigens
Remyelination	The process of forming oligodendrocytes to create new myelin sheaths around demyelinated axons within the CNS
T cells	A lymphocyte which possesses specific cell-surface antigen receptors, they function to recognise foreign tissues and infected cells and directs the immune system in response; produced in the thymus
Type I interferons	A subgroup of interferons that help regulate the immune system

APPENDIX

RSID	SNPID	CHR	Position	A1	A0	Freq (A1)	Beta (A1)	SE	P	Info
.	1_145044288	1	145044288	C	T	0.09	0.15	0.03	4.44×10^{-8}	0.50
rs34584371	2_153037808	2	153037808	A	G	0.12	0.08	0.02	4.01×10^{-6}	0.99
rs72932144	2_177553322	2	177553322	T	C	0.32	0.06	0.01	5.30×10^{-7}	1.00
rs11691194	2_177553829	2	177553829	A	G	0.32	0.06	0.01	4.06×10^{-7}	1.00
rs955972	2_177555937	2	177555937	A	G	0.34	0.05	0.01	1.48×10^{-6}	1.00
rs955973	2_177556099	2	177556099	G	A	0.42	0.05	0.01	6.74×10^{-7}	0.99
rs17786781	2_177556469	2	177556469	G	A	0.32	0.06	0.01	4.10×10^{-7}	1.00
rs2885628	2_177561074	2	177561074	T	A	0.33	0.05	0.01	1.43×10^{-6}	1.00
rs10207615	2_177561365	2	177561365	A	C	0.32	0.06	0.01	3.13×10^{-7}	1.00
rs13413215	2_177563646	2	177563646	T	A	0.32	0.06	0.01	3.04×10^{-7}	1.00
rs72929573	2_177568452	2	177568452	A	G	0.32	0.06	0.01	1.67×10^{-7}	0.99
rs1398972	2_177569251	2	177569251	C	G	0.32	0.06	0.01	1.49×10^{-7}	0.99
rs72929579	2_177570137	2	177570137	A	C	0.32	0.06	0.01	1.51×10^{-7}	0.99
rs6433609	2_177572596	2	177572596	T	C	0.32	0.06	0.01	1.85×10^{-7}	0.99
rs113441701	2_177572715	2	177572715	A	G	0.32	0.06	0.01	1.80×10^{-7}	0.99
rs72929586	2_177573625	2	177573625	T	C	0.32	0.06	0.01	1.50×10^{-7}	0.99
rs7635898	3_55408967	3	55408967	C	A	0.09	0.09	0.02	3.87×10^{-6}	0.84
rs816545	3_156092170	3	156092170	T	C	0.05	0.11	0.02	5.42×10^{-6}	0.93
rs73221623	3_196370840	3	196370840	T	C	0.10	0.08	0.02	4.43×10^{-6}	0.97
.	6_32237926	6	32237926	C	T	0.78	-0.06	0.01	9.60×10^{-6}	1.00
.	6_32245370	6	32245370	G	A	0.78	-0.06	0.01	9.65×10^{-6}	1.00
.	6_32259527	6	32259527	A	G	0.78	-0.06	0.01	9.75×10^{-6}	1.00
rs9268154	6_32266021	6	32266021	A	T	0.78	-0.06	0.01	9.65×10^{-6}	1.00
rs9268155	6_32266310	6	32266310	C	T	0.78	-0.06	0.01	9.68×10^{-6}	1.00
.	6_32279938	6	32279938	G	A	0.78	-0.06	0.01	9.50×10^{-6}	1.00
.	6_32289390	6	32289390	C	A	0.78	-0.06	0.01	9.34×10^{-6}	1.00
.	6_32300809	6	32300809	G	A	0.78	-0.06	0.01	6.95×10^{-6}	1.00
.	6_32305979	6	32305979	T	G	0.78	-0.06	0.01	6.33×10^{-6}	1.00
.	6_32316016	6	32316016	G	T	0.78	-0.06	0.01	6.93×10^{-6}	1.00
.	6_32318610	6	32318610	A	G	0.78	-0.06	0.01	7.37×10^{-6}	1.00
.	6_32320153	6	32320153	A	G	0.78	-0.06	0.01	7.02×10^{-6}	1.00
.	6_32336187	6	32336187	C	T	0.78	-0.06	0.01	6.83×10^{-6}	1.00
.	6_32336495	6	32336495	A	T	0.78	-0.06	0.01	7.18×10^{-6}	1.00

