

Characterising the effects of sex  
interaction, pleiotropy and local  
population structure on  
**ALS GWAS**

Ross Byrne



**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

A thesis submitted to The University of Dublin  
for the degree of Doctor of Philosophy

Department of Genetics  
2021



I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: R Byrne Date: 18/03/2021

## Acknowledgements:

First, I would like to express my gratitude to my supervisor Russell McLaughlin for his unwavering support, encouragement and mentorship throughout my PhD. His willingness to listen and give advice no matter the size of the problem has made navigating both my project and the everyday challenges of being a scientist much less daunting. It is impossible to doubt yourself when backed by Russell's enthusiasm and positivity, which makes him a fantastic supervisor. I cannot thank him enough for the opportunities and perspective he has given me.

For their continued mentorship and support throughout my PhD I would like to give a special thank you to both Orla Hardiman and Daniel Bradley. Dan and his group in the Molecular Population Genetics lab generously welcomed me into their lab space in my first two years as a postgraduate and my experience would not have been the same without the great environment of collaboration and camaraderie he has fostered. I cannot thank him enough for introducing me to the field of human population genetics and sharing his advice so freely. Orla and her multidisciplinary research group at the Academic unit of Neurology have enriched my understanding of ALS far beyond genetics, tying in the unique and exciting perspectives from their multifaceted fields of study. I am indebted to her for welcoming me to into this wonderful research team and introducing me to the wider world of ALS research.

I would like to thank everyone in the project MinE consortium for welcoming me to their ranks and sharing such interesting conversations and meetings over the years. It goes without saying that much of the work in this thesis could not have been achieved without the resources and ideas this group shared. A special thanks to my collaborators and colleagues in Jan Veldink's group in UMC Utrecht for generously hosting me in their lab during my brief visit and for sharing data that was essential for my project. Thank you to Jan, Wouter, Rick, Kevin, Brendan, Joke and Kristel for this and all they have done to support me over the years. I would also be greatly remiss to go without thanking each individual who has donated their DNA to make our research possible, especially the many patients who are a driving force in our efforts to understand and hopefully cure ALS.

I would like to thank my great colleagues and friends in the McLaughlin and Bradley labs (past and present) for brightening my days and teaching me so much. I could not have made it this far without all of your help. If I listed all each of you have done for me here I would need a second volume so I will keep it brief. In no particular order a special thank you to Matthew, who guided me from the brink of far too many computer disasters; To Rui, who introduced me to finescale population genetics and high performance computing;

To Lara, who shared her vast expertise about human genetics and history so freely; To Valeria, who keeps us all together and is always there for everyone; To Victoria, for showing me that you can do both sport and a PhD; To Marta, who brightened the lab with pictures of cows; To Kevin, who has weathered hundreds of my half-baked ideas and always given helpful advice; To Andrew, who is always up to chat about anything; To Mark, whose infectious sense of humour has kept me sane some days; To Jenny, whose kindness and work ethic are inspiring; To Ciaran, who reminds me to question everything (by questioning literally everything); To both Eppie and Pier, who always lift the mood when they visit; To Conor, who always has a smile; To Emily, who tells great stories; To Bruno, who has tirelessly worked through many computer issues with me; and to Laura and Iseult, who have only joined us recently, but are sure to do great things. Thank you all for all the boardgames, trips to the pub and good times.

I could not have achieved what I have without my family. Thank you to my parents Siobhan and Des who raised me in an environment that supported learning and growth; to my brother Oisín and my sister Fiona, who have taught me so much over the years and to my late grandparents whose spirit and determination are a source of great motivation.

For giving me a second home at college I would like to thank my teammates and coach at Dublin University Fencing Club. Travelling, competing, training and simply hanging out with all of you has kept me active, motivated and lifted my spirits over the years. A special thanks to my coach Colm Flynn and my friends and teammates Achilleas Floudas and Killian Hanlon (who deserves a special thanks for encouraging me to pursue genetics) for sharing their experience from their PhDs and their constant support and advice.

As life is more than just work, a huge thank you to all of my friends who have given me plenty to smile about these past few years and support through stressful times. This has been invaluable for keeping me sane throughout my PhD, particularly through the recent challenges of the pandemic. An especially huge thank you to Fearghal, Camille, Colm, Donnchda, Seán, Brianna and Sam for countless hours of boardgames and chats throughout these months of lockdown and better times; to my housemates past and present for creating a wonderful welcoming place to live; to my friends from my undergraduate for being there no matter how much time or space separates us; and to Eoin and every member of the Genetics Department D&D group for the weekly games and a space to relax after a hard day in the lab. Most importantly, a massive thank you to my partner Camille for being there with me through all the new experiences, challenges and adventures we've shared together these past 6 years, I cannot stress how much it means to me.

I dedicate this thesis to the memory of my friend Gemma McGee for teaching me what determination and hard work can achieve.

## Summary:

This thesis focuses on the interface between disease genetics and population genetics through analysis of both individual level genotype data and summary statistics from genome wide association studies (GWAS) for amyotrophic lateral sclerosis (ALS). Work throughout aims to expand our understanding of the genetic architecture of ALS and improve the robustness of both GWAS and downstream analyses.

**Chapter 1** provides a general background and context for this thesis, outlining key information concerning modern techniques and advances in GWAS analysis; the genetics of ALS; and issues surrounding population structure from its impact on GWAS to its detection. This chapter concludes with a statement of the major aims of the thesis.

**Chapter 2** details two analyses to further understand the genetic architecture of ALS. These analyses explore i.) the genetic overlap of ALS with secondary psychiatric and cognitive traits, and ii.) the possibility of a sex-dependent genetic architecture in ALS. Work in this chapter makes several potentially important discoveries about the genetics of ALS showing evidence of a novel genetic correlation between ALS and bipolar disorder; that ALS fits well within the “p-factor” model for psychiatric traits; that ALS heritability is enriched in genes expressed in the frontal lobe and that ALS heritability differs across sexes, being both less heritable and less polygenic in males. We also identify several novel ALS loci through sex-specific and multi-trait genome wide scans, which have plausible roles in ALS. While these results give perspective on the role of genetics in the observed sex-differences and extra-motor symptoms in ALS patients, residual population structure appears to confound these analyses despite careful use of standard correction methods, motivating further study.

**Chapter 3** details work carried out to characterise local population structure and recent admixture events shaping the genomes of modern Irish individuals using Irish samples from an ALS GWAS (n=991) with associated geographic data, and external data from Britain and mainland Europe. This study identifies finescale population structure in Ireland at a higher resolution than prior work finding robust genetic clusters that segregate with geography by leveraging haplotype sharing methods. When considering external data from Britain and Europe we find that genes mirror geography across Britain and Ireland, and identify signatures of past admixtures in Ireland dating to the Viking and Norman invasions, and the Ulster Plantations. The subtle local population structure identified in this chapter is missed by standard methods using unlinked markers, which has implications for the design of GWAS.

**Chapter 4** details work characterising recent population structure, demographic change

and admixture in the Netherlands using Dutch samples from an ALS GWAS ( $n=1,626$ ) and external data from neighbouring European countries. This study reveals subtle local population structure based on patterns of haplotype sharing, identifying splits both between and within provinces. A novel method developed within shows that north-south structure in the country is strong and persistent, while east-west structure is more transient. This structure correlates with opposing clines of Germanic and Belgian ancestry across the country and may be partially caused by a lowered rate of migration across the Rhine, Meuse and Waal rivers, which we infer from genetic and spatial data. Effective population size estimates from identity-by-descent (IBD) sharing indicate that the population in the Netherlands has been growing rapidly in the past 50 generations, with significant uptick in the 17th century corresponding to a period of economic growth. This pattern is conserved across northern and southern groups, however the effective population of the north is consistently lower across time. Finally signatures of a population crash corresponding to the Black Death are only evident when data was analysed by province, potentially highlighting the importance of population structure to demographic inference based on IBD sharing.

**Chapter 5** builds on work from Chapters 3 and 4, and investigates the application of haplotype sharing principal components (PCs) as covariates to the problem of correcting confounding from population structure in a reanalysis of the project MinE GWAS for ALS. This is made possible at GWAS dataset scale using a fast and scalable a fast scalable implementation of ChromoPainter exploiting the Positional Burrows Wheeler Transform (PBWT-paint; <https://github.com/richarddurbin/pbwt/blob/master/pbwtPaint.c>) to speed up haplotype matching. Inflation due to confounding from population structure (measured using the LD score intercept) is reduced both in a small Dutch subset of the data ( $n=4,753$ ) and the full multi-country dataset ( $n=35,985$ ) when using these haplotype sharing PCs, while power to detect ALS risk variants is unaffected. Additionally both polygenic risk scores (PRS) calculated from GWAS corrected with these haplotype-based PCs and heritability estimates corrected with haplotype-based PCs are less affected by residual stratification than those corrected using standard SNP-based PCs. Strikingly, evidence from this analysis shows that the concentration of heritability in low frequency variants suggested by original analysis of the data may be overestimated, with implications for the genetic architecture of ALS. This chapter also proposes and validates a novel pipeline for quickly identifying population clusters in large datasets. The methods explored throughout this chapter will likely aid in the analysis of large GWAS and population genetics datasets.

Finally, **Chapter 6** highlights and discusses future directions opened by this thesis



## Table of Contents:

Acknowledgements:.....	ii
Summary:.....	v
Table of Contents:.....	vii
List of Figures:.....	x
List of Tables:.....	xii
List of abbreviations:.....	xiii
<b>Chapter 1 - General Introduction.....</b>	<b>1</b>
1.1 - <i>The genetics of ALS</i> .....	2
1.1.1 - Non-genetic risk and the multistep model.....	2
1.1.2 - The emerging genetic architecture of ALS.....	3
1.1.3 - Whole genome sequencing efforts.....	4
1.2 - <i>GWAS: A window into complex trait architecture and epidemiology</i> .....	5
1.2.1 - Genetic architecture.....	5
1.2.2 - Estimating heritability from individual level GWAS data.....	6
1.2.3 - Estimating heritability from GWAS summary statistics.....	7
1.2.4 - Genomic prediction: polygenic risk scores.....	11
1.2.5 - Non-additive effects and the role of rare variation.....	12
1.2.6 - Beyond polygenicity: The omnigenic model and pleiotropy.....	13
1.3 - <i>Population structure</i> .....	15
1.3.1 - Ancestry informative markers:.....	16
1.3.2 - Principal component analysis:.....	17
1.3.3 - Model based clustering: STRUCTURE and ADMIXTURE:.....	17
1.3.4 - Haplotype sharing methods:.....	19
1.4 - <i>Aims</i> .....	22
<b>Chapter 2 - The Extended Genetic Architecture of ALS.....</b>	<b>24</b>
2.1 - <i>Introduction</i> .....	24
2.1.1 - Background.....	24
2.1.2 - Research aims.....	30
2.2 - <i>Methods</i> .....	31
2.2.1 - Estimating genetic correlation between ALS and psychiatric traits.....	31
2.2.2 - Partitioned heritability analysis.....	31
2.2.3 - Genomic SEM – Multi-trait models of ALS and psychiatric disorders.....	32
2.2.4 - Detecting pleiotropic loci with multi-trait analysis.....	32
2.2.5 - Characterising heterogeneity in ALS genetics by sex.....	33
2.2.6 - Local heritability estimation and contrast polygenicity.....	35
2.2.7 - Latent Causal Variable Analysis: Understanding the causal relationship between ALS and secondary traits.....	36
2.2.8 - Functional annotation and analysis.....	36
2.3 - <i>Results</i> .....	37
2.3.1 - Characterising the genetic overlap between ALS and an extended set of neuropsychiatric and cognitive traits.....	37
2.3.2 - Sex specific architecture.....	54
2.4 - <i>Discussion</i> .....	68
2.4.1 - Genetic overlap with psychiatric and cognitive traits.....	68
2.4.2 - Differential genetic architecture of ALS in males and females.....	72
<b>Chapter 3 - Finescale Irish Population Structure and Migration.....</b>	<b>78</b>
3.1 - <i>Introduction</i> .....	78
3.1.1 - Background.....	78
3.1.2 - Research aims.....	80
3.2 - <i>Methods</i> .....	81
3.2.1 - Datasets.....	81
3.2.2 - Quality control.....	82
3.2.3 - Phasing.....	83
3.2.4 - fineSTRUCTURE analysis.....	83
3.2.5 - Cluster robustness.....	84

3.2.6 - Estimating admixture dates.....	85
3.2.7 - Ancestry proportion estimation.....	86
3.2.8 - ADMIXTURE analysis.....	86
3.2.9 - PCA and t-SNE analysis.....	87
3.2.10 - Mapping samples.....	87
3.2.11 - Statistical analyses.....	88
<b>3.3 - Results.....</b>	<b>89</b>
3.3.1 - Finescale population structure in Ireland.....	89
3.3.2 - The genetic structure of Ireland in the context of Britain.....	95
3.3.3 - Evidence of migration into Ireland.....	103
<b>3.4 - Discussion.....</b>	<b>107</b>
<b>Chapter 4 - Dutch Population Structure, Movement and Demographic Change.....</b>	<b>109</b>
<b>4.1 - Introduction.....</b>	<b>109</b>
4.1.1 - Background.....	109
4.1.2 - Research aims.....	111
<b>4.2 - Methods.....</b>	<b>112</b>
4.2.1 - Data and quality control.....	112
4.2.2 - fineSTRUCTURE analysis.....	113
4.2.3 - Cluster robustness and differentiation.....	113
4.2.4 - Ancestry profiles.....	114
4.2.5 - Identity by descent analysis.....	115
4.2.6 - Inferring admixture dates.....	116
4.2.7 - ADMIXTURE analysis.....	117
4.2.8 - Estimating mean pairwise IBD sharing.....	117
4.2.9 - Estimating recent changes in population size.....	117
4.2.10 - Estimating effective migration surfaces.....	118
<b>4.3 - Results.....</b>	<b>119</b>
4.3.1 - The genetic structure of the Dutch population.....	119
4.3.2 - Genome flux and stasis in the Netherlands.....	123
4.3.3 - Genomic signatures of Dutch mobility.....	127
<b>4.4 - Discussion.....</b>	<b>129</b>
<b>Chapter 5 - Detecting and Correcting Confounding in GWAS Using Haplotype Sharing</b>	
<b>Methods.....</b>	<b>135</b>
<b>5.1 - Introduction.....</b>	<b>135</b>
5.1.1 - Background.....	135
5.1.2 - Research aims.....	138
<b>5.2 - Methods.....</b>	<b>140</b>
5.2.1 - Datasets and initial quality control.....	140
5.2.2 - Phasing.....	141
5.2.3 - Haplotype painting analysis.....	141
5.2.4 - Principal component analysis and t-SNE.....	142
5.2.5 - Benchmarking PBWT-paint against ChromoPainter.....	142
5.2.6 - GWAS analyses.....	143
5.2.7 - Estimating variance explained in phenotype.....	144
5.2.8 - Estimating confounding with LD score regression.....	145
5.2.9 - Clustering haplotype sharing datasets using the Louvain method for community detection.....	145
5.2.10 - Polygenic risk scores and residual confounding.....	146
5.2.11 - Stringent heritability estimation using haplotype sharing covariates.....	147
<b>5.3 - Results.....</b>	<b>148</b>
5.3.1 - Correcting GWAS confounding in a single population dataset (Dutch).....	148
5.3.2 - Detecting population structure in large datasets with scalable methods.....	153
5.3.3 - Correcting multi-population GWAS structure with haplotype sharing methods.....	162
5.3.4 - Addressing bias from residual stratification in polygenic methods using haplotype sharing PCs.....	165
<b>5.4 - Discussion.....</b>	<b>171</b>
<b>Chapter 6 - Discussion.....</b>	<b>178</b>
<b>6.1 - Future directions.....</b>	<b>178</b>
6.1.1 - Replication and functional validation of putative ALS loci.....	178
6.1.2 - Updating multi-trait analysis for ALS in the face of growing GWAS datasets.....	179

6.1.3 - Towards greater diversity in the study of ALS genetics.....	179
6.1.4 - Evaluating correction of confounding using haplotype sharing PCs in other traits and simulation studies.....	180
6.1.5 - Application of haplotype sharing methods to rare variant association studies.....	181
6.1.6 - Expanding our understanding of human history using large scale GWAS data .....	182
6.2 - Concluding remark.....	184
<b>References.....</b>	<b>185</b>
<b>Appendix - Supporting material.....</b>	<b>200</b>
<i>Appendix material for Chapter 2 .....</i>	<i>200</i>
<i>Appendix material for Chapter 3 .....</i>	<i>201</i>
<i>Appendix material for Chapter 4 .....</i>	<i>209</i>
<i>Appendix material for Chapter 5 .....</i>	<i>219</i>
<b>Appendix 2 – Publications.....</b>	<b>228</b>

## List of Figures:

Figure 1.1: LD score regression explained.....	8
Figure 1.2: Comparing trait polygenicity using HESS.....	11
Figure 1.3: Subtle allele frequency differences between European countries.....	16
Figure 1.4: Toy illustration of chromosome painting.....	21
Figure 2.1: Stratified ALS heritability estimates for Irish population.....	26
Figure 2.2: Comparison of power and correlation between ALS GWAS.....	29
Figure 2.3: Genetic correlations of ALS with psychiatric traits and cognition.....	38
Figure 2.4: Common and two genetic factor model for ALS and psychiatric traits.....	39
Figure 2.5: ALS multi-trait analyses are enriched for genes expressed in brain.....	42
Figure 2.6: Cell type specific heritability enrichments.....	51
Figure 2.7: Cell type specific heritability enrichment correlations.....	52
Figure 2.8: Genotype by sex interaction in ALS.....	55
Figure 2.9: Inflation in sex stratified GWAS for ALS.....	56
Figure 2.10: Genome wide sex-specificity scan for ALS.....	57
Figure 2.11: Genome wide sex-differentiation scan for ALS.....	63
Figure 2.12: Sex stratified local heritability compared.....	65
Figure 2.13: Heritability vs chromosome length partitioned by sex.....	66
Figure 2.14: Illustration of sex differentiated liability threshold models of disease.....	76
Figure 3.1: Fine-grained population structure in Ireland.....	90
Figure 3.2: Genes mirror geography in the British Isles.....	96
Figure 3.3: Principal components 2 and 3 of combined Irish and British coancestry matrix.....	97
Figure 3.4: Inter-island exchange of haplotypes between the north of Ireland and northern Britain.....	99
Figure 3.5: ADMIXTURE analysis for PoBI/Irish cluster groups with ancient British samples.....	100
Figure 3.6: t-distributed stochastic neighbour embedding (t-SNE) of Irish and British coancestry matrix.....	102
Figure 3.7: All-Ireland GLOBETROTTER admixture date estimates for European and British surrogate admixing populations.....	104
Figure 3.8: GLOBETROTTER breakdown for clusters in the Republic of Ireland.....	105
Figure 4.1: The genetic structure of the people of the Netherlands.....	120
Figure 4.2: ChromoPainter-PCs relationship to geography.....	121
Figure 4.3: The ancestry profile of the Netherlands.....	123
Figure 4.4: The changing genomic structure of the Dutch population over time.....	124
Figure 4.5: Dutch effective population size over time.....	127
Figure 4.6: The effective migration surface of the Netherlands.....	128
Figure 5.1: Phenotype stratification captured by ChromoPainter PCs and SNP PCA in Dutch dataset.....	149
Figure 5.2: Comparing the relationship of cp-PCA and SNP PCA to Dutch geography.....	150
Figure 5.3: Stratified QQ-plots under unlinked and linked correction methods.....	152
Figure 5.4: Runtimes for PBWT-paint compared to projected runtimes for ChromoPainter.....	154
Figure 5.5: Benchmark of PBWT-paint vs ChromoPainter in Irish and Dutch data.....	155
Figure 5.6: Describing structure in large multi population datasets using t-SNE initialised with PBWT-paint PCs and SNP PCs.....	156
Figure 5.7: Louvain community detection method for identifying population subgroups in preliminary data.....	158
Figure 5.8: Detecting population structure in large haplotype sharing datasets with the Louvain community detection method.....	160
Figure 5.9: Louvain method for community detection identifies extremely subtle splits missed by fineSTRUCTURE in the “indivisible” SEE cluster.....	161
Figure 5.10: Phenotype stratification captured by PBWT-paint PCs and SNP PCA in multi-population dataset.....	163
Figure 5.11: Comparing GWAS power and inflation when corrected with haplotype sharing or SNP PCA.....	164
Figure 5.12: PRS model fit and distribution is affected by correction method.....	167
Appendix Figure 2.1: Comparing male and female per chromosome heritability estimates from GREML shows strong differences in distribution of genetic effects.....	200
Appendix Figure 3.1: Irish fineSTRUCTURE tree cluster details.....	201
Appendix Figure 3.2: t-SNE projection of British and Irish SNP data.....	202
Appendix Figure 3.3: PoBI/Irish fineSTRUCTURE tree cluster details.....	203
Appendix Figure 3.4: PoBI maximum concordance fineSTRUCTURE tree cluster details.....	204
Appendix Figure 3.5: European maximum concordance fineSTRUCTURE tree cluster details.....	205
Appendix Figure 3.6: Comparison of Linked vs Unlinked fineSTRUCTURE in Ireland at 166,139 SNPs.....	206
Appendix Figure 4.1: Total variation distance (TVD) tree for k=16 split in the Netherlands.....	209
Appendix Figure 4.2: SOURCEFIND ancestry gradients.....	210

Appendix Figure 4.3: Dutch and European ADMIXTURE CV-error plot .....	211
Appendix Figure 4.4: Old and recent IBD sharing per province and per cluster.....	212
Appendix Figure 4.5: Ratio of estimated $N_e$ /Census is stable over the past 3 generations.....	213
Appendix Figure 4.6: Convergence of MCMC chains for EEMS run in The Netherlands.....	214
Appendix Figure 4.7: ADMIXTURE modelling for Dutch and European samples. ....	215
Appendix Figure 4.8: Ancestry profile per Dutch cluster group.....	216
Appendix Figure 4.9: Geographic distribution of South Holland clusters from the SHOL cluster group. ....	217
Appendix Figure 5.1: LDSC intercepts from ALS GWAS using haplotype PCs vs SNP PCs. ....	219
Appendix Figure 5.2: Louvain clustering third iteration breakdown. ....	220
Appendix Figure 5.3: Correlation of per chromosome ALS heritability estimates corrected using PBWT-paint PCs and SNP PCs.....	221
Appendix Figure 5.4: Heritability estimates corrected with PBWT-paint retain evidence of polygenicity. ....	222

## List of Tables:

Table 2.1: MTAG hits for ALS and secondary traits. ....	40
Table 2.2: Functional annotations enriched in ALS and cognition MTAG analysis. ....	43
Table 2.3: cFDR hits for ALS and secondary traits. ....	44
Table 2.4: Tissue expression annotations enriched for cFDR(ALS Cognition) hits. ....	48
Table 2.5: Gene ontology terms enriched for cFDR (ALS Cognition) hits. ....	49
Table 2.6: Latent Causal Variable analysis of ALS and secondary traits. ....	53
Table 2.7 Sex-specific ALS GWAS scan results. ....	58
Table 2.8: Tissue specific upregulation for sex-specific genes. ....	59
Table 2.9: Gene ontology annotations enriched for sex-specific genes. ....	60
Table 2.10: Sex-partitioned SNP-based heritability and inflation. ....	64
Table 3.1: Mean pairwise $F_{ST}$ between Irish cluster groups. ....	89
Table 3.2: Total Variation Distance between Irish cluster groups. ....	91
Table 3.3: Prediction of longitude and latitude using GCTA GRM and ChromoPainter. ....	93
Table 4.1: Date and source estimates for admixture into the Netherlands. ....	125
Table 5.1: Measures of inflation in Dutch only GWAS of ALS with haplotypic (cp-PCA) and unlinked (SNP PCA) ancestry covariates on HapMap SNPs. ....	151
Table 5.2: Best fit PRSice model details for SNP and PBWT-paint PC correction methods. ....	165
Table 5.3: Measures of spatial autocorrelation for PRS under SNP PC and PBWT-paint PC correction. ....	168
Table 5.4: Total unpartitioned GREML SNP-based heritability estimates from 2016 ALS GWAS under SNP and PBWT-paint PC corrections. ....	169
Table 5.5: MAF partitioned GREML SNP-based heritability estimates from the 2016 ALS GWAS under SNP and PBWT PC correction. ....	170
Appendix Table 3.1: Europe GLOBETROTTER table. ....	207
Appendix Table 3.2: British GLOBETROTTER table. ....	208
Appendix Table 4.1: Mean pairwise $F_{ST}$ ( $\times 10^{-3}$ ) for Dutch clusters and European groups from Sawcer <i>et al.</i> (Sawcer <i>et al.</i> 2011). ....	218
Appendix Table 5.1: Sample breakdown for the 2016 ALS GWAS dataset. ....	223
Appendix Table 5.2: Geographic clustering of SNP and cp-PCs for Dutch only dataset. ....	224
Appendix Table 5.3: GREML per chromosome heritability estimates from 2016 ALS GWAS under SNP and PBWT-paint PC corrections. ....	226

## List of abbreviations:

<b>ADHD</b> - Attention deficit hyperactivity disorder	<b>LD</b> - Linkage disequilibrium
AIC - Akaike's information criteria	<b>LDSC</b> - Linkage disequilibrium score regression
<b>AIM</b> - Ancestry informative marker	<b>LLD</b> - Lower linkage disequilibrium
<b>ALS</b> - Amyotrophic lateral sclerosis	<b>LOCO</b> - Leave one chromosome out
<b>ANX</b> - Anxiety disorder	<b>LMM</b> - Linear mixed model
<b>AS</b> - Anglo-Saxon	<b>MAF</b> - Minor allele frequency
<b>BIP</b> - Bipolar disorder	<b>maxFDR</b> - Maximum false discovery rate
<b>CAD</b> - Coronary artery disease	<b>MCMC</b> - Markov chain Monte Carlo
<b>CBD</b> - Corticobasal degeneration;	<b>MS</b> - Multiple sclerosis
<b>CE</b> - Common era	<b>MTAG</b> - Multi trait analysis of GWAS
<b>cFDR</b> - Conditional false discovery rate	<b>Ne</b> - Effective population size
<b>CFI</b> - <b>Comparative fit index</b>	<b>NIV</b> - Non-invasive ventilation
<b>cM</b> - Centimorgan	<b>NNLS</b> - Non-negative least squares
<b>Cog</b> - Cognition	<b>NUTS</b> - Nomenclature of territorial units for statistics
<b>cp</b> - ChromoPainter	<b>PBWT</b> - Positional Burrows-Wheeler transform
<b>cp-PC</b> - ChromoPainter principal component	<b>PC</b> - Principal component
<b>cp-PCA</b> - ChromoPainter principal component analysis	<b>PCA</b> - Principal component analysis
<b>DNA</b> - Deoxyribonucleic acid	<b>PoBI</b> - People of the British Isles (Study)
<b>DWLS</b> - Diagonally weighted least squares	<b>PRS</b> - Polygenic risk score
<b>EEMS</b> - Estimate effective migration surface	<b>PSP</b> - Progressive supranuclear palsy
<b>EM</b> - Expectation maximisation	<b>PTSD</b> - Post traumatic stress disorder
<b>EN</b> - Enteral nutrition	<b>QC</b> - Quality control
<b>FDR</b> - False discovery rate	<b>RM</b> - Roman
<b>fs</b> - fineSTRUCTURE	<b>SA</b> - Structured association
<b>F<sub>ST</sub></b> - Fixation index	<b>SCZ</b> - Schizophrenia
<b>FTD</b> - Frontotemporal dementia	<b>SEM</b> - Structural equation modelling
<b>FUMA</b> - Functional mapping and annotation of genome wide summary statistics	<b>s-LDSC</b> - Stratified LD score regression
<b>GCI-GREML</b> - Genotype-covariate interaction GREML	<b>SNP</b> - Single nucleotide polymorphism
<b>GREML</b> - Genomic-relatedness-based restricted maximum likelihood	<b>SNP PC</b> - SNP genotype principal component
	<b>SNP PCA</b> - SNP genotype principal component analysis
	<b>t-SNE</b> - t-distributed stochastic neighbour embedding