.	6_32392906	6	32392906	A	C	0.79	-0.06	0.01	9.69×10^{-6}	1.00
.	6_32393235	6	32393235	C	G	0.79	-0.06	0.01	9.71×10^{-6}	1.00
.	6_32397309	6	32397309	G	A	0.74	-0.06	0.01	6.32×10^{-6}	1.00
.	6_32397784	6	32397784	A	G	0.74	-0.06	0.01	6.27×10^{-6}	1.00
.	6_32398748	6	32398748	A	C	0.74	-0.06	0.01	6.29×10^{-6}	1.00
.	6_32399159	6	32399159	C	T	0.74	-0.06	0.01	6.27×10^{-6}	1.00
.	6_32402686	6	32402686	T	C	0.74	-0.06	0.01	6.31×10^{-6}	1.00
.	6_32406473	6	32406473	G	A	0.74	-0.06	0.01	7.46×10^{-6}	1.00
.	6_32406704	6	32406704	G	T	0.74	-0.06	0.01	7.32×10^{-6}	1.00
.	6_32409058	6	32409058	C	T	0.80	-0.06	0.01	5.91×10^{-6}	0.99
.	6_32411726	6	32411726	A	G	0.75	-0.06	0.01	3.89×10^{-6}	1.00
.	6_32412580	6	32412580	C	T	0.75	-0.06	0.01	5.90×10^{-6}	1.00
.	6_32415109	6	32415109	G	T	0.75	-0.06	0.01	4.11×10^{-6}	1.00
rs12209200	6_57217336	6	57217336	T	C	0.08	0.10	0.02	1.03×10^{-6}	0.84
rs4720446	7_4982335	7	4982335	C	G	0.88	-0.07	0.02	7.56×10^{-6}	0.98
rs10965046	9_21518275	9	21518275	G	A	0.14	0.07	0.02	7.92×10^{-6}	0.98
rs76585251	9_24750193	9	24750193	T	G	0.07	0.10	0.02	2.20×10^{-6}	0.95
rs78870428	10_25662804	10	25662804	G	A	0.10	0.09	0.02	6.60×10^{-6}	0.86
rs117374511	10_134496477	10	134496477	T	C	0.06	0.10	0.02	6.97×10^{-6}	0.91
rs2268134	12_13997305	12	13997305	G	T	0.09	0.08	0.02	7.22×10^{-6}	1.00
rs11055642	12_13999533	12	13999533	C	T	0.09	0.08	0.02	7.18×10^{-6}	1.00
rs4280084	12_13999655	12	13999655	C	T	0.09	0.08	0.02	7.03×10^{-6}	1.00
rs1861787	12_14000568	12	14000568	T	G	0.09	0.09	0.02	6.07×10^{-6}	1.00
rs1861788	12_14000592	12	14000592	G	A	0.13	0.08	0.02	1.45×10^{-6}	1.00
rs71459105	12_14001448	12	14001448	T	C	0.09	0.08	0.02	6.11×10^{-6}	1.00
rs71459107	12_14003283	12	14003283	A	G	0.13	0.07	0.02	5.50×10^{-6}	1.00
rs11055646	12_14005719	12	14005719	C	T	0.13	0.08	0.02	5.62×10^{-7}	1.00
rs2268138	12_14007212	12	14007212	C	A	0.09	0.09	0.02	2.40×10^{-6}	1.00
rs7297313	12_14008506	12	14008506	C	A	0.13	0.08	0.02	5.83×10^{-7}	1.00
rs11055654	12_14022098	12	14022098	A	G	0.09	0.08	0.02	7.69×10^{-6}	1.00
rs11055660	12_14029371	12	14029371	G	A	0.14	0.07	0.02	5.45×10^{-6}	0.99
rs73162646	12_130464398	12	130464398	C	T	0.13	0.08	0.02	1.52×10^{-6}	0.99
rs8008067	14_72941207	14	72941207	A	G	0.07	0.10	0.02	2.39×10^{-6}	0.88
rs8041424	15_97359414	15	97359414	T	C	0.06	0.11	0.02	1.86×10^{-6}	0.95
rs1604686	15_97385210	15	97385210	G	A	0.09	0.09	0.02	6.43×10^{-6}	0.95
rs7199663	16_86608899	16	86608899	C	A	0.16	0.07	0.01	1.90×10^{-6}	0.95
rs62048038	16_88727961	16	88727961	A	T	0.08	0.09	0.02	4.79×10^{-6}	0.92
.	17_70738077	17	70738077	A	G	0.75	-0.06	0.01	6.43×10^{-6}	0.99
rs62092559	18_19975654	18	19975654	G	C	0.06	0.11	0.02	1.47×10^{-6}	0.95
rs112519464	18_19977947	18	19977947	A	T	0.05	0.11	0.02	2.18×10^{-6}	0.97
rs34069471	18_19982454	18	19982454	C	A	0.05	0.12	0.02	4.70×10^{-7}	0.97
rs62094163	18_19984094	18	19984094	T	C	0.05	0.12	0.02	9.83×10^{-7}	0.97
rs73401039	18_19988535	18	19988535	T	C	0.05	0.11	0.02	3.08×10^{-6}	0.99
rs73401040	18_19989378	18	19989378	T	C	0.05	0.11	0.02	4.06×10^{-6}	0.99
rs1893251	18_19991432	18	19991432	T	C	0.05	0.12	0.02	2.92×10^{-7}	0.98
rs17602961	18_36627629	18	36627629	A	C	0.05	0.11	0.02	5.27×10^{-6}	0.99
rs1396651	18_36646561	18	36646561	A	G	0.06	0.11	0.02	5.77×10^{-6}	0.99
rs1509215	18_36646764	18	36646764	T	C	0.05	0.11	0.02	5.30×10^{-6}	0.99
rs77249263	18_36663604	18	36663604	G	A	0.06	0.11	0.02	5.94×10^{-6}	0.99
rs62096323	18_56304224	18	56304224	G	C	0.10	0.08	0.02	4.15×10^{-6}	0.97
rs55704994	18_56307599	18	56307599	T	G	0.10	0.08	0.02	4.79×10^{-6}	0.97
rs10221425	18_56317005	18	56317005	T	C	0.14	0.07	0.02	5.65×10^{-6}	0.98
rs62094974	18_56318728	18	56318728	A	G	0.14	0.07	0.02	5.48×10^{-6}	0.98

rs11872221	18_56319582	18	56319582	T	A	0.14	0.07	0.02	4.88×10^{-6}	0.98
rs75479243	19_2739091	19	2739091	T	C	0.08	0.10	0.02	9.17×10^{-6}	0.85
rs28752178	22_36833819	22	36833819	C	T	0.11	0.08	0.02	2.62×10^{-6}	0.99

Supplementary Table 1: Results from GWAS in MS in ORCADES/VIKING dataset, where SNP p-value < 1×10^{-5}

Results from a GWAS of MS on the ORCADES/VIKING dataset (cases: 112, controls: 4223). Only SNPs which have passed the suggestive threshold of 1×10^{-5} are included. However, it should be noted that only one of these SNPs passed the genome-wide significance threshold (SNPID chr1_145044288), and so beta values should be read with caution. The RSID is the Reference SNP cluster ID. SNPID is the chromosome number and base pair number, which is used to identify a SNP if an RSID has not been assigned, for example. Chr is the chromosome the SNP is located on. Position is the base pair position the SNP can be found at. A1 is the reference allele, and A0 is the non-reference allele. Freq is the frequency of the A1 allele within the dataset. Beta is the effect size estimate. SE is the standard error of the effect size estimate. The p-value is the calculated probability or level of significance of the effect estimate. The Info measure describes the information within the imputed genotypes relative to the information that would be presented if only the allele frequencies were known – for example, an info measure of 1 would denote that all genotypes are completely certain, whereas an info score of 0 would denote that the genotype probabilities are completely uncertain (Purcell et al., 2007).

			Controls					Cases				T-Test	
pT	Pop	N SNPs	N Inds	Mean (95% CI)	SD	Min	Max	N Inds	Mean (95% CI)	SD	Min	Max	Statistic	p-value
A	GS	127	8708	7.20 (7.19 - 7.21)	0.60	4.88	9.58	30	7.60 (7.39 - 7.81)	0.56	6.31	8.75	-3.86	5.74×10^{-4}
	OR	127	2118	7.24 (7.21 - 7.27)	0.61	5.19	9.29	94	7.63 (7.51 - 7.75)	0.58	5.82	9.31	-6.35	5.86×10^{-9}
	VIK	127	2156	7.26 (7.23 - 7.28)	0.62	5.30	9.56	16	7.66 (7.42 - 7.89)	0.45	6.58	8.52	-3.51	3.05×10^{-3}
B	GS	126	8708	7.17 (7.16 - 7.18)	0.60	4.84	9.50	30	7.57 (7.36 - 7.77)	0.56	6.31	8.75	-3.88	5.42×10^{-4}
	OR	126	2118	7.21 (7.19 - 7.24)	0.61	5.15	9.28	94	7.60 (7.49 - 7.72)	0.57	5.82	9.28	-6.43	4.07×10^{-9}
	VIK	126	2156	7.23 (7.20 - 7.26)	0.62	5.26	9.52	16	7.62 (7.39 - 7.86)	0.44	6.58	8.48	-3.53	2.93×10^{-3}
C	GS	101	8708	6.29 (6.28 - 6.30)	0.54	4.39	8.55	30	6.69 (6.50 - 6.87)	0.50	5.20	7.54	-4.35	1.52×10^{-4}
	OR	101	2118	6.36 (6.33 - 6.38)	0.55	4.40	8.34	94	6.70 (6.60 - 6.80)	0.48	5.34	7.79	-6.66	1.30×10^{-9}
	VIK	101	2156	6.33 (6.31 - 6.36)	0.55	4.76	8.12	16	6.69 (6.40 - 6.98)	0.55	5.59	7.56	-2.57	2.13×10^{-2}