**GRM** - Genetic relationship matrix

**GTE<sub>x</sub>** - Genotype-Tissue expression

**GWAS** - Genome wide association study

**HESS** - Heritability estimation from  
summary statistics

**IA** - Iron age

**IBD** - Identity by descent

**IMH** - Irish modal haplotype

**LCV** - Latent causal variable

**TVD** - Total variation distance

**UK** - United Kingdom

**VNR** - Verbal numeric reasoning

**WGS** - Whole genome sequencing



# Chapter 1 - General Introduction

The study of human genetics has benefited greatly in recent years from large scale international collaboration and data-sharing leading to the availability of ever growing genotype array datasets (e.g. the UK Biobank (Bycroft et al. 2018)), and a vast number of publicly available resources to maximise the availability and interpretation of results emerging from these datasets (e.g. the GWAS catalog (Buniello et al. 2019) and LD hub (Zheng et al. 2017)). This emerging data has significantly advanced our understanding of the spectrum of genetic variation within human populations (population genetics), and how it contributes to a wide range of traits (disease genetics). Indeed, owing to these sharing practices, the paradigm in genome wide association studies (GWAS) has moved from the insular analysis of individual variants affecting individual traits to understanding genetic pleiotropy and the shared genetic architectures of traits by integrating data from many studies. This paradigm may be particularly effective for understanding the genetic roots of rare diseases such as the late onset neurodegenerative disease amyotrophic lateral sclerosis (ALS), which due to its relative sparsity in the population may struggle with scale compared to GWAS for other common traits, limiting the power to study its genetic variation in isolation.

However, regardless of sample size, study design may have unintended impact on both single-trait and multi-trait analysis, with distributions of variables not necessarily related to the trait such as sample ancestry and sex potentially affecting results. Improving our understanding of these variables and how they impact both single-trait and multi-trait analysis is thus an important area of focus. The work laid out in this thesis focuses on applying the paradigm of multi-trait analysis to the study of ALS, and exploring and addressing how latent population structure and sex affect our interpretation of data from a large GWAS dataset for ALS (van Rheenen et al. 2016). In addition, we conduct in-depth regional population genetic analysis of two countries from this dataset (Ireland and the Netherlands) to improve our understanding of historical events shaping modern populations.

## 1.1 - The genetics of ALS

The principal trait of interest in this thesis is amyotrophic lateral sclerosis (ALS). ALS is a late-onset and fatal neurodegenerative disease with a lifetime risk of 1:400 in the general population (Hardiman, van den Berg, and Kiernan 2011), qualifying it as a rare disease. While twin studies suggest that it is highly heritable: 0.61 (0.38 - 0.78) (Al-Chalabi et al. 2010), ALS only presents within families in between 5-20% of cases (Hardiman, van den Berg, and Kiernan 2011; Ryan et al. 2018), with the remainder showing no family history, complicating genetic analysis of the disease. Strikingly SNP-based heritability estimates are substantially lower than twin studies (~8.5% vs ~60%), highlighting additional difficulties in characterising the genetic causes of the disease (van Rheenen et al. 2016). These complications are reflected in the rate of discovery of ALS-related genes; while the first pathogenic mutations in familial cases were discovered in *SOD1* over 25 years ago via linkage analysis (Rosen et al. 1993), the next major ALS gene *TARDBP* took over 15 years to discover (Sreedharan et al. 2008). Although improvements in sequencing and genotyping technology have led to an increased rate of discovery in subsequent years, the percentage of cases explained by genetics remains low. The most common ALS mutation, the *C9orf72* repeat expansion, is only found in roughly 7% of sporadic, and 40% of familial cases (Majounie et al. 2012), hence explaining less than 15% of cases. A recent review estimated that only ~70% of familial and ~15% of sporadic cases have mutations in known ALS genes (Chia, Chiò, and Traynor 2018), meaning the majority of cases (~80%) are unexplained genetically.

### 1.1.1 - Non-genetic risk and the multistep model

One likely reason for the sparsity of genetically explained cases could be the involvement of non-genetic components, such as environmental or lifestyle risks. While it is generally agreed that non-genetic factors play a role in ALS, much of the evidence for specific lifestyle and environmental risks is inconclusive (reviewed by Ingre et al (Ingre et al. 2015)). Intriguingly as noted by Al-Chalabi et al (Al-Chalabi et al. 2014), even in individuals with known mutations, onset typically occurs suddenly in late life, suggesting that non-genetic factors accumulated across a lifetime play a role. In further support of this, investigation of disease registers from 5 countries yielded a linear relationship between log incidence and log age (Al-Chalabi et al. 2014), consistent with a multistep model for the disease whereby six insults are needed to develop ALS. More recent work around this concept of a multistep process of ALS pathology has shown that large effect mutations such as the *C9orf72* repeat expansion can lower the slope of the above regression, accounting for more than one step in this multistep process (Chiò et al. 2018).

The authors argue that focusing on cases where genetics account for the majority of the steps (for example *SOD1* which accounts for 4 of 6 steps) may increase our chances of identifying clear environmental factors as fewer factors will be involved (Chiò et al. 2018). Together these and other observations demonstrate the complexity of ALS as a multifactorial disease.

### 1.1.2 - The emerging genetic architecture of ALS

Genetic work in ALS is not limited to gene discovery. A large GWAS of ALS performed in 2016 explored its genetic architecture, finding that heritability is roughly proportional to chromosome length (van Rheenen et al. 2016). This means that variants contributing to the disease are uniformly spread throughout the genome, suggesting that ALS is polygenic, or driven by many variants. The existence of a polygenic component of ALS has prompted exploration of hypotheses of genetic overlap between ALS and other traits using methods such as bivariate LD score regression (Bulik-Sullivan, Finucane, et al. 2015) (introduced in section 1.2.3). A prime example of this is work demonstrating the genetic correlation of the polygenic components of schizophrenia and ALS (R. L. McLaughlin et al. 2017), which bolstered previous evidence of increased rates of psychotic illness in first degree relatives of ALS patients (S. Byrne et al. 2013). Leveraging knowledge of this correlation, the authors used pleiotropy-informed conditional false discovery rate (cFDR) analysis to identify novel ALS loci including *TNIP1*, which has since been validated in both a cross-ethnic meta-analysis of ALS and a meta-analysis with North American samples (Benyamin et al. 2017; Nicolas et al. 2018). More recently studies have leveraged “pleiotropic enrichment” methods, which evaluate if the proportion of SNPs associated with a given phenotype increases as a function of association with a second phenotype, indicating genetic overlap. One such study focused on the overlap between ALS and diseases of the frontotemporal dementia (FTD) spectrum (Karch et al. 2018) and identified significant overlap with FTD, progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD). A subsequent wider-reaching study which looked at pleiotropic enrichment with ALS across 65 GWAS traits confirmed the above genetic overlaps, and added several more traits to the growing list (coronary artery disease, memory, C-reactive protein, celiac disease, body mass index and verbal numeric reasoning), identifying 59 novel pleiotropic ALS loci (Broce et al. 2018). While these pleiotropic loci are unlikely to contribute a large amount to genetic risk for ALS, they may contribute incrementally to our overall understanding of its disease biology. These studies highlight the power of contextualising relatively small GWAS for a rare trait with a number of large scale GWAS.

A second key feature that GWAS has revealed about ALS is that heritability is enriched in lower frequency variants (van Rheenen et al. 2016), signposting a potentially important role for rare variants. In support of this observation, gene burden analysis for excess rare variants has since identified two new ALS-associated genes, *NEK1* and *KIF5A* (Kenna et al. 2016; Nicolas et al. 2018). Several recent studies have also observed that a significant number of ALS patients have 2 or more rare mutations in known ALS genes, suggesting an oligogenic basis to the disease (Morgan et al. 2017; Pang et al. 2017). The growing importance of understanding the role of rare variants in ALS has served as a motivation for the establishment of the international whole-genome sequencing (WGS) consortium Project MinE (van Rheenen et al. 2018).

### 1.1.3 - Whole genome sequencing efforts

The Project MinE sequencing consortium was established to uncover genetic risk factors for ALS through whole genome sequencing 15,000 patients and 7500 controls (van Rheenen et al. 2018). At the time of writing this consortium has generated and analysed WGS data for 4366 ALS patients and 1832 matched controls (van der Spek et al. 2019), and made the results of initial analysis publicly available in a databrowser (<http://databrowser.projectmine.com/>) to disseminate results and allow use as a replication cohort for other sequencing studies. The major advantage of WGS over SNP arrays typically used in GWAS analysis comes in its ability to identify rarer variation which is expected to be crucial in ALS (van Rheenen et al. 2016, 2018). However use of this technology may come with complications of its own. The pilot phase has identified that rare variants show tight spatial clustering in a subsample from the Netherlands (van Rheenen et al. 2018). Given previous work showing that rare variants may be subject to differential confounding than common variants (Mathieson and McVean 2012), this structuring of rare variants may bias results. Hence, careful adjustment will be crucial for the success of the project. For example it is conceivable that false positive associations could result from subtle population stratification if a rare variant's geographic range overlapped the geographic range of an environmental risk factor for ALS. As such proper control for local population structure may prove essential in future efforts to explore the association of rare variants to ALS. In the pilot study it appears that false discovery rates are not significantly increased in this small primary cohort (van Rheenen et al. 2018), possibly suggesting that burden tests which aggregate many rare variants are less susceptible to confounding from structure. However, with the addition of more samples this analysis may have to be revisited. Hence there is a need for careful research into the current methods for identifying and characterising population structure.

## 1.2 - GWAS: A window into complex trait architecture and epidemiology

In principle, genome wide association studies (GWAS) assay the association between genetic variants spread across the entire genome and a measured trait (e.g. height) in a large sample of individuals with the goal of identifying loci that contribute to the heritability of that trait. These studies typically use a regression framework, which usually takes the form of a simple linear model for continuous traits, a logistic model for binary traits or a linear mixed model in cohorts with significant family and population structure (Yang et al. 2014). As millions of concurrent tests are performed in a given GWAS, stringent p-value thresholds based on the number of independent common loci are applied to correct for multiple testing and reduce false positives. For GWAS in Europeans a p-value threshold of  $5 \times 10^{-8}$  is typically used, however as lower frequency alleles (e.g. <1% minor allele frequency) are assayed this may need to be adjusted (Fadista et al. 2016). External replication of significant loci in an independent sample is a gold standard requirement for these studies to reduce artifactual findings. Given that GWAS in humans are not controlled experiments but instead observational studies, multiple confounding covariates (e.g. age, sex, ancestry etc) must be accounted for to produce robust results. In particular sample ancestry is commonly corrected by fitting genetic principal components (PCs) as covariates (Price et al. 2006; Patterson, Price, and Reich 2006) in the regression, which capture broad ancestral differences that could lead to allele frequency differences between cases and controls unrelated to the studied trait (we will cover population structure more thoroughly in section 1.3). Owing to these many precautions, GWAS findings are highly replicable across studies as highlighted by the large numbers of replicated hits recorded in the GWAS catalog (Marigorta et al. 2018; Buniello et al. 2019), providing reliable insights into the genetic roots of diseases and other traits.

### 1.2.1 - Genetic architecture

In addition to reliably identifying loci of interest in multiple diseases, GWAS results have enabled us to explore the genetic architecture of human traits in unprecedented detail. A trait's genetic architecture is defined by several factors including the number of genetic variants that confer risk for the trait, the distribution of their effects, the frequencies of these variants and the manner in which they interact (Gratten et al. 2014; Timpson et al. 2018). Trait architectures are typically classified as monogenic, oligogenic or polygenic based on whether one, a few or many variants drive them, respectively (Badano and Katsanis 2002). For example, Mendelian disorders are driven by high impact variants in a single gene, and are thus considered to be monogenic, while complex traits such as

height, which are driven by many lower impact variants spread throughout the genome, are considered polygenic (Yang et al. 2010). Understanding the genetic architecture of a disease is crucial for properly designing studies of that disease, for example when considering a rare variant-driven oligogenic disease (driven by a few genes), whole genome sequencing (WGS) might be essential for identifying the causal mutation given that WGS can identify low frequency variants not covered by standard genotyping arrays. In contrast, WGS might not be cost effective for studying a common variant polygenic disease whose variants are likely to individually be low effect, requiring large sample sizes to detect, when this variation could instead be adequately described by a much cheaper technology such as a SNP array.

### 1.2.2 - Estimating heritability from individual level GWAS data

The past decade has seen the development of a wide range of methods for estimating trait heritability from both individual level data (e.g. GCTA GREML (Yang et al. 2011)) and more widely accessible GWAS summary statistics (LD score regression (LDSC) (Bulik-Sullivan, Loh, et al. 2015), Heritability Estimation from Summary Statistics (HESS) (Shi, Kichaev, and Pasaniuc 2016) and LDAK (Speed et al. 2017)), which have collectively revealed that the vast majority of complex traits are highly polygenic. Strategies aggregating the effects of all SNPs across the genome when estimating the phenotypic variance explained by GWAS for complex traits (the earliest example is height (Yang et al. 2010)) have reconciled much (but not all) of the famous “missing heritability” attributed to early GWAS (Maher 2008; Manolio et al. 2009). These approaches demonstrate that properly modelling a trait’s genetic architecture is crucial for estimating heritability. Before this, researchers noted that significant SNPs identified by GWAS could explain only a fraction of the phenotypic variance attributed to total genetic effects by twin studies (Maher 2008; Manolio et al. 2009). The phenotypic variance explained by only significant SNPs under an additive model is termed the GWAS heritability ( $h^2_{\text{GWS}}$ ), while the variance explained by all genotyped SNPs collectively is termed the SNP-based heritability ( $h^2_{\text{SNP}}$ ) (Yang et al. 2017), which yields a closer value to heritability estimates from family and twin studies.

The first method developed for estimating SNP-based heritability for quantitative (Yang et al. 2010) and binary (S. H. Lee et al. 2011) traits from individual level genetic data is commonly referred to as GREML (Genomic-relatedness-based Restricted Maximum Likelihood) and uses a mixed linear model to investigate the relationship between phenotypic similarity and genetic similarity. In this approach genetic similarity is represented by a pairwise “genetic relationship matrix” (GRM) constructed from all SNPs

genotyped in the dataset. GREML is typically carried out in distantly related individuals which minimises bias originating from a shared environment between related samples which may otherwise lead to overestimation of heritability (Yang et al. 2011, 2017). Notably this method can only capture the phenotypic variance explained by causal variants adequately tagged by the SNP set used, meaning  $h^2_{\text{SNP}}$  will be smaller than the true narrow-sense heritability ( $h^2$ ; proportion of phenotypic variance explained by additive effects) for most SNP arrays due to inability to tag rare causal variants (Yang et al. 2017).

Application of the base GREML model can be biased if the set of causal variants have a different minor allele frequency (MAF) spectrum or different linkage disequilibrium (LD) levels to the SNP set used to construct the GRM (Yang et al. 2015), motivating the development of LD and MAF stratified GREML model (GREML-LDMS). Applying GREML-LDMS to imputed GWAS data for height demonstrated a moderate increase in  $h^2_{\text{SNP}}$  from 45% (Yang et al. 2010) to 56% (Yang et al. 2015), indicating that rare variation not captured by SNP arrays harbour some of the missing heritability in height. Despite the power and flexibility of this method it is limited in its usage by restrictions on sharing sensitive individual level genetic data between research groups and computational cost when running on large sample sets, motivating development of quick approximate methods using summary statistics such as LD score regression (Bulik-Sullivan, Loh, et al. 2015; Zheng et al. 2017), HESS (Shi, Kichaev, and Pasaniuc 2016) and LDAK (Speed et al. 2017).

### 1.2.3 - Estimating heritability from GWAS summary statistics

The LD score regression (LDSC) model was developed to distinguish inflation caused by polygenicity from confounding (Bulik-Sullivan, Loh, et al. 2015). This model draws from the idea that variants in LD with a causal variant will have an inflated GWAS test statistic ( $\chi^2$ ) as a result of this linkage. Hence under an infinitesimal or polygenic architecture with causal variants spread across the entire genome, there is an expected linear relationship between how much variation a given SNP tags (LD score = the sum of  $r^2$  between the SNP and surrounding SNPs) and its GWAS test statistic (Figure 1.1). A linear regression of  $\chi^2$  vs LD score under this model has a slope proportional to the trait heritability, as described in the central equation from the paper (Bulik-Sullivan, Loh, et al. 2015):

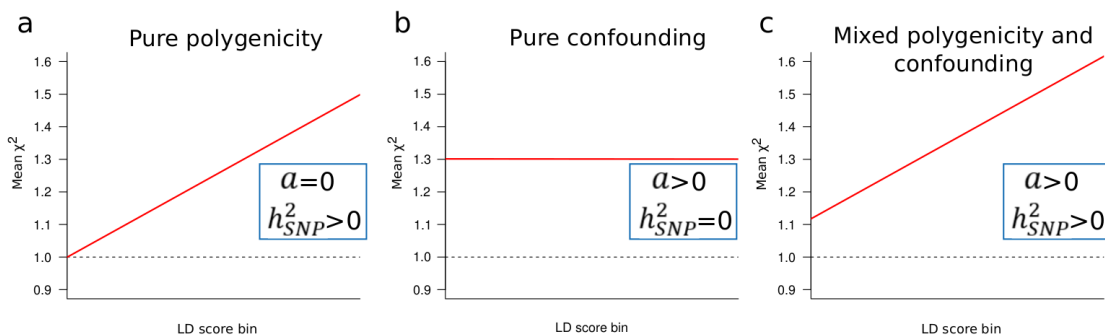
$$E[\chi^2 | l_j] = \frac{N h^2 l_j}{M} + N a + 1, \quad (1)$$

where  $N$  is the number of individuals in the GWAS;  $M$  is the number of SNPs;  $h^2/M$  is the per-SNP heritability;  $l_j$  is the LD score of variant  $j$  and  $a$  is a measure of confounding due to population structure. GWAS inflation caused by cryptic relatedness or population structure is not expected to correlate with LD, and should thus only affect the intercept of the model ( $Na + 1$ ). This model is particularly useful for investigating complex trait heritability as it does not rely on individual level data, enabling researchers to apply it to publicly-available summary statistics, provided a suitable population-matched LD reference panel is available. Moreover, LDSC also provides a direct estimate for bias due to confounding (Figure 1.1), making it a useful tool for assessing how well-controlled a GWAS is from summary statistics alone (a property we will exploit in Chapter 5).

## LD score regression

$$\text{Model: } E[x^2 | l_j] = \frac{Nh_{SNP}^2}{M} \cdot l_j + Na + 1$$

Y
Slope
X
Intercept



**Figure 1.1: LD score regression explained.**

The LD score regression model is a linear model describing changes in GWAS statistics (chi-squared, y axis) as a function of the LD score (x axis) with slope proportional to SNP heritability and intercept equal to one plus a term proportional to confounding. Model plots of LD score regression demonstrate how the model distinguishes inflation resulting from polygenicity from confounding in practice.

- Where inflation of summary statistics are completely due to polygenic signal, SNPs with stronger LD scores will be more likely to tag a causal SNP resulting in a positive slope. The intercept will equal one in this case, as the confounding term  $a$  equals zero.
- Where inflation is entirely due to confounding there will be no relationship between chi-squared and LD score, hence the slope will be zero. The confounding term describes all inflation here and hence the intercept will be raised above 1.
- Where there is inflation of GWAS statistics due to a mixture of confounding and polygenicity we will see both a positive relationship between LD score and chi-squared statistic (polygenicity), and an intercept above 1 (confounding). This is typically the case for real datasets.



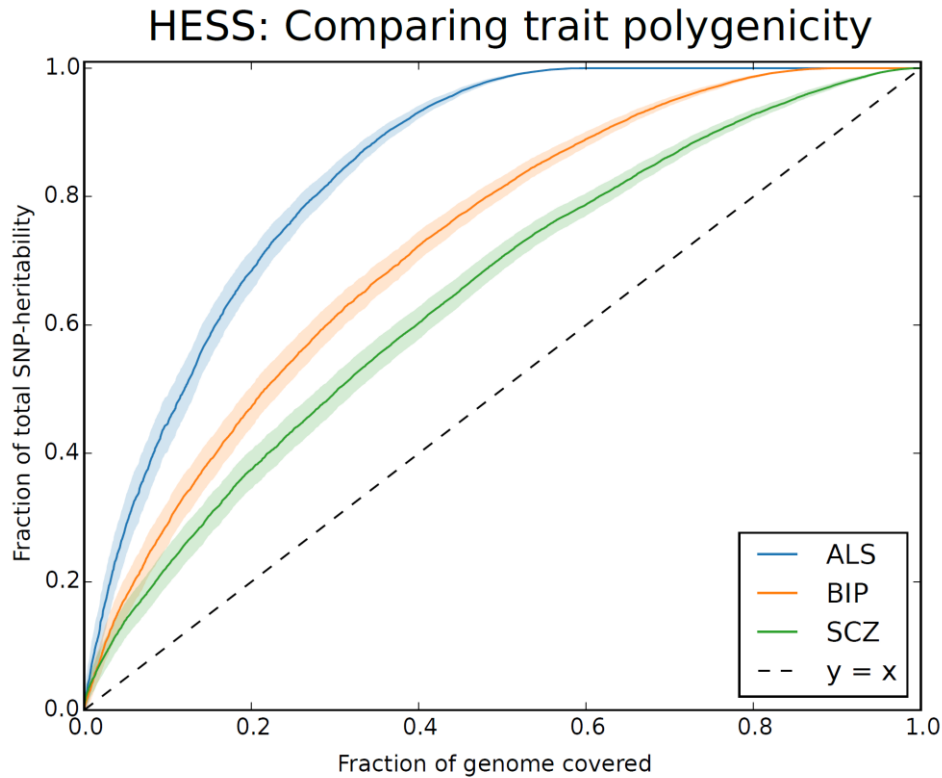
In addition to estimating genome wide heritability, an extension of LD score regression known as stratified LD score regression (s-LDSC) has been developed to test for enrichment of heritability within genomic annotations (Finucane et al. 2015) (e.g. coding regions, promoters etc.), yielding answers to questions about which specific regions of the genome are most involved in trait heritability, and which classes of variants contribute most to risk for a given disease. This method has revealed important features of the architecture of complex traits and disease biology that might otherwise have been missed by standard pathway analyses and other approaches solely focusing on loci achieving significance. Of note studies using this method have identified a shared heritability enrichment across complex traits in conserved and coding regions (Finucane et al. 2015), pointed towards the enrichment of heritability for complex diseases in regions under strong negative selection (Gazal et al. 2017; Zeng et al. 2018) and demonstrated the relative enrichment of heritability in older and conserved regulatory regions (enhancers and promoters) (Hujuel et al. 2019). In addition, this method has revealed heritability enrichments in genes specific to a tissue or cell type for a number of traits (Finucane et al. 2018) (e.g. CNS for schizophrenia). Some authors have argued for using these functional annotations known to be enriched for polygenic burden to weight SNPs and increase power in future GWAS (Kichaev et al. 2019), which demonstrates how growing knowledge on the general architecture of complex traits can feed into future discovery.

However, there are caveats for the use of the LD score regression model, for example if the causal SNPs for a trait have a relationship with LD score, then the heritability estimates will become biased (downwardly biased where causal SNPs have LD scores below the genome-wide median LD score, and upwardly biased where the causal SNPs have LD scores above the genome-wide median LD score) (J. J. Lee, McGue, et al. 2018). Given the emerging evidence that SNPs with lower LD (LLD) tend to have higher per SNP heritability (Gazal et al. 2017), this could mean estimates of heritability from LDSC are regularly downward biased by LD-dependent architecture. On top of this, LDSC is based on an infinitesimal model, meaning it assumes all SNPs contribute to trait heritability, which may be violated for many traits.

Unlike LDSC, which assumes all SNPs contribute to trait heritability, HESS makes no assumptions about the distribution of effect sizes, meaning it should be more robust to a wider range of genetic architectures (Shi, Kichaev, and Pasaniuc 2016). HESS can estimate local SNP heritability ( $h^2_{g,local}$ ) robustly from summary statistics, accounting for LD structure at each locus using eigenvectors of an LD matrix (Shi, Kichaev, and Pasaniuc 2016). This LD matrix can be calculated either within sample or from a reference dataset,

making the method versatile. In this way HESS can dissect which regions of the genome contribute most to heritability. Applied to 30 traits across 1,703 approximately independent loci HESS showed that a large fraction of trait heritability is found in common variants (Shi, Kichaev, and Pasaniuc 2016). As HESS estimates the contribution heritability of each region of the genome, it can also be used to compare the degree of polygenicity of traits. This can be achieved visually by ordering loci by their relative heritability and plotting the cumulative fraction of total heritability they explain together vs the cumulative fraction of the genome they cover (Figure 1.2). Traits closer to the 1:1 line show higher polygenicity as a large fraction of their genome contributes to heritability, while curves deviating drastically from this line suggest fewer SNPs drive the architecture. In Figure 1.2 we see that ~80% of heritability for the neurodegenerative disease amyotrophic lateral sclerosis (ALS) is explained by ~20% of the genome, suggesting it is much less polygenic than the neuropsychiatric diseases bipolar disorder and schizophrenia, which require 40% and 60% of the genome to explain the same fraction of heritability. Extreme examples of highly polygenic vs weakly polygenic traits from the initial HESS study of this method are height, which had heritability spread quite evenly across the genome and rheumatoid arthritis which had a lot of its heritability concentrated in a small fraction of SNPs (Shi, Kichaev, and Pasaniuc 2016).

HESS can additionally be applied to multiple traits to identify shared regions enriched across all these traits, making it suitable for identifying pleiotropic loci important in the architecture of human disease. Moreover the model has been expanded to allow investigation of local genetic correlations between traits even in the absence of genome-wide genetic correlation (Shi et al. 2017), enabling a more nuanced look at shared genetic architecture than genome-wide methods such as bivariate LD score regression (Bulik-Sullivan, Finucane, et al. 2015).



**Figure 1.2: Comparing trait polygenicity using HESS.**

The cumulative fraction of SNP-based heritability versus the fraction of the genome covered as estimated in HESS provides a visualisation of relative polygenicity for traits. For illustrative purposes polygenicity curves for amyotrophic lateral sclerosis (ALS) (van Rheenen et al. 2016), bipolar disorder (BIP) (Ruderfer et al. 2018) and schizophrenia (SCZ) (Ruderfer et al. 2018) calculated following the methods in Chapter 2 are plotted. Traits with high polygenicity are expected to trend towards the 1:1 line (dashed) as heritability is spread throughout the genome, while traits with steep inclines are explained by fewer variants and hence less polygenic. Here schizophrenia appears to be the most polygenic of the traits measured, while ALS appears to be the least polygenic. This demonstrates the utility of developing reliable local estimates of heritability in assessing the genetic architecture of traits.