D	GS	61	8708	4.32 (4.31 - 4.33)	0.44	2.90	6.05	30	4.63 (4.46 - 4.80)	0.45	3.65	5.60	-3.82	6.54×10^{-4}
	OR	61	2118	4.38 (4.36 - 4.40)	0.46	3.11	5.89	94	4.66 (4.56 - 4.76)	0.47	3.61	5.62	-5.62	1.67×10^{-7}
	VIK	61	2156	4.37 (4.35 - 4.39)	0.47	3.18	5.93	16	4.67 (4.43 - 4.91)	0.46	3.97	5.28	-2.61	1.96×10^{-2}

Supplementary Table 2: Polygenic risk score p-value threshold group descriptive statistics

Descriptive statistics for the polygenic risk scores at the four different p-value threshold (pT) groups: pT A = 0.05; pT B = 0.0005; pT C = 0.000005; pT D = 0.00000005, for the three population groups (pop = population; OR = ORCADES; VIK = VIKING; GS = Generation Scotland). Two-sample t-test statistics for comparing the mean values of cases and controls is also included.

							RAF				GS/ORCADES		GS/VIKING
RSID	CHR	BP	RA	OR	OR p-Value	GS	ORC	VIK	χ2	p-value	χ2	p-value
rs4648356	1	2709164	C	1.16	1.00×10^{-14}	0.67	0.68	0.70	2.02	0.16	16.49	4.89×10^{-5}
rs912961	1	10356848	G	1.34	3.00×10^{-10}	0.33	0.32	0.35	4.04	0.04	6.22	0.01
rs233100	1	85772009	G	1.09	1.00×10^{-6}	0.56	0.56	0.55	0.04	0.84	2.08	0.15
rs11810217	1	93148377	A	1.15	5.80×10^{-15}	0.27	0.26	0.24	1.86	0.17	15.37	**8.86×10^{-5}**
rs12048904	1	101331536	A	1.1	4.00×10^{-8}	0.38	0.38	0.37	0.03	0.86	1.38	0.24
rs11581062	1	101407519	G	1.13	2.50×10^{-10}	0.30	0.28	0.30	8.26	4.06×10^{-3}	0.19	0.66
rs1335532	1	117100957	A	1.2	6.86×10^{-22}	0.86	0.86	0.88	0.00	0.99	6.01	0.01
rs3761959	1	157669278	G	1.1	2.90×10^{-6}	0.53	0.53	0.53	0.24	0.62	0.16	0.69
rs1323292	1	192541021	A	1.12	2.30×10^{-8}	0.82	0.81	0.79	4.20	0.04	32.43	**1.23×10^{-8}**
rs7522462	1	200881595	G	1.11	1.90×10^{-9}	0.70	0.68	0.66	8.44	3.66×10^{-3}	25.42	**4.61×10^{-7}**
rs6718520	2	43325570	A	1.17	3.00×10^{-8}	0.46	0.40	0.47	55.88	**7.72×10^{-14}**	2.27	0.13
rs12466022	2	43359061	C	1.1	6.20×10^{-10}	0.72	0.71	0.74	2.53	0.11	5.66	0.02
rs7592560	2	68647001	A	1.1	5.10×10^{-11}	0.54	0.58	0.57	15.87	6.79×10^{-5}	7.34	0.01
rs17174870	2	112665201	G	1.1	1.30×10^{-8}	0.74	0.74	0.71	0.00	0.98	24.13	**9.00×10^{-7}**
rs882300	2	136976255	C	1.19	1.00×10^{-7}	0.53	0.51	0.50	4.77	0.03	17.44	**2.97×10^{-5}**
rs281783	2	200751582	A	1.11	0.00018	0.80	0.81	0.80	4.81	0.03	0.14	0.71
rs10201872	2	231106724	A	1.13	1.80×10^{-10}	0.17	0.16	0.14	3.57	0.06	21.77	**3.08×10^{-6}**
rs9821630	3	16970938	G	1.09	3.90×10^{-6}	0.28	0.30	0.25	9.75	1.79×10^{-3}	14.32	**1.55×10^{-4}**
rs11129295	3	27788780	T	1.11	1.20×10^{-9}	0.35	0.42	0.37	85.49	2.33×10^{-20}	5.35	0.02
rs669607	3	28071444	C	1.13	1.90×10^{-15}	0.47	0.43	0.43	29.77	**4.86×10^{-8}**	24.60	**7.04×10^{-7}**
rs1500710	3	56914065	A	1.09	5.20×10^{-5}	0.58	0.53	0.58	45.36	**1.64×10^{-11}**	0.01	0.92
rs771767	3	101748638	A	1.12	8.60×10^{-9}	0.26	0.26	0.26	0.18	0.67	0.80	0.37
rs9657904	3	105586714	T	1.4	2.00×10^{-10}	0.79	0.75	0.75	31.24	**2.28×10^{-8}**	33.85	**5.95×10^{-9}**
rs2293370	3	119219934	G	1.16	2.70×10^{-9}	0.81	0.79	0.82	10.51	1.19×10^{-3}	5.78	0.02
rs4285028	3	121660664	A	1.11	1.80×10^{-8}	0.74	0.73	0.75	1.22	0.27	2.88	0.09
rs4308217	3	121793187	C	1.1	5.70×10^{-8}	0.68	0.65	0.65	8.45	3.65×10^{-3}	7.56	0.01
rs9282641	3	121796768	G	1.2	1.00×10^{-11}	0.92	0.92	0.91	0.35	0.56	5.06	0.02
rs4680534	3	159698945	C	1.12	6.00×10^{-6}	0.34	0.34	0.35	0.00	0.97	0.62	0.43
rs2243123	3	159709651	C	1.09	7.20×10^{-6}	0.28	0.27	0.30	5.73	0.02	5.75	0.02
rs10936599	3	169492101	C	1.1	7.00×10^{-7}	0.76	0.74	0.75	8.93	2.80×10^{-3}	0.52	0.47
rs228614	4	103578637	G	1.09	1.40×10^{-7}	0.54	0.59	0.54	29.15	6.71×10^{-8}	0.21	0.65
rs6821894	4	186571441	T	1.08	9.20×10^{-5}	0.62	0.59	0.60	17.36	**3.09×10^{-5}**	7.22	0.01
rs6897932	5	35874575	C	1.12	1.08×10^{-18}	0.72	0.69	0.74	12.80	3.47×10^{-4}	9.75	1.80×10^{-3}
rs350058	5	40211802	A	1.14	1.00×10^{-4}	0.08	0.09	0.07	11.22	8.10×10^{-4}	3.68	0.06