#### 1.2.4 - Genomic prediction: polygenic risk scores

As well as identifying and describing the genetic factors that contribute to trait heritability, GWAS summary statistics can also theoretically be leveraged to predict the disease liability of individuals in an independent cohort for a given trait (Wray, Goddard, and Visscher 2007). The most common method of genomic prediction is known as a polygenic risk score (PRS), which is simply a score calculated from the weighted sum of effect alleles an individual has for a given trait, with weights derived from a GWAS of the same trait. Methods for calculation of these scores vary; while some methods select SNPs for inclusion based on a p-value threshold, and prune correlated SNPs (e.g. PRSice

(Euesden, Lewis, and O'Reilly 2015)), others directly model the correlation structure between SNPs and use all variants to achieve the best prediction (e.g. the Bayesian method LDpred (Vilhjálmsson et al. 2015)).

While early PRS for disease had little use in a clinical setting due to small, non-actionable risk increases in individuals with extreme scores (Ripatti et al. 2010), more recent studies have shown potential clinical utility. For example, a recent study in the UK Biobank showed a three-fold increase in risk for coronary artery disease (CAD) in individuals with high PRS which is comparable to the risk conferred by rare monogenic mutations (Khera et al. 2018). Given the prevalence of individuals with these high PRS scores was 20 fold higher than those with rare mutations conferring the same risk increase, PRS for CAD may become a useful predictor in diagnostic models (Khera et al. 2018). However, there are still several obstacles to application of PRS in the clinical setting. Most notably PRS have been shown to transfer poorly across ancestries, with lower predictive power in populations with divergent ancestry from the reference GWAS cohort (A. R. Martin et al. 2017). Given the huge bias towards exclusively European GWAS at present, and the resulting significantly attenuated predictive power of PRS from these GWAS in non-European individuals (Duncan et al. 2019), use of PRS in a clinical setting is likely to result in significant health disparities across populations (A. R. Martin et al. 2019). In addition to reduced predictive power across ancestries, PRS also show evidence of potential bias from residual population structure, with PRS for a number of traits clustering geographically in both Finland (Kerminen et al. 2019) and the UK (Haworth et al. 2019; Abdellaoui et al. 2019). Some of this clustering of PRS in the UK Biobank may be partially explained by extremely recent socio-economic driven migration (Abdellaoui et al. 2019), which could theoretically result in further disparities in their clinical use. Hence, while the improvements in power of PRS heralded by the large training and testing sets emerging from biobanks suggest a possible clinical application, until GWAS become more diverse and potentially better controlled for population structure, PRS will be of limited clinical use for large numbers of individuals, and will likely exacerbate damaging health inequalities due to the above systematic biases.

#### 1.2.5 - Non-additive effects and the role of rare variation

While additive models accounting for polygenicity (e.g.  $h^2_{\text{SNP}}$ ) have made huge strides in closing the gap between twin studies and early estimates of heritability from GWAS data for complex traits, there is still a notable disparity. Some researchers have posited that the remaining missing heritability is held by non-additive effects (e.g. dominance and epistasis). However, GWAS results from 79 quantitative complex traits have been used to

demonstrate that dominance genetic variation contributes very little to phenotypic variance (Z. Zhu et al. 2015), and precise estimates of the variance attributable to epistasis are expected to require huge samples (millions of individuals) (Yang et al. 2017). Alternatively, rare variants not captured by SNP panels or imputation may hold the remaining heritability. Recent application of heritability estimation methods coupled with whole genome sequencing (WGS) have further closed the gap between the SNP heritability and twin study heritability for height and BMI thanks to the inclusion of rarer variation (Wainschtein et al. 2019). This emerging work implies that many of the causal genetic variants contributing to variation in height and BMI are rare and in low LD, challenging the prior observation that polygenic diseases are mainly driven by mainly common variants (Shi, Kichaev, and Pasaniuc 2016) and motivating future adoption of WGS in GWAS design once adequate sample sizes become available.

#### 1.2.6 - Beyond polygenicity: The omnigenic model and pleiotropy

While GWAS signals are enriched in functional regions of the genome, the ubiquity of weak effect variants spread broadly across the entire genome in complex traits has brought into question the idea that all associated variants should fall in biologically relevant genes and pathways. This motivated the development of a new theoretical framework for considering their genetic architecture termed the “omnigenic” model (Boyle, Li, and Pritchard 2017). This model builds on Fisher’s infinitesimal model (Fisher 1918), in which every variant is considered to have a non-zero effect on the phenotype. The framework of this model is that some genes will have direct effect on trait biology (i.e. produce a protein or transcript that directly affects a relevant cellular process leading to variation in the trait) which are termed “core” genes, while all other genes expressed in a relevant tissue will have some small but non zero effect through network interactions with the core genes (these genes are termed “peripheral”). Hence under this framework phenotypic variation caused by the expression of core genes can be partitioned into cis regulatory effects on core gene expression and trans regulatory effects on core genes mediated through peripheral genes (Xuanyao Liu, Li, and Pritchard 2019). As trans effects on expression are expected to cumulatively contribute a large fraction of the heritability of gene expression, the model suggests that they should too explain the majority of trait heritability (Xuanyao Liu, Li, and Pritchard 2019).

One important advancement that the omnigenic model offers over standard polygenic models is a conceptual framework for interpreting pleiotropy, the phenomenon where genetic variants affect multiple traits. Where two traits have no shared or co-regulated core genes, but simply share peripheral genes with no direct effect on trait biology, we

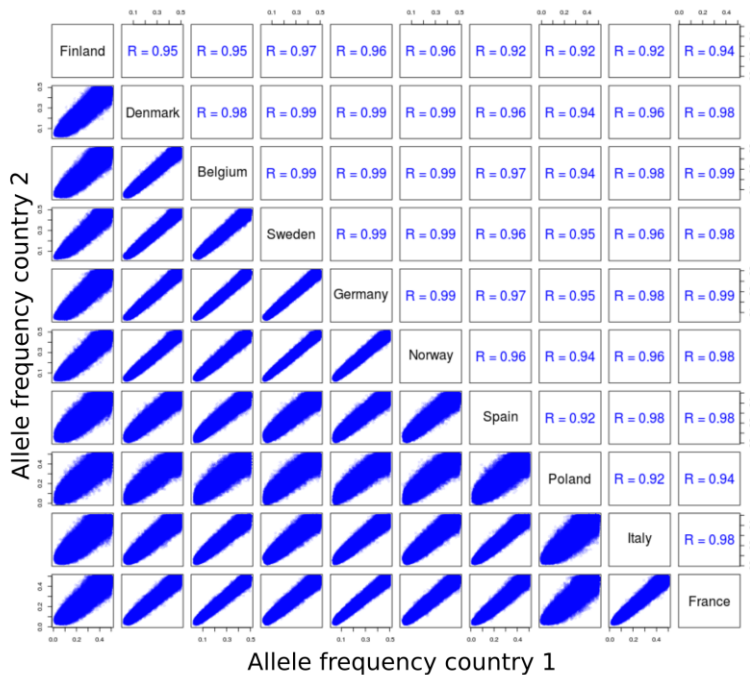
expect these sporadically pleiotropic variants not to correlate in terms of direction of effect. In contrast where the core genes of both traits are shared or co-regulated, indicating some degree of biological overlap, then the pleiotropic variants can have correlated directions of effect on the two traits. This means methods such as bivariate LD score regression (Bulik-Sullivan, Finucane, et al. 2015) which look at the signed correlation of SNP effects across multiple traits should preferentially identify sources of biologically informative pleiotropy, as opposed to sporadic pleiotropy.

Polygenic risk has been shown to correlate for a large number of diseases both globally (Bulik-Sullivan, Finucane, et al. 2015; Zheng et al. 2017) and locally (Shi et al. 2017), suggesting shared genetic components influence multiple diseases, which may point to shared causal genetic mechanisms, or result from random pleiotropy with no biological overlap. On top of this shared polygenic architecture spread across many low effect SNPs, a large number of variants appear to have a statistically significant impact on more than one trait (Watanabe et al. 2019). These observations imply that to fully understand the genetic contribution of variants to disease, traits are best studied together rather than in isolation, a paradigm which is being adopted by multi trait meta analyses methods such as multi-trait analysis of GWAS (MTAG) (Turley et al. 2018), which leverage correlated summary statistics from multiple traits to enhance power to detect causal SNPs. Proper model specification is crucial to maximise yield from multi-trait analysis, for example MTAG assumes SNPs have a homogeneous effect on each correlated traits studied making it unsuitable in cases where some SNPs only affect some traits. More sophisticated multivariate methods such as Genomic Structural Equation Modelling (Genomic SEM) (Grotzinger et al. 2019) allow researchers to specify and compare a range of multivariate genetic architectures, and estimate which SNPs diverge from these models via estimates of heterogeneity, making these methods more flexible. Importantly this allows construction of composite polygenic risk scores for related traits with heterogeneous loci removed, which would be disproportionately predictive of some subtraits. Composite psychiatric polygenic risk scores constructed with Genomic SEM outperformed univariate scores in the UK Biobank (Grotzinger et al. 2019), supporting a general shared polygenic component for psychiatric diseases (see the p-factor model (Caspi et al. 2014)). These composite GWAS could also potentially improve pathway enrichment analyses and point towards shared mechanisms.

### 1.3 - Population structure

Population structure refers to the existence of subpopulations in a dataset or population which (in a genetic context) are distinguishable based on differences in their observed genotypes (Hellwege et al. 2017). Population structure is caused by non-random mating, potentially due to geographic isolation or assortative mating arising from cultural differences. Once subpopulations are isolated, mechanisms such as genetic drift, mutation and selection act independently on resulting groups leading to their genetic differentiation over a number of generations (however it is important to note that populations are rarely entirely split, and migration slows the effects of these processes; see Figure 1.3 for the scale of differentiation in Europe). Differentiated populations harbour systematic differences in allele frequencies as a result of these processes, which are measurable using metrics such as the average fixation index ( $F_{ST}$ ) across a set of unlinked markers (Holsinger and Weir 2009; Hellwege et al. 2017). Markers with large frequency differences between subpopulations can drive spurious associations in genetic association studies where sampling of cases and controls from subpopulations is uneven. This systematic difference in allele frequencies between cases and controls due to ancestry or population structure is referred to as population stratification (Price et al. 2006). Correction for population stratification is essential for robust large scale genome wide association studies (Marchini et al. 2004; Campbell et al. 2005) which has prompted the development of a range of techniques to quantify it and adjust for it. Standard methods of identifying population structure include the use of ancestry informative markers (AIMS), dimensional reduction techniques such as principal component analysis (PCA) (Price et al. 2006; Patterson, Price, and Reich 2006) and model based clustering methods such as such as STRUCTURE and ADMIXTURE (Pritchard, Stephens, and Donnelly 2000; Alexander, Novembre, and Lange 2009).

### Between country allele frequency correlations



**Figure 1.3: Subtle allele frequency differences between European countries**  
 Correlation plots comparing allele frequencies for ~110,000 independent genotype markers between 10 European countries (Data from Sawcer et al. (Sawcer et al. 2011)) provide an intuition of how subtle the differences resulting from forces such as genetic drift are in Europe. While some variants show large frequency differences between countries, the majority of variants have broadly similar allele frequencies. This is visible in the strong linear relationships seen in the bottom panels and high pairwise correlation coefficients seen in the top panels. In spite of how subtle these differences are, however, these populations are distinguishable via PCA and other methods discussed for detecting population structure in this section.

#### 1.3.1 - Ancestry informative markers:

AIMs are markers with extremely large frequency differences between populations. A major advantage of AIMs is they enable identification of broad ancestry differences using very few markers, allowing for correction for population structure in association studies that have not generated dense genotype data. Notably continental ancestry can be accurately estimated using as few as 128 AIMs (Kosoy et al. 2009), demonstrating their power. A set of 300 AIMs was shown to outperform self-reported ancestry in correcting for population stratification in a cohort of European Americans (Price, Butler, et al. 2008), confirming their potential for use in small association studies. However while AIMs can be identified that can efficiently separate many ancestry groups, for large association studies sampling from a range of countries, with potentially unknown ancestry groups (making



marker selection tricky), it is more prudent to use a much larger set of markers to detect and correct for population structure.

### 1.3.2 - Principal component analysis:

Principal Component Analysis (PCA) of genetic data is a model-free approach for detecting population structure, which works by decomposing high dimensional genetic variation from many markers into a smaller number of uncorrelated orthogonal dimensions explaining the largest portion of variance between samples. When data are properly prepared, the first few PCs generally describe the ancestral similarities and differences between samples with no prior assumptions. Use of PCA to describe ancestral clines of allele frequencies and their relationship to geography was first proposed by Cavalli-Sforza (Menozzi, Piazza, and Cavalli-Sforza 1978). The method has famously since been used on SNP genotype data to demonstrate that genes mirror geography in Europe (Novembre et al. 2008), and is regularly exploited as a continuous covariate to correct for population structure in GWAS analysis due to its relative simplicity and effectiveness (Price et al. 2006). As it reduces the sources of variation to a few key components, PCA can be used for many thousands of markers without prior knowledge of informativeness, hence making discovery of AIMs redundant in large GWAS datasets. PCA, however, assumes independence of variables, and as such the method requires pruning of markers in linkage disequilibrium (LD) to obtain an approximately independent SNP set. This pruning process potentially leads to the loss of information about population structure that might be captured by patterns of correlation or linkage disequilibrium between markers. In addition to model-free approaches like PCA, several methods for characterising and correcting population structure use explicit models to characterise subgroups in a genetic dataset.

### 1.3.3 - Model based clustering: STRUCTURE and ADMIXTURE:

Model based clustering approaches assume that  $K$  genetic sub-populations exist in a heterogeneous population (or dataset), with each defined by its own set of distinct allele frequencies. Using this assumption they aim to assign individuals to one (no admixture model) or a combination (admixture model) of those sub-populations based on the probability of their observed genotypes in the sub-populations. For example in a no-admixture model, the likelihood of an individual originating from sub-population  $K$  carrying a given allele is equal to the frequency of that allele in sub-population  $K$ . However to begin with neither the allele frequencies, nor the membership of the sub-populations is known, and estimation of each is conditional on the state of the other. Because of this, methods such as STRUCTURE (Pritchard, Stephens, and Donnelly 2000) initialise parameters of

interest such as sub-population origin of an individual ( $Z$ ) and sub-population allele frequencies ( $P$ ) using priors, then sample new solutions for each parameter sequentially using a Markov Chain Monte Carlo (MCMC) model over many iterations (i.e. when sampling  $P$  on the  $n$ th MCMC iteration, the values of  $Z$  from the  $(n-1)$ th iteration would be used to define the probability distribution  $P$  is drawn from). For a model allowing admixture a third parameter  $Q$  may also be considered, which is the proportion of an individual's genome originating from a population subgroup (Pritchard, Stephens, and Donnelly 2000; Alexander, Novembre, and Lange 2009). ADMIXTURE adopts a similar likelihood model to STRUCTURE, but uses a maximum likelihood approach rather than an MCMC approach, greatly improving speed (Alexander, Novembre, and Lange 2009).

The subpopulations derived from model based clustering can in theory be used to control for population stratification in association testing in a process called Structured Association (SA) (Pritchard et al. 2000). SA tests the null hypothesis that allele frequencies are independent of phenotype within subpopulations against the alternative hypothesis that allele frequencies depend on phenotype in the given subpopulation. As subpopulations are expected to be ancestrally homogeneous, and hence harbour no population stratification, this test should not be biased by stratification. However SA relies on correct assignment of subpopulations, potentially opening it to bias where  $K$  is misspecified. Hence SA can be error prone in situations where estimating  $K$  is difficult, for example where one of the sub-populations is underrepresented. STRUCTURE and ADMIXTURE are also computationally intense, and often cannot converge in cases where  $K$  is greater than 10, limiting their usefulness in correcting population structure where many ancestral groups are present (Lawson et al. 2012). In addition many model-based clustering approaches (for example ADMIXTURE and STRUCTURE) assume independence of markers in the modelled populations (exceptions discussed below), potentially limiting their power to detect patterns of subtle population structure captured by patterns of linked markers.

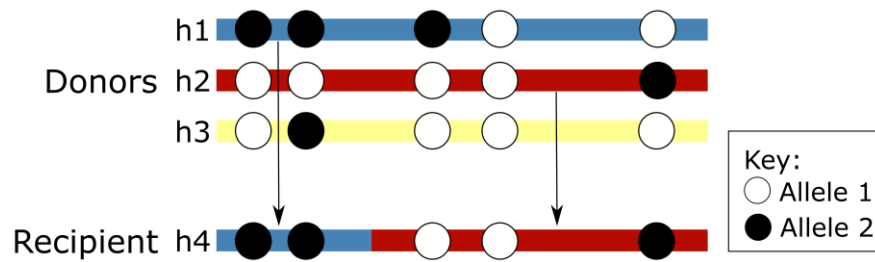
#### 1.3.4 - Haplotype sharing methods:

Markers located on the same chromosome are physically connected and inherited together, unless separated by recombination, in what is known as a haplotype. In a population many haplotypes are shared in common among individuals with shared ancestry, leading to correlation between markers known as linkage disequilibrium (LD). The existence of LD violates the independence assumptions of PCA and many model-based clustering algorithms, potentially rendering them unsuitable for capturing true patterns of population structure (Lawson et al. 2012). Recently a range of methods exploiting linkage information between markers to improve detection of population structure have been developed. One approach, ChromoPainter (illustrated in Figure 1.4), uses a Hidden Markov Model to reconstruct the chromosomes of each individual as a series of haplotypic “chunks” from their nearest neighbouring haplotype in the sample population (Lawson et al. 2012). In doing so ChromoPainter can summarise the ancestral relationships between samples based on the degree of haplotype sharing with other individuals in the sample in a “coancestry matrix”. The model is flexible in that the linked chunks reduce to single markers in cases where no LD is present, making it suitable for use in both densely and sparsely genotyped data. The coancestry matrix contains rich ancestral information that can be projected into PC space to demonstrate population structure or clustered using fineSTRUCTURE (a companion model-based clustering algorithm that uses an MCMC approach to partition the data into distinct homogeneous groups), making it a versatile summary of population structure. Unlike STRUCTURE and ADMIXTURE, fineSTRUCTURE does not rely on a user-specified K, and is capable of splitting data into over 100 groups (>10 groups rarely reaches convergence in STRUCTURE) (Lawson et al. 2012). As well as providing a finer resolution of clustering, fineSTRUCTURE also generates a hierarchical tree representing the relationships between clusters, allowing greater understanding of the history of relationships between these groups. This method has recently been employed to greatly improve our understanding of subtle population structure in a number of countries.

ChromoPainter and fineSTRUCTURE have been used to great effect in detecting subtle regional population structure at a resolution not previously possible in a range of countries to date including Britain, Finland, Ireland (see Chapter 3), Japan, Spain, the Netherlands (see Chapter 4), Italy and France (Leslie et al. 2015; Kerminen et al. 2017; R. P. Byrne et al. 2018; Gilbert et al. 2017; Takeuchi et al. 2017; Bycroft et al. 2019; Raveane et al. 2019; Pierre et al. 2020; R. P. Byrne et al. 2020). These studies all identify subtle differences at short geographic ranges based on haplotype sharing profiles, suggesting

that population structure exists on a finer scale than typically detected using unlinked markers. One important implication of these studies is that PCA and popular model-based clustering algorithms applied to unlinked markers fail to identify subtle genetic structure existing within populations, and might be expected to underperform when used to correct for confounding in association studies. Notably a study of residual stratification in 36 published GWAS studies using LD score regression intercepts (Bulik-Sullivan, Loh, et al. 2015) confirms the limitations of SNP based PCA, revealing residual stratification in 8 studies despite correction using PCs (Bhatia et al. 2016). Additionally polygenic signals of adaptation for traits such as height have recently been shown to be significantly biased, likely due to uncorrected population stratification (Sohail et al. 2019; Berg et al. 2019), in spite of use of standard correction procedures. Recent simulations have also shown that common variant PCA is only suitable for correcting old longstanding population structure, and does not fully correct confounding from complex recent population structure in GWAS (Zaidi and Mathieson 2020). Moreover GWAS of birth location in the UK Biobank, which is essentially a proxy for population stratification, shows significant GWAS associations (Haworth et al. 2019) and genetic correlations with behavioural traits (Cook, Mahajan, and Morris 2020) despite correction with large numbers of SNP PCs. Leveraging the results of fine-scale population genetics based on haplotype sharing between individuals may prove more suitable than SNP-based PCA to address this residual confounding in association studies due to the capacity to capture both recent and deep structure. Moreover, as haplotype sharing approaches tend to identify more recent population structure, they may also be more suitable for capturing and correcting the subtle stratification of rare variants (Mathieson and McVean 2012). Given the rare variant architecture of ALS discussed above (van Rheenen et al. 2016), development of appropriate corrections for differential confounding of rare variants may be warranted to further our understanding of the disease. On top of this the extended role of environmental effects in ALS likely makes careful correction for population structure increasingly relevant for this trait. However, to date the use of haplotype sharing methods to correct population structure in genetic association studies has not been explored in detail, despite their substantially improved power to detect subtle structure. We will explore this untapped application in detail in Chapter 5.

# Chromosome painting



**Figure 1.4: Toy illustration of chromosome painting**

A toy illustration of the basic process of chromosome painting employed by methods such as ChromoPainter (Lawson et al. 2012). Here haplotypes (rectangles) made up of biallelic variants (dots) are shown. The recipient haplotype (h4) can be “painted” as a mosaic of copied “chunks” from the best matching donor haplotypes (h1 and h2), as indicated by the colour pallet and arrows. Each copied “chunk” can be intuitively thought of as an independent instance of recent shared ancestry between the donor and the recipient haplotypes at that locus. This “chunk” sharing provides more information about the relationship of h4 to the donor haplotypes than the independent markers do alone (each donors shares exactly 3 alleles with h4, which is uninformative). The copying process used here is governed by the Li and Stephen’s Hidden Markov Model (HMM) (Li and Stephens 2003) which sequentially moves along the chromosome finding the best matching donor and switching “chunks” based on a parameterised recombination rate. Mismatches are permitted based on a mutation parameter. Figure adapted from Li and Stephens (Li and Stephens 2003) for illustrative purposes

The utility of haplotype sharing in population genetics is far from limited to the detection of population structure or potential sources of confounding in association studies.

Companion methods have been developed which leverage the output of ChromoPainter to date recent admixture events based on the length distribution of haplotypes from admixing sources (Hellenthal et al. 2014), and to estimate the proportion of an individual’s genome most closely related to a reference population (Leslie et al. 2015; Chacón-Duque et al. 2018). Moreover following the recent surge of methods developed to accurately estimate regions of the genome at which unrelated individuals share a common ancestor from SNP data (identity by descent, IBD) (Gusev et al. 2009; B. L. Browning and Browning 2011, 2013a, 2013b), methods have leveraged this haplotypic sharing to make inferences about demographic change in the past 100 generations. Seminal works by Palamara et al developed a framework for describing the relationship between the demographic history of populations and the distribution of IBD segment lengths, enabling inferences to be made about population size fluctuations across time and migrations between populations from observed IBD sharing in populations (Palamara et al. 2012; Palamara and Pe’er 2013). Non-parametric approaches to estimate population growth and decline based on the

distribution of time to most recent common ancestor given an observed segment have also emerged which achieve similar resolution with much shorter runtimes (S. R. Browning and Browning 2015). Aside from general interest, the emergence of such powerful techniques for dissecting recent events in human population history may prove important for contextualising the study of recently arising rare disease causing alleles, as seen in a recent study of haplotype sharing in Finland (A. R. Martin et al. 2018), where the authors showed increased haplotype sharing among individuals sharing rare disease-causing variants. Thus the interface between population genetics and medical genetics is a rich emerging field.

## 1.4 - Aims

The research detailed in this thesis has four major foci which together culminate in the shared aims of improving our understanding of disease genetics (in particular ALS), and its interface with population genetics:

- i.) Understanding of the genetic overlap of ALS with comorbid secondary psychiatric and cognitive phenotypes
- ii.) Evaluating the role of sex in the genetic risk of ALS
- iii.) Characterising finescale population structure and demographic history from medical datasets
- iv.) Investigating novel haplotype sharing methods for addressing GWAS confounding from multiscale population structure.

While this work is important for properly understanding the genetics of ALS and other diseases, it also holds a general interest for historians as haplotype-based methods allow us to date and characterise recent migrations, estimate changes in population size and reveal nuanced pictures of recent history. In this thesis I will outline the results of my work on characterising the shared genetic architecture of ALS and secondary complex traits using GWAS summary statistics and exploring its heterogeneity across sexes using individual data (Chapter 2) which contributes towards my first aim and second aims; I will discuss my work dissecting the local population structure and recent demographic change in Ireland and the Netherlands using haplotype sharing and other complementary methods (Chapters 3 and 4) which contribute towards my third aim; finally I will investigate

the application of haplotype sharing matrixes on large ( $n=36k$ ) GWAS datasets to correct for both finescale and broadscale inflation from population stratification and other technical artifacts (Chapter 5) which will contribute towards my fourth aim. Notably large samples from homogeneous populations such as the UK Biobank demonstrate substantial population structure (Diaz-Papkovich et al. 2019), suggesting that there is a great importance to understand and correct this local structure if we are to robustly study disease associations in coming years where similar large cohorts are likely to emerge as central resources.

# Chapter 2 - The Extended Genetic Architecture of ALS

## 2.1 - Introduction

### 2.1.1 - Background

While ALS is primarily a motor system disease, extra motor symptoms are common, with cognitive and behavioural changes comorbid in between 30%-50% of patients (Rippon et al. 2006; van der Hulst, Bak, and Abrahams 2015; Beeldman et al. 2016; van Es et al. 2017). Systematic meta-analysis has shown that all cognitive domains except visuoperceptive functions are significantly affected in ALS patients compared to controls (Beeldman et al. 2016), including significant effects on fluency, social cognition, verbal memory, executive function and language. Additionally comorbidity with frontotemporal dementia (FTD) is high in ALS, with ~15% of ALS patients fulfilling the Neary criteria for diagnosis with FTD (Phukan et al. 2012). These extra motor symptoms form an important axis of the disease which can impact caregiver burden, and have a significant negative effect on the success of life-prolonging treatments such as enteral nutrition (EN) and non-invasive ventilation (NIV) (Chiò et al. 2012). Interestingly recent emerging evidence suggests that lower cognitive performance may have a shared genetic component with ALS: higher polygenic risk scores for ALS were negatively associated with measures of verbal numeric reasoning in the UK Biobank (Hagenaars et al. 2018); and LD score regression of years of schooling as a proxy measure of cognition versus ALS showed a nominally significant negative correlation with ALS in a separate study (Bandres-Ciga et al. 2019) (though the latter was not significant after multiple testing). However it remains unclear whether this genetic overlap has a causal relationship with ALS, and what the extent of pleiotropy between these traits is. The widespread nature of extra motor cognitive symptoms and evidence for a possible involvement of genetic factors warrants further work characterising the genetic overlap between ALS and cognitive function.