rs4613763	5	40392728	C	1.21	2.50×10^{-16}	0.13	0.15	0.14	12.13	4.95×10^{-f04}	9.21	2.41×10^{-3}
rs9292777	5	40437948	T	1.19	1.00×10^{-9}	0.59	0.58	0.59	3.49	0.06	0.17	0.68
rs756699	5	133446575	A	1.12	6.20×10^{-7}	0.88	0.88	0.88	0.60	0.44	0.02	0.88
rs6879677	5	140954954	A	1.08	0.00054	0.40	0.36	0.37	20.80	**5.11×10^{-6}**	8.46	3.64×10^{-3}
rs1062158	5	141523000	A	1.09	2.30×10^{-6}	0.62	0.59	0.65	13.43	2.48×10^{-4}	14.86	**1.16×10^{-4}**
rs2546890	5	158759900	A	1.11	2.33×10^{-17}	0.51	0.53	0.54	1.92	0.17	10.33	1.31×10^{-3}
rs10866713	5	158918894	A	1.17	7.00×10^{-7}	0.21	0.20	0.18	5.18	0.02	22.10	**2.59×10^{-6}**
rs4075958	5	176784512	A	1.11	4.90×10^{-7}	0.25	0.25	0.27	0.00	0.99	6.57	0.01
rs11755724	6	7118990	A	1.08	2.60×10^{-6}	0.36	0.34	0.35	8.24	4.10×10^{-3}	0.64	0.43
rs9260119	6	29910189	A	1.21	1.00×10^{-11}	0.45	0.42	0.43	13.86	**1.97×10^{-4}**	7.00	0.01
rs9271069	6	32575700	A	2.77	$< 1 \times 10^{-50}$	0.17	0.23	0.21	103.00	**3.36×10^{-24}**	44.93	**2.05×10^{-11}**
rs854917	6	90127390	A	1.09	0.00017	0.73	0.75	0.73	8.27	4.03×10^{-3}	0.48	0.49
rs12212193	6	90996769	G	1.09	3.80×10^{-8}	0.45	0.53	0.47	97.45	**5.53×10^{-23}**	3.88	0.05
rs11962089	6	105612220	G	0.69	8.00×10^{-6}	0.11	0.11	0.08	0.60	0.44	38.82	**4.65×10^{-10}**
rs802734	6	128278798	A	1.1	5.50×10^{-9}	0.67	0.65	0.63	8.20	4.19×10^{-3}	29.00	**7.24×10^{-8}**
rs9399141	6	135495574	C	1.12	1.60×10^{-6}	0.23	0.22	0.26	2.16	0.14	12.56	3.94×10^{-4}
rs11154801	6	135739355	A	1.15	1.00×10^{-13}	0.36	0.36	0.36	0.02	0.89	0.32	0.57
rs17066096	6	137452908	G	1.14	6.00×10^{-13}	0.24	0.28	0.28	37.87	**7.58×10^{-10}**	28.79	**8.07×10^{-8}**
rs13192841	6	137967214	A	1.1	1.30×10^{-8}	0.29	0.29	0.28	1.01	0.32	2.87	0.09
rs1738074	6	159465977	C	1.14	1.56×10^{-20}	0.55	0.53	0.54	4.46	0.03	0.85	0.36
rs6952809	7	2448493	A	1.08	3.60×10^{-6}	0.31	0.30	0.30	2.63	0.10	2.32	0.13
rs1843938	7	3113034	A	1.09	1.10×10^{-5}	0.44	0.46	0.42	2.81	0.09	9.81	1.73×10^{-3}
rs2214543	7	10796892	G	1.09	0.00016	0.74	0.75	0.73	4.04	0.04	0.07	0.79
rs2066992	7	22768249	C	1.18	6.30×10^{-5}	0.95	0.97	0.95	31.44	**2.05×10^{-8}**	0.02	0.89
rs11984075	7	37436854	G	1.13	1.10×10^{-5}	0.11	0.10	0.11	4.82	0.03	0.42	0.52
rs7789940	7	75951230	G	1.87	6.00×10^{-6}	0.30	0.27	0.30	12.13	4.96×10^{-4}	0.00	0.99
rs354033	7	149289464	G	1.1	4.70×10^{-9}	0.74	0.78	0.79	37.44	**9.43×10^{-10}**	57.01	**4.33×10^{-14}**
rs6986386	8	9421789	A	1.09	1.60×10^{-5}	0.23	0.24	0.23	0.88	0.35	0.44	0.51
rs1520333	8	79401038	G	1.11	1.60×10^{-7}	0.24	0.23	0.25	1.48	0.22	4.05	0.04
rs4410871	8	128815029	G	1.11	7.70×10^{-9}	0.70	0.70	0.72	0.05	0.82	9.37	2.21×10^{-3}
rs2019960	8	129192271	C	1.1	5.20×10^{-9}	0.23	0.23	0.20	0.20	0.65	14.06	**1.77×10^{-4}**
rs6984045	8	131092413	C	1.59	2.00×10^{-6}	0.03	0.04	0.03	6.89	0.01	2.34	0.13
rs2150702	9	5893861	G	1.16	3.00×10^{-8}	0.49	0.54	0.53	30.07	**4.18×10^{-8}**	20.87	**4.92×10^{-6}**
rs290986	9	93563536	A	1.12	9.10×10^{-7}	0.80	0.80	0.78	0.00	1.00	10.63	1.11×10^{-3}
rs10984447	9	121984553	A	1.17	8.00×10^{-6}	0.74	0.74	0.74	0.01	0.91	0.06	0.80
rs3780792	9	136835343	G	1.6	1.00×10^{-6}	0.33	0.34	0.36	2.09	0.15	15.70	**7.43×10^{-5}**
rs12722489	10	6102012	C	1.24	6.41×10^{-15}	0.84	0.82	0.80	3.63	0.06	41.40	**1.24×10^{-10}**
rs7090512	10	6110829	C	1.19	4.60×10^{-20}	0.28	0.30	0.29	8.62	3.32×10^{-3}	5.07	0.02
rs793108	10	31415106	A	1.09	2.60×10^{-6}	0.49	0.54	0.45	44.97	**2.00×10^{-11}**	14.97	**1.09×10^{-4}**
rs2503875	10	43814049	A	1.66	2.00×10^{-7}	0.29	0.29	0.27	0.84	0.36	2.21	0.14
rs7912269	10	78727604	A	1.16	1.40×10^{-5}	0.94	0.94	0.96	3.14	0.08	41.92	**9.53×10^{-11}**
rs1250550	10	81060317	A	1.1	6.30×10^{-9}	0.36	0.36	0.36	0.31	0.58	0.60	0.44
rs7923837	10	94481917	G	1.1	4.90×10^{-9}	0.62	0.66	0.64	26.39	**2.80×10^{-7}**	6.46	0.01
rs17824933	11	60760612	G	1.18	4.00×10^{-9}	0.23	0.22	0.24	4.36	0.04	2.01	0.16
rs650258	11	60832282	C	1.12	2.00×10^{-11}	0.64	0.66	0.61	5.21	0.02	17.21	**3.35×10^{-5}**
rs694739	11	64097233	A	1.08	0.00014	0.62	0.62	0.56	0.10	0.75	51.29	**7.96×10^{-13}**
rs4409785	11	95311422	G	1.11	6.30×10^{-7}	0.19	0.22	0.14	29.83	**4.72×10^{-8}**	55.00	**1.20×10^{-13}**
rs491111	11	116238034	G	1.08	0.00048	0.65	0.65	0.65	0.43	0.51	0.12	0.73
rs630923	11	118754353	C	1.11	2.80×10^{-7}	0.84	0.83	0.85	4.78	0.03	1.74	0.19
rs7941030	11	122522375	C	1.09	1.60×10^{-5}	0.39	0.38	0.41	1.94	0.16	6.92	0.01
rs1800693	12	6440009	C	1.14	6.42×10^{-23}	0.40	0.38	0.38	11.12	8.54×10^{-4}	8.78	3.04×10^{-3}
rs4149584	12	6442643	T	1.58	5.00×10^{-6}	0.02	0.02	0.02	4.71	0.03	0.00	1.00
rs10466829	12	9876091	A	1.09	1.40×10^{-8}	0.50	0.46	0.51	27.72	**1.40×10^{-7}**	1.28	0.26
rs703842	12	58162739	A	1.23	5.00×10^{-11}	0.66	0.68	0.65	6.67	0.01	1.15	0.28
rs1790100	12	123656725	G	1.11	7.00×10^{-7}	0.21	0.20	0.21	1.03	0.31	0.00	0.95
rs17594362	13	42139245	T	1.12	3.70×10^{-6}	0.12	0.12	0.09	0.74	0.39	25.18	**5.23×10^{-7}**
rs806321	13	50841323	T	1.09	5.00×10^{-7}	0.53	0.55	0.56	6.53	0.01	10.45	1.23×10^{-3}
rs9596270	13	50842440	T	1.35	7.00×10^{-7}	0.93	0.90	0.92	64.84	**8.14×10^{-16}**	9.73	1.81×10^{-3}
rs4902647	14	69254191	C	1.11	9.30×10^{-12}	0.53	0.54	0.54	2.04	0.15	0.74	0.39
rs2300603	14	76005557	T	1.11	2.00×10^{-8}	0.76	0.73	0.71	20.94	**4.75×10^{-6}**	48.62	**3.11×10^{-12}**
rs2119704	14	88487689	C	1.27	2.20×10^{-10}	0.93	0.93	0.93	0.14	0.71	0.27	0.60
rs449295	16	1074443	A	1.12	8.40×10^{-8}	0.17	0.19	0.16	7.42	0.01	1.80	0.18

rs7200786	16	11177801	A	1.15	8.50×10^{-17}	0.44	0.44	0.45	0.02	0.89	2.08	0.15
rs7191700	16	11406803	C	1.15	6.00×10^{-7}	0.66	0.68	0.67	3.97	0.05	1.23	0.27
rs11864333	16	11475576	A	1.09	3.60×10^{-5}	0.50	0.54	0.53	24.88	6.11×10^{-7}	14.62	1.31×10^{-4}
rs8049603	16	23067260	T	1.19	1.00×10^{-6}	0.23	0.21	0.23	9.02	2.68×10^{-3}	0.00	0.98
rs386965	16	79652541	G	1.09	3.90×10^{-6}	0.21	0.21	0.27	0.14	0.70	74.54	5.93×10^{-18}
rs13333054	16	86011033	A	1.12	1.30×10^{-8}	0.24	0.25	0.26	1.03	0.31	7.01	0.01
rs17445836	16	86017663	G	1.25	4.00×10^{-9}	0.77	0.76	0.79	1.13	0.29	15.59	7.88×10^{-5}
rs2293152	17	40481529	C	1.22	4.00×10^{-8}	0.60	0.60	0.62	0.05	0.82	10.42	1.25×10^{-3}
rs9891119	17	40507980	C	1.1	1.80×10^{-10}	0.35	0.33	0.39	5.92	0.01	26.49	2.65×10^{-7}
rs1373089	17	44915265	A	1.08	4.00×10^{-5}	0.49	0.54	0.52	39.19	3.84×10^{-10}	15.24	9.49×10^{-5}
rs8070463	17	45768836	T	1.15	1.00×10^{-7}	0.46	0.48	0.45	4.29	0.04	3.07	0.08
rs180515	17	58024275	G	1.11	8.80×10^{-8}	0.36	0.36	0.36	1.23	0.27	0.01	0.94
rs8081176	17	78283987	C	1.09	1.50×10^{-5}	0.31	0.33	0.34	6.86	0.01	15.75	7.22×10^{-5}
rs12456021	18	56213390	A	1.1	3.60×10^{-6}	0.19	0.23	0.18	34.31	4.70×10^{-9}	1.83	0.18
rs7238078	18	56384192	T	1.11	2.50×10^{-9}	0.77	0.83	0.76	86.01	1.79×10^{-20}	0.76	0.38
rs1077667	19	6668972	C	1.16	9.40×10^{-14}	0.78	0.82	0.80	25.78	3.82×10^{-7}	8.47	3.62×10^{-3}
rs2278442	19	10444826	A	1.08	0.00012	0.65	0.62	0.69	15.81	**7.01×10^{-5}**	33.81	6.09×10^{-9}
rs8112449	19	10520064	G	1.1	1.20×10^{-6}	0.67	0.70	0.65	14.87	1.15×10^{-4}	4.92	0.03
rs10411936	19	16548375	A	1.16	2.00×10^{-7}	0.30	0.32	0.34	6.85	0.01	24.55	7.23×10^{-7}
rs874628	19	18304700	A	1.12	1.30×10^{-8}	0.68	0.71	0.73	16.92	3.91×10^{-5}	46.48	9.26×10^{-12}
rs7255066	19	45146103	C	1.1	1.20×10^{-6}	0.25	0.22	0.24	20.40	**6.29×10^{-6}**	2.58	0.11
rs6509314	19	47696626	T	1.1	4.60×10^{-7}	0.70	0.70	0.68	0.21	0.65	6.92	0.01
rs281380	19	49214470	G	1.08	1.90×10^{-6}	0.33	0.39	0.32	74.35	6.55×10^{-18}	1.87	0.17
rs2303759	19	49869051	G	1.11	5.20×10^{-9}	0.24	0.23	0.23	1.40	0.24	0.50	0.48
rs6074022	20	44740196	C	1.17	3.56×10^{-12}	0.25	0.27	0.26	8.05	4.54×10^{-3}	0.26	0.61
rs2762932	20	52768391	G	1.15	8.10×10^{-7}	0.14	0.15	0.17	1.73	0.19	39.42	3.41×10^{-10}
rs2248359	20	52791518	C	1.12	2.50×10^{-11}	0.59	0.65	0.58	51.56	6.96×10^{-13}	1.05	0.30
rs6062314	20	62409713	T	1.18	1.30×10^{-7}	0.92	0.91	0.93	1.84	0.18	8.58	3.39×10^{-3}
rs2283792	22	22131125	C	1.09	4.70×10^{-9}	0.51	0.52	0.51	0.74	0.39	0.01	0.93
rs2072711	22	37268555	A	1.12	6.30×10^{-5}	0.17	0.15	0.16	11.16	8.34×10^{-4}	4.78	0.03
rs140522	22	50971266	T	1.09	1.70×10^{-8}	0.32	0.33	0.29	1.08	0.30	13.72	2.12×10^{-4}