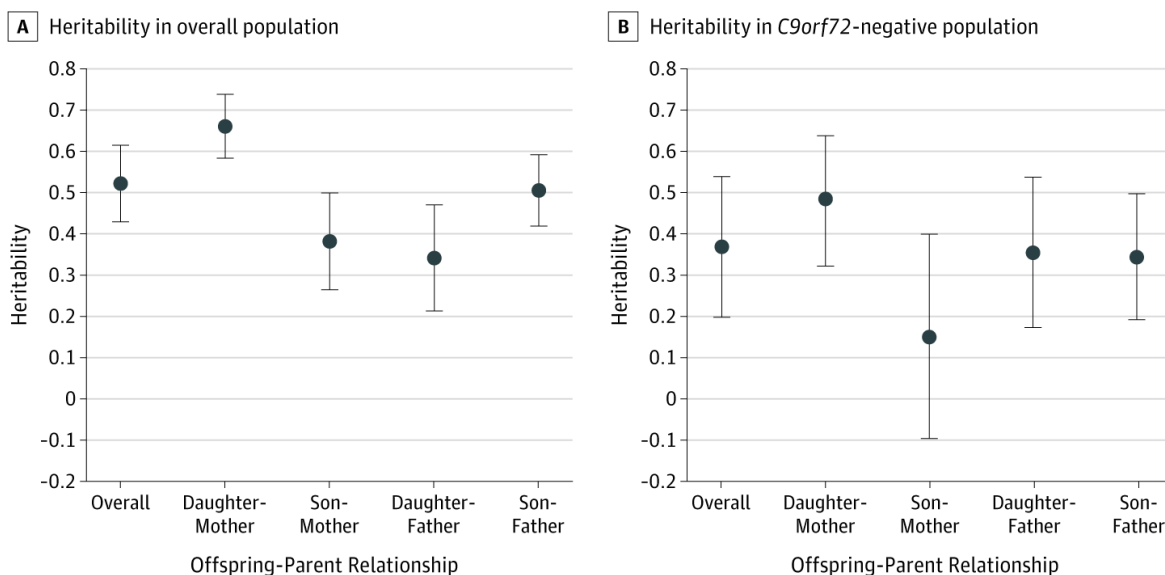
Similarly to cognitive and behavioural symptoms, neuropsychiatric diseases also appear to be widely concomitant with ALS. Neuropsychiatric disease is significantly enriched in ALS kindreds (S. Byrne et al. 2013), and a prior diagnosis of depression, bipolar disorder, schizophrenia, or anxiety has been strongly linked with a higher risk of diagnosis with ALS in subsequent years (Turner et al. 2016). ALS has recently been shown to have a genetic overlap with schizophrenia (R. L. McLaughlin et al. 2017), raising the possibility that shared genetic risk factors drive these associations between ALS and other psychiatric traits (though no additional links with secondary psychiatric traits were identified in the



aforementioned study, likely due to power). Since these studies, larger GWAS have emerged for many of the studied secondary traits, alongside reasonably powered GWAS for untested traits such as anxiety (Otowa et al. 2016) and post-traumatic stress disorder (Duncan et al. 2018) and better measures of cognitive performance (Davies et al. 2018; J. J. Lee, Wedow, et al. 2018) motivating an updated investigation of the genetic overlap of ALS with psychiatric and cognitive traits. Moreover, methodological advances including powerful multi-trait methods to improve power and detect pleiotropic loci (Turley et al. 2018; Grotzinger et al. 2019) and models for determining the causality of a genetic correlation (O'Connor and Price 2018) have been developed, meaning we are currently in a strong position to untangle the nature of these genetic overlaps. As a central aim of this chapter we will dissect the nature of the genetic overlap between ALS and cognitive and psychiatric traits. A firm understanding of the genetic basis of the relationship between ALS and these secondary traits will likely improve our understanding of disease aetiology and potentially inform treatment of extra motor symptoms.

In addition to extramotor symptoms, the role of sex in ALS, and its interaction with genotype is well documented throughout the literature (Trojsi et al. 2020), but rarely accounted for in the design of genetic studies. The genetic risk of ALS appears to be partially sex-dependent, as evidenced by sex-specific heritability estimates, which show increased heritability in mother-daughter pairings (Figure 2.1) (Ryan et al. 2019). This sex by gene interaction can have substantial impact on disease outcome, as observed in the lower survival rates of males with spinal onset ALS harbouring the *C9orf72* repeat expansion compared to females (Rooney, Fogh, et al. 2017). Additionally sex appears to affect the onset of the disease, with higher observed rates of ALS in males pre-menopause age (Manjaly et al. 2010), suggesting a potential protective role of female sex hormones in ALS risk. Conversely, testosterone, the male sex hormone may contribute to ALS risk; It has been suggested that high levels of pre-natal testosterone, regardless of sex, increase ALS risk, likely by conferring motor neurone vulnerability (Vivekananda et al. 2011) (measured using index-to-ring finger ratio as a proxy). As ALS is a multifactorial disease which appears to be governed by a multistep process (Al-Chalabi et al. 2014; Chiò et al. 2018), requiring more than one insult (~6) to develop across a patient's lifespan, it is apparent that environmental and lifestyle exposures contribute to disease onset in addition to an individual's baseline genetic or biological risk. It is possible thus that a combination of sex-differentiated hormonal exposures, lifestyle exposures and biology could all modify the risk for developing ALS conferred by genetic variants in men and women. In a GWAS context, this would lead to an apparent sex dependent heterogeneity in the effect sizes of risk loci in males and females that might decrease

power to detect these loci in combined analysis, particularly where cohorts are unbalanced for sex. Conceptually, this is analogous to studying a composite of two distinct diseases with shared phenotype but differing genetic risk factors, a scenario which has been shown by simulation to lead to lower power to detect risk loci and deflated estimates of SNP heritability (Wray and Maier 2014). This may potentially explain some of the discordance between pedigree-based heritability estimates (~52% (42%-62%) Figure 2.1 (Ryan et al. 2019)) and SNP based estimates of heritability in ALS (~8% (van Rheenen et al. 2016)). Hence, given the potential gain in power, and importance to the proper design of future GWAS, the second central aim in this chapter is to assess the impact of sex on GWAS studies of ALS, and address this potential heterogeneity using a sex stratified GWAS design for ALS. In doing so we provide further evidence for a sex specific architecture and highlight the potential power increases for future large ALS GWAS by stratifying by sex.



**Figure 2.1: Stratified ALS heritability estimates for Irish population.**

Mean pedigree-based heritability estimates for ALS in A) the whole population studied and B) The population not harbouring the C9orf72 mutation. Estimates were calculated in the overall population, and independently for subpopulations with different parent-offspring pairings (Daughter-Mother, Mother-Son, Father-Daughter and Son-Father). Notably Mother-Daughter pairings show the highest average heritability, while heritability calculated in opposite sex parent-offspring pairings (Mother-Son and Father-Daughter) is lowest, suggesting a sex-specific component to heritability.

Figure reproduced from Ryan et al. (Ryan et al. 2019) for illustration purposes.

Figure source: <https://jamanetwork.com/journals/jamaneurology/fullarticle/2737804>

In addition to showing a higher heritability in females than males, the results of Ryan et al. (Ryan et al. 2019) (Figure 2.1) show a difference in ALS heritability estimates in like-sexed (e.g. Mother-Daughter and Father-Son) and unlike-sexed (e.g. Mother-Son and Father-Daughter) parent-offspring pairings, suggesting that ALS may not be perfectly genetically correlated between sexes. Following Falconer (Falconer 1967), we can estimate the pedigree-based genetic correlation between sexes with the formula:

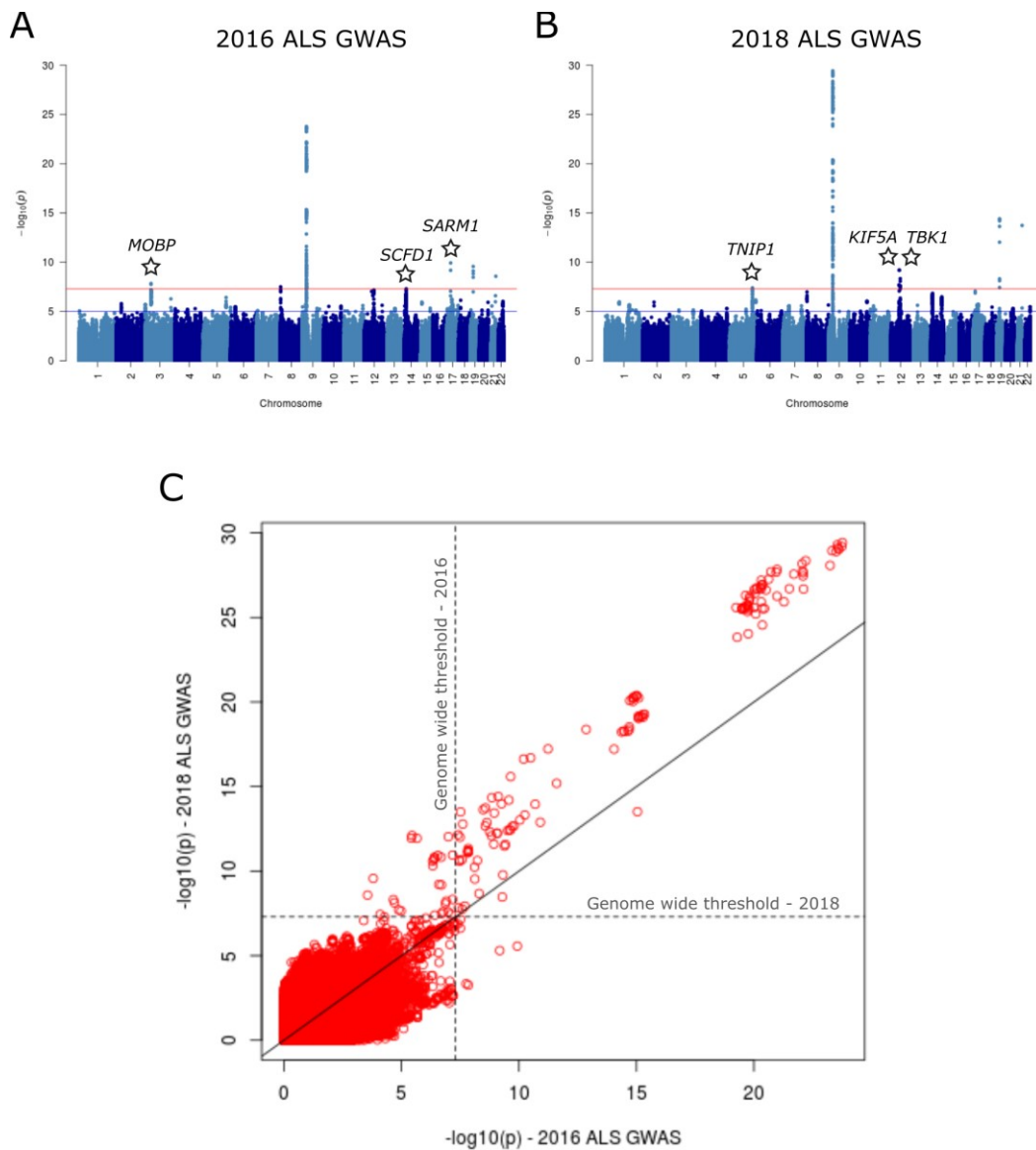
$$r_g = \frac{h_m \cdot h_f(\text{Unlike})}{h_m \cdot h_f(\text{Like})}, \quad (2)$$

where  $r_g$  is the genetic correlation between sexes,  $h_m$  and  $h_f$  are the square routes of the heritability estimates for males and females respectively in like-sexed and unlike-sexed pairs. This returns an estimate of 62.8% (95% c.i: 48%-73.2% ) for the pedigree-based genetic correlation between males and females for ALS providing empirical evidence that the trait is explained by different genetic factors in males and females and further motivating the study of potential sex-based heterogeneity in ALS genetics in a GWAS context.

Further supporting the potential importance of exploring sex based heterogeneity in ALS genetics, there is mounting evidence of heterogeneity in the genetic architecture of ALS across recent GWAS. In the past four years (2016-2020) two moderately sized GWAS for ALS (van Rheenen et al. “2016 ALS GWAS”; n=36,052 (van Rheenen et al. 2016) and Nicolas et al. “2018 ALS GWAS”; n=80,610 (Nicolas et al. 2018)) have emerged, enabling more in depth study of the genetic architecture of ALS. The first of these GWAS (n=36,052) established firm evidence that ALS is a polygenic trait, showing that SNP-based heritability is roughly 8.5% and is proportional to chromosome size (van Rheenen et al. 2016), suggesting that larger GWAS should return additional trait-associated loci. However, the subsequent 2018 ALS GWAS (n=80,610) identified very few additional loci (*KIF5A*, *TNIP1*, *TBK1*), and in fact lost association signal at an equal number of loci (*MOBP*, *SARM1* and *SCFD1*), despite a doubling in sample size, meaning GWAS in ALS may have diminishing returns. The low yield of novel variants may be explained partially by the concentration of heritability in low frequency variants observed in the 2016 ALS GWAS (van Rheenen et al. 2016), suggesting a rare variant architecture for the disease that may require more sensitive methods such as whole genome sequencing to identify further causal variants. However the loss of association at *MOBP*, *SARM1* and *SCFD1*, each tagged by common variants, suggests that additional factors are at play.

One possible answer to this discrepancy is study design, as the initial study applied a powerful linear mixed model (LMM) approach to control for population structure, enabling

the joint analysis of all samples, while the subsequent study used a meta-analysis design, requiring splitting of cohorts and enforcing more stiff assumptions of homogeneity of effect across cohorts. The use of a meta-analysis may have resulted in lower power in the 2018 ALS GWAS, potentially explaining the loss of association at these variants, however the doubling in sample size appears to more than compensate for this (Figure 2.2; the 2018 ALS GWAS is better powered on average). Instead it is more probable that the sensitivity of meta analysis to heterogeneity of effect is responsible here, which would manifest if *MOBP* and *SCFD1* have heterogeneous effects across patients. Investigating the sampling schemes of each reveals that there is a clear discrepancy between these GWAS in terms sex balance of case and control cohorts; While the 2016 ALS GWAS has ~50% female controls and ~40% female cases (van Rheenen et al. 2016), which is close to balanced, the 2018 ALS GWAS has ~60% female controls and only ~40% female cases (Nicolas et al. 2018) making it more heavily sex biased. In the case of sex-dependent heterogeneity in ALS, where a variant contributed to ALS risk in only males for example, we would expect the frequency of this risk allele to be equal in female cases and controls, and hence such a sex imbalance could substantially increase the risk allele frequency in controls, diluting the signal in the combined GWAS. This would be particularly impactful in the 2018 ALS GWAS where only 40% of the controls are male. Hence a sex-dependent architecture as discussed above could reconcile some of the differences between these GWAS.