Supplementary Table 3: Risk allele frequencies in mainland Scotland, Orkney and Shetland

Risk allele frequencies (RAF) and Pearson's chi-squared test values (with corresponding p-values) for all (n=126) variants included in the PGRS calculation. The chi-squared p-values are shown for two RAF comparisons between population controls: Generation Scotland (GS) and ORCADES (ORC); Generation Scotland and VIKING (VIK). Significant results, corrected for multiple testing at 126 loci over two population comparisons (corrected p-value significance level of $<1.98 \times 10^{-4}$), are shown in bold. Significant p-values from RAF that are higher in Generation Scotland than ORCADES or VIKING are highlighted in dark red, and significant p-values from RAF that are higher in ORCADES or VIKING than Generation Scotland are highlighted in light blue. *NB: RAF values are rounded to 2 decimal places. This can result in values (such as rs4149584) that appear the same due to rounding but give differing p-values between populations.*

					95% CI	
A. pT 0.0005 ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-9.49	1.63	-5.82	5.82×10^{-9}	-12.69	-6.30

PGRS	1.14	0.21	5.49	4.14×10^{-8}	0.74	1.55
Age	-0.01	0.01	-1.65	0.10	-0.03	0.00
Sex	-0.67	0.27	-2.48	0.01	-1.20	-0.14
PC1	-1.81	6.41	-0.28	0.78	-14.37	10.76
PC2	-10.73	11.60	-0.93	0.36	-33.47	12.01

					95% CI	
B. pT 0.0005 VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-14.32	3.68	-3.89	0.00	-21.54	-7.10
PGRS	0.90	0.43	2.09	0.04	0.06	1.75
Age	0.03	0.02	1.65	0.10	-0.01	0.07
Sex	0.86	0.67	1.29	0.20	-0.45	2.16
PC1	1.52	47.27	0.03	0.97	-91.13	94.17
PC2	58.31	32.81	1.78	0.08	-5.99	122.61

					95% CI	
C. pT 0.0005 GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-14.20	2.43	-5.85	5.01×10^{-9}	-18.96	-9.44
PGRS	1.04	0.31	3.40	0.00	0.44	1.64
Age	0.02	0.01	1.75	0.08	0.00	0.05
Sex	-1.20	0.49	-2.43	0.02	-2.16	-0.23
PC1	6.62	21.32	0.31	0.76	-35.17	48.41
PC2	-47.06	23.24	-2.02	0.04	-92.61	-1.50

					95% CI	
D. pT 0.0005 ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-10.13	1.55	-6.55	5.93×10^{-11}	-13.17	-7.10
PGRS	1.10	0.20	5.40	6.61×10^{-8}	0.70	1.50

					95% CI	
E. pT 0.0005 VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-10.70	3.26	-3.29	0.00	-17.09	-4.32
PGRS	0.93	0.43	2.18	0.03	0.09	1.77

					95% CI	
F. pT 0.0005 GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-12.99	2.26	-5.75	9.17×10^{-9}	-17.43	-8.56
PGRS	1.00	0.30	3.34	0.00	0.41	1.58

					95% CI	
G. pT 0.0005 without rs9271069 ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-8.10	1.74	-4.65	3.29×10^{-6}	-11.52	-4.69
PGRS	1.01	0.24	4.24	2.20×10^{-5}	0.54	1.48
Age	-0.01	0.01	-1.79	0.07	-0.03	0.00
Sex	-0.65	0.27	-2.44	0.01	-1.17	-0.13
PC1	-3.74	6.20	-0.60	0.55	-15.89	8.41

PC2	-8.55	11.16	-0.77	0.44	-30.42	13.32

					95% CI	
H. pT 0.0005 without rs9271069 VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-13.77	4.03	-3.42	6.38×10^{-4}	-21.68	-5.87
PGRS	0.86	0.51	1.70	0.09	-0.13	1.85
Age	0.03	0.02	1.63	0.10	-0.01	0.07
Sex	0.88	0.66	1.33	0.18	-0.42	2.18
PC1	-1.35	45.89	-0.03	0.98	-91.29	88.60
PC2	56.53	31.45	1.80	0.07	-5.11	118.16

					95% CI	
I. pT 0.0005 without rs9271069 GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-13.60	2.67	-5.09	3.53×10^{-7}	-18.83	-8.36
PGRS	1.00	0.35	2.81	4.99×10^{-3}	0.30	1.69
Age	0.02	0.01	1.69	0.09	0.00	0.05
Sex	-1.19	0.49	-2.42	0.02	-2.16	-0.23
PC1	7.15	21.27	0.34	0.74	-34.54	48.84
PC2	-46.63	23.06	-2.02	0.04	-91.83	-1.42

					95% CI	
J. pT 0.0005 without rs9271069 ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-8.68	1.67	-5.20	2.01×10^{-7}	-11.95	-5.41
PGRS	0.95	0.23	4.09	4.27×10^{-5}	0.49	1.40

					95% CI	
K. pT 0.0005 without rs9271069 VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-10.16	3.67	-2.76	5.71×10^{-3}	-17.36	-2.95
PGRS	0.89	0.50	1.76	0.08	-0.10	1.88

					95% CI	
L. pT 0.0005 without rs9271069 GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-12.42	2.52	-4.92	8.54×10^{-7}	-17.37	-7.48
PGRS	0.95	0.35	2.74	6.17×10^{-3}	0.27	1.63

					95% CI	
M. rs9271069 alone ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-1.47	0.47	-3.16	1.56×10^{-3}	-2.39	-0.56
PGRS	1.43	0.36	3.93	8.36×10^{-5}	0.72	2.14

Age	-0.01	0.01	-1.58	0.11	-0.03	0.00
Sex	-0.61	0.27	-2.30	2.14×10^{-2}	-1.13	-0.09
PC1	-0.42	6.27	-0.07	0.95	-12.70	11.87
PC2	-10.60	11.21	-0.95	0.34	-32.57	11.36

					95% CI	
N. rs9271069 alone VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-7.95	1.76	-4.51	6.39×10^{-6}	-11.41	-4.50
PGRS	1.07	0.84	1.27	0.20	-0.58	2.72
Age	0.03	0.02	1.62	0.11	-0.01	0.07
Sex	0.85	0.66	1.29	0.20	-0.45	2.15
PC1	-7.83	44.25	-0.18	0.86	-94.55	78.89
PC2	56.26	30.93	1.82	0.07	-4.36	116.89

					95% CI	
O. rs9271069 alone GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-6.79	0.72	-9.43	$< 2 \times 10^{-16}$	-8.20	-5.38
PGRS	1.13	0.58	1.96	0.05	0.00	2.27
Age	0.02	0.01	1.73	0.08	0.00	0.05
Sex	-1.17	0.49	-2.38	0.02	-2.14	-0.21
PC1	5.33	21.35	0.25	0.80	-36.51	47.18
PC2	-45.33	22.99	-1.97	0.05	-90.40	-0.27

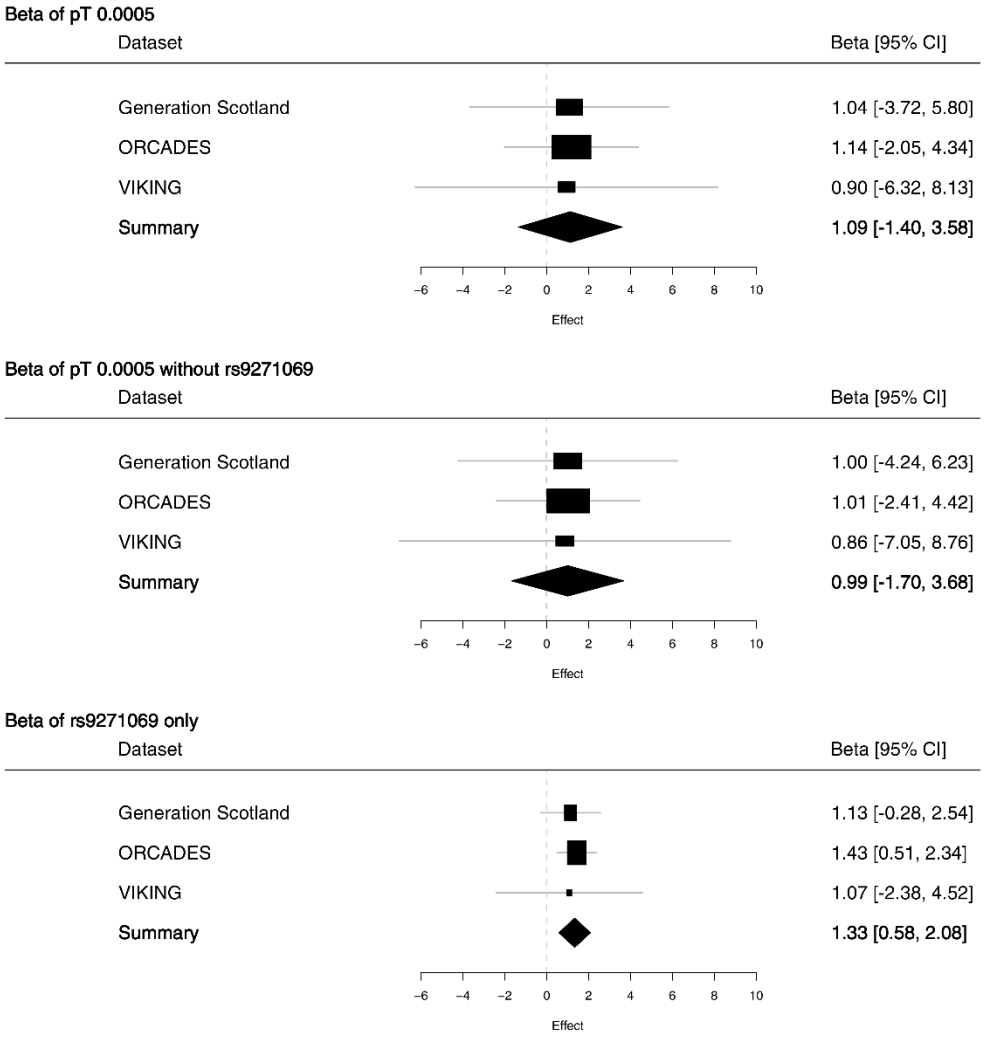
					95% CI	
P. rs9271069 alone ORCADES	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-2.36	0.17	-13.81	$< 2 \times 10^{-16}$	-2.69	-2.02
PGRS	1.41	0.35	3.98	7.00×10^{-5}	0.71	2.10

					95% CI	
Q. rs9271069 alone VIKING	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-4.11	0.38	-10.86	$< 2 \times 10^{-16}$	-4.85	-3.37
PGRS	1.14	0.82	1.40	0.16	-0.46	2.74

					95% CI	
R. rs9271069 alone GS	**Beta**	**SE**	**Z score**	**p-value**	**Lower**	**Upper**
(Intercept)	-5.92	0.25	-23.96	$< 2 \times 10^{-16}$	-6.40	-5.44
PGRS	1.15	0.58	1.99	0.05	0.02	2.29

Supplementary Table 4: Logistic regression results for predicting MS risk

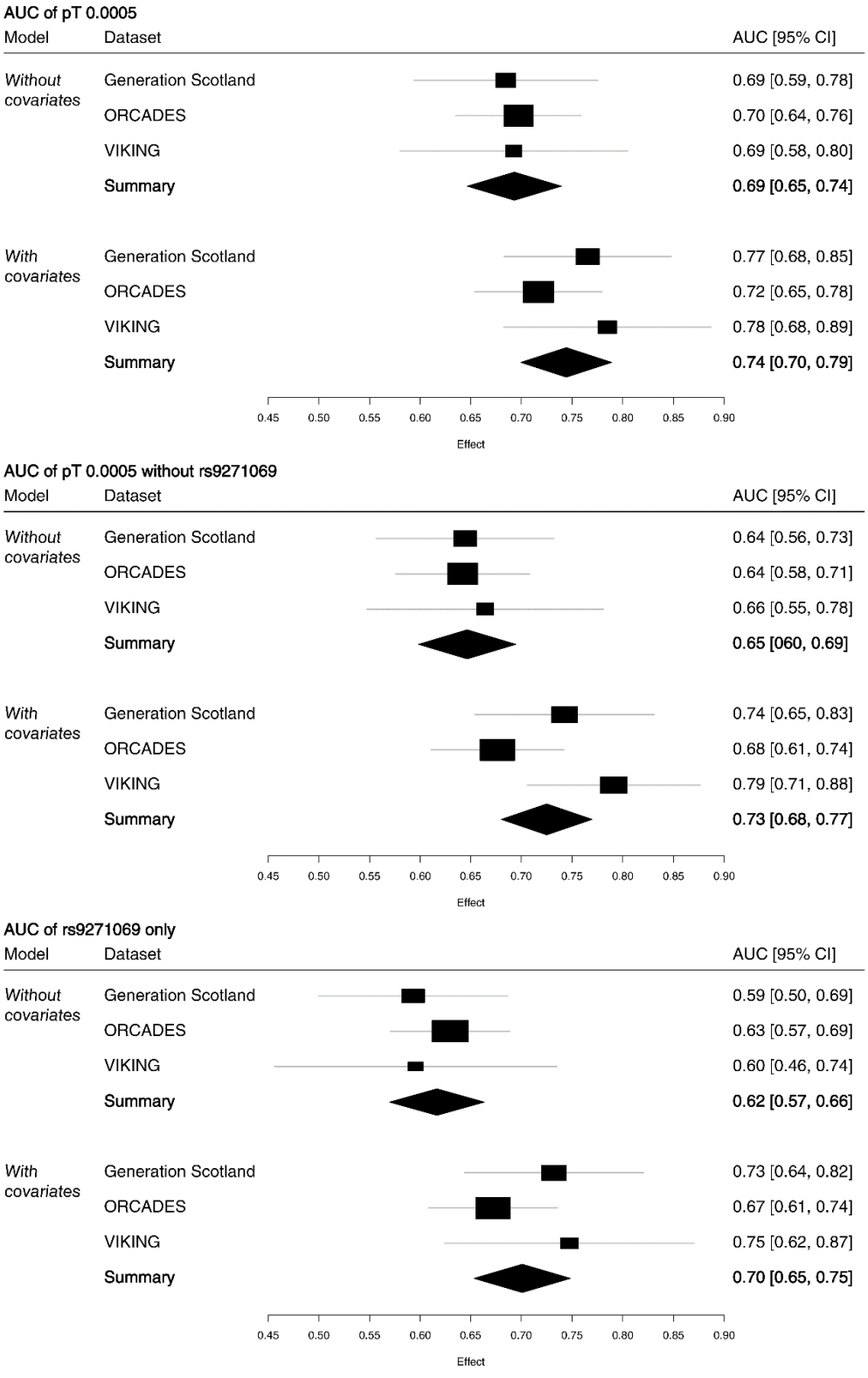
Logistic regression results for the model MS ~ PGRS + age + sex + PC1 + PC2 (letters A, B, C, G, H, I, M, N, O) and the model MS ~ PGRS (letters D, E, F, J, K, L, P, Q, R), using PGRS calculated from three SNP sets: pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1), using three datasets: Generation Scotland (n cases = 30, n controls = 8708), ORCADES (n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158).



Supplementary Figure 1: Meta-analysed beta scores from MS risk prediction models

Meta-analysed beta values for Generation Scotland (n cases = 30, n controls = 8708), ORCADES

(n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158) using beta values produced from the model MS~PGRS+age+sex+PC1+PC2, using three different SNP sets - pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1).


AUC of pT 0.0005
Model
Dataset
AUC [95% CI]
Without covariates
Generation Scotland
0.69 [0.59, 0.78]
ORCADES
0.70 [0.64, 0.76]
VIKING
0.69 [0.58, 0.80]
Summary
0.69 [0.65, 0.74]
With covariates
Generation Scotland
0.77 [0.68, 0.85]
ORCADES
0.72 [0.65, 0.78]
VIKING
0.78 [0.68, 0.89]
Summary
0.74 [0.70, 0.79]
0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
Effect
AUC of pT 0.0005 without rs9271069
Model
Dataset
AUC [95% CI]
Without covariates
Generation Scotland
0.64 [0.56, 0.73]
ORCADES
0.64 [0.58, 0.71]
VIKING
0.66 [0.55, 0.78]
Summary
0.65 [060, 0.69]
With covariates
Generation Scotland
0.74 [0.65, 0.83]
ORCADES
0.68 [0.61, 0.74]
VIKING
0.79 [0.71, 0.88]
Summary
0.73 [0.68, 0.77]
0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
Effect
AUC of rs9271069 only
Model
Dataset
AUC [95% CI]
Without covariates
Generation Scotland
0.59 [0.50, 0.69]
ORCADES
0.63 [0.57, 0.69]
VIKING
0.60 [0.46, 0.74]
Summary
0.62 [0.57, 0.66]
With covariates
Generation Scotland
0.73 [0.64, 0.82]
ORCADES
0.67 [0.61, 0.74]
VIKING
0.75 [0.62, 0.87]
Summary
0.70 [0.65, 0.75]
0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
Effect

Supplementary Figure 2: Meta-analysed AUC values

Meta-analysed area under the curve (AUC) values for Generation Scotland (n cases = 30, n controls = 8708), ORCADES (n cases = 94, n controls = 2120) and VIKING (n cases = 16, n controls = 2158) using the model MS~PGRS+age+sex+PC1+PC2 and the model MS~PGRS. AUC values were calculated in three different SNP sets - pT 0.0005 (n SNPS = 126), pT 0.0005 without rs9271069 (n SNPs = 125) and rs9271069 only (n SNPs = 1).