**Figure 2.2: Comparison of power and correlation between ALS GWAS.**

A-B.) Manhattan plots for overlapping SNPs in the A.) 2016 ALS GWAS and B.) 2018 ALS GWAS with significant hits unique to one (of the two) GWAS marked with stars. Notably three loci passing significance in the 2016 ALS GWAS do not replicate in the larger 2018 ALS GWAS. C.) A scatterplot comparing  $-\log_{10}(p)$ -values from the 2016 (x-axis) and 2018 (y-axis) GWAS shows that they are highly correlated. The 2018 ALS GWAS has greater inflation for the majority of shared hits as indicated by the deflection from the 1:1 line, suggesting this GWAS has greater power. Notably a handful of variants pass genome wide significance thresholds (dashed lines) in one GWAS but not the other (identifiable on the Manhattan plots above).

Data: van Rheenen et al. 2016 (van Rheenen et al. 2016) and Nicolas et al. 2018 (Nicolas et al. 2018)

### 2.1.2 - Research aims

This chapter presents work carried out on ALS GWAS summary statistics, and individual level data with the global aim of expanding our understanding of the genetic architecture of ALS and identifying further putative variants by leveraging known overlaps with other phenotypes and the existence of subgroups in the data. We approach this global aim through the lens of two minor aims:

i.) To interrogate the nature and extent of the shared genetic architecture between ALS with cognitive and neuropsychiatric traits;

and

ii.) To interrogate the possibility of a sex-differentiated genetic architecture of ALS and identify associated variants.

Through this research we hope to identify putative ALS variants missed by standard single trait models and those ignoring potential sex-mediated heterogeneity, and better understand potential subgroups of the disease.

## 2.2 - Methods

### 2.2.1 - Estimating genetic correlation between ALS and psychiatric traits

GWAS summary statistics were downloaded for ALS (van Rheenen et al. 2016), schizophrenia (Ruderfer et al. 2018), bipolar disorder (Ruderfer et al. 2018), major depressive disorder (Wray et al. 2018) (minus 23&me samples), ADHD (Demontis et al. 2019), anxiety (Otowa et al. 2016), PTSD (Duncan et al. 2018) and cognitive performance (Davies et al. 2018; J. J. Lee, Wedow, et al. 2018) to test for genetic correlations between ALS with extended neuropsychiatric and cognitive phenotypes. Summary statistics were formatted for use with the LD score regression software using the “munge\_sumstats.py” script provided with the software, retaining only SNPs present in the HapMap phase 3 dataset, which are typically well imputed. We ran bivariate LD score regression to estimate the genetic correlation between ALS and all secondary traits using precomputed LD scores for European individuals ([https://data.broadinstitute.org/alkesgroup/LDSCORE/eur\\_w\\_ld\\_chr.tar.bz2](https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2)), with both constrained (following McLaughlin et al. (R. L. McLaughlin et al. 2017)) and free intercepts.

### 2.2.2 - Partitioned heritability analysis

We ran stratified LD score regression to test for enrichment of ALS (and secondary trait) heritability in functional annotations (Finucane et al. 2015) and genes highly expressed in specific cell types (Finucane et al. 2018). For this analysis the annotations for the functional enrichment were downloaded from <https://data.broadinstitute.org/alkesgroup/LDSCORE/> and annotations for the tissue specific gene expression analysis were downloaded from: [https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC\\_SEG\\_ldscores/Multi\\_tissue\\_gene\\_expr\\_1000Gv3\\_ldscores.tgz](https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/Multi_tissue_gene_expr_1000Gv3_ldscores.tgz). For the tissue specific analysis the functional annotation was conditioned upon as a baseline model. A follow up analysis was also performed to look at enrichment of heritability in an annotation of brain region specific gene expression due to a significant enrichment in the frontal lobe in the multi tissue analysis. For this we used the GTEx brain specific annotation ([https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC\\_SEG\\_ldscores/GTEx\\_brain\\_1000Gv3\\_ldscores.tgz](https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/GTEx_brain_1000Gv3_ldscores.tgz)). For all analyses including ALS the intercept was constrained to 1 as in McLaughlin et al. (R. L. McLaughlin et al. 2017).

### 2.2.3 - Genomic SEM – Multi-trait models of ALS and psychiatric disorders

Genomic structural equation modelling (Grotzinger et al. 2019) (GenomicSEM; <https://github.com/MichelNivard/GenomicSEM>) was used to determine how well a single common factor model could be fit to the genetic covariance matrix of ALS (van Rheenen et al. 2016), schizophrenia (Ruderfer et al. 2018), bipolar disorder (Ruderfer et al. 2018), major depressive disorder (Wray et al. 2018), ADHD (Demontis et al. 2019), anxiety (Otowa et al. 2016), PTSD (Duncan et al. 2018). Summary statistics were formatted using the internal “munge” function of GenomicSEM, keeping only variants from the HapMap3 panel, and excluding the MHC locus. We ran multivariable LDSC to produce the genetic covariance (S) and sampling covariance (V) matrices for these summary statistics using the “ldsc” function in GenomicSEM. We ran exploratory factor analysis using factanal on the smoothed genetic covariance matrix to identify possible factor models, followed by fitting these models in confirmatory factor analysis in genomic SEM to determine the most likely number of latent factors explaining the trait covariance. We assessed relative model fit using Akaike’s information criteria (AIC; smaller is better) and the comparative fit index (CFI; larger is better). Based on the exploratory factor analysis we fit both a common factor model and a correlated two factor model to these matrices using the diagonally weighted least squares (DWLS) estimation method. Due to a Heywood case (negative residual variance) for MDD in the two factor model we constrained its residual variance to be above zero.

### 2.2.4 - Detecting pleiotropic loci with multi-trait analysis

MTAG (Turley et al. 2018) was run using standard settings to identify pleiotropic loci between summary statistics for ALS (van Rheenen et al. 2016; Nicolas et al. 2018) and correlated traits schizophrenia (Ruderfer et al. 2018), bipolar disorder (Ruderfer et al. 2018) and cognitive performance (J. J. Lee, Wedow, et al. 2018). Verbal numeric reasoning (Davies et al. 2018) was excluded from the analysis as the summary statistics did not contain an entry for the effect allele frequency which is required to run MTAG. As bipolar disorder and schizophrenia have a strong known genetic correlation (Bulik-Sullivan, Finucane, et al. 2015) we ran a combined MTAG analysis for these traits with ALS to improve power, while ALS and cognition were run as a separate analysis. Summary statistics were formatted to the MTAG format, in R, and cleaned and standardised within the MTAG program, filtering for minor allele frequency (>1%), variants that were not SNPs, strand ambiguous variants and flipping SNPs with non-concordant effect alleles. To evaluate the reliability of our results we ran maximum false discovery



rate (maxFDR; described in the MTAG paper (Turley et al. 2018)) for each MTAG run. Variants were filtered to those achieving genome-wide significance in MTAG runs for both ALS datasets for each set of the secondary traits to alleviate false positives potentially caused by the relatively low sample sizes in these datasets. We clumped variants achieving genome wide significance in the multi-trait analysis and recorded the closest genes to the index variants using plink v1.9 (--clump).

We also applied the conditional false discovery rate (cFDR) method (Andreassen et al. 2013) to consolidate and supplement the multi-trait associations found using MTAG. This method tests for association in a principal phenotype conditional on levels of association with a second conditional phenotype, leveraging the expectation of pleiotropy between the traits. The cFDR is defined here as the probability that a SNP is null for the principal phenotype given its p-value is below a set of thresholds in both the conditional and principal phenotypes. We used ALS (van Rheenen et al. 2016) as the principal phenotype here and the correlated traits from above as conditional phenotypes. To mirror the MTAG analysis, where ALS was analysed alongside bipolar and schizophrenia simultaneously we used the summary statistics for a joint GWAS of both bipolar and schizophrenia versus controls (Ruderfer et al. 2018). Variants passing a cFDR threshold (cFDR<0.01) were clumped using plink v1.9 ( $r^2 = 0.5$ ) to identify independent loci, and annotated with the nearest gene.

### 2.2.5 - Characterising heterogeneity in ALS genetics by sex

To test the hypothesis of a gene by sex interaction in ALS we followed a methodology similar to Tropf et al. and Robinson et al. (Tropf et al. 2017; Robinson et al. 2017) and used GCTA GREML (Yang et al. 2011) to fit a genotype-covariate interaction (GCI-GREML) model (NB: This is implemented in the --gxe switch in GCTA but is not strictly a gene by environment interaction). We ran this analysis on autosomal individual level data from the 2016 ALS GWAS (van Rheenen et al. 2016), partitioning by MAF given that a MAF-dependent architecture was seen in the original ALS GWAS paper, and fitting 10 principal components (PCs) as covariates. For this analysis genetic relationship matrices (GRMs) were constructed from all SNPs passing QC in the original GWAS (van Rheenen et al. 2016) (note: only autosomal SNPs are present in this data). We used the likelihood ratio test in GREML to compare the model with gene by sex interaction to the baseline model with no interaction.

To further characterise the role of sex in the genetic architecture of ALS we separated the data by sex and ran two independent mixed linear model GWAS for ALS (male-only and

female-only) using the --mlma switch in GCTA (Yang et al. 2011). We ran this analysis using the leave one chromosome out approach, whereby each SNP was run in a mixed linear model including a GRM from all chromosomes excluding the chromosome where the target SNP resided. Summary statistics from these GWAS were scanned for “sex-specific” association with ALS, and difference in effect size and direction across sexes (“sex-difference”) using two scanning approaches proposed in a previous work on sex differences in anthropomorphic traits (e.g. height) (Randall et al. 2013):

i.) In the “sex-specific” scan, association p-values from the sex-specific ALS GWAS were scanned for significant association in one sex, but not the other, correcting for multiple testing using the Benjamini-Hochberg procedure accepting SNPs at a 5% FDR (R function p.adjust). Here a variant was classified as suggestively sex-specific if it passes the 5% FDR p-value threshold in one but not the other sex and strongly sex-specific if it passes the 5% FDR threshold in one sex and doesn’t reach nominal significance ( $p < 0.05$ ) in the other sex. For our two datasets across 7.3 million SNPs the p-value threshold corresponding to a 5% FDR was  $p = 9.03 \times 10^{-7}$ .

ii.) For our “sex-differentiation” scan we are interested in finding variants that have substantially different effects on ALS risk in males and females (perhaps even opposite direction of effect). The degree of difference in GWAS effects in males and females was assessed using a test statistic from Randall et al (Randall et al. 2013):

$$Z_{diff-beta} = \frac{b_{female} - b_{male}}{\sqrt{(se(b_{female}))^2 + se(b_{male})^2 - 2 \cdot r \cdot se(b_{female}) \cdot se(b_{male})}} \sim N(0,1), \quad (3)$$

where  $b_{female}$  and  $b_{male}$  are the effect sizes of the SNP in the female-only ALS GWAS and male-only ALS GWAS and  $r$  is the Spearman correlation between the male and female effect sizes. Variants were tested at 5% FDR.

SNPs passing either scan were then clumped using plinkv1.9 (--clump) to retain only the top associated SNP per LD block (250kb range,  $0.5 r^2$ ). We then found the gene/genes closest to these lead SNPs using a hg19 gene list with plinkv1.9 (--clump-range), which we cross-checked with UCSC.

To look at global differences in architecture we ran univariate LD score regression for the male and female only ALS GWAS to test the relative heritability of ALS in males and females, using precomputed LD scores as above

([https://data.broadinstitute.org/alkesgroup/LDSCORE/eur\\_w\\_ld\\_chr.tar.bz2](https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2)), and HESS as described below. For conversion to the liability scale we used the average lifetime risk of developing ALS (1 in 400) (Johnston et al. 2006) as the prevalence, following previous publications using the dataset (van Rheenen et al. 2016; R. L. McLaughlin et al. 2017).

Finally we constructed GRMs using HapMap3 SNPs for the male only, female only and full dataset using GCTA (Yang et al. 2011; S. H. Lee et al. 2011) to estimate the sex-specific heritability in males and females based on the individual-level data (GREML), and the genetic correlation between males and females (bivariate GREML). We first ran GREML (S. H. Lee et al. 2011) on the sex specific GRMs, fitting 10 PCs as covariates. We also split the data by chromosome, and minor allele frequency and estimated heritability by sex per chromosome and allele frequency bin to compare enrichment in each sex, and further explore the sex specific architecture of ALS.

To compare heritability estimates between sexes we calculated a z-score for the difference in estimates using the following equation :

$$Z_{diff-heritability} = \frac{h_{female}^2 - h_{male}^2}{\sqrt{se(h_{female}^2)^2 + se(h_{male}^2)^2}} \sim N(0,1) \quad (4)$$

#### 2.2.6 - Local heritability estimation and contrast polygenicity

HESS (0.5.4-beta; <https://github.com/huwenboshi/hess>) was run on summary statistics for ALS, schizophrenia and bipolar disorder to estimate local heritability (shared and unshared) and compare the degrees of polygenicity between these associated traits. Additionally we ran HESS on the female-specific and male-specific GWAS for ALS described above to see if they had differing degrees of polygenicity. Summary statistics were first formatted using the munge.py script supplied with the LD score regression software, and the necessary CHR and BP columns required by the software were added. For all HESS analyses an external reference panel composed of European samples from the 1000 Genomes Project was used to approximate the LD matrix between SNPs in the summary statistic datasets

(<https://ucla.box.com/shared/static/l8cjb15jsnghhcn0gdej026x017aj9u.gz>) and

approximately independent loci previously defined in Berisa et al. (Berisa and Pickrell 2016) were used (<https://bitbucket.org/nygcresearch/ldetect-data/src/master/EUR/>).

Where a partition defined this way had no SNPs in one or more datasets we excluded it from analysis including those traits. When estimating local genetic covariance between

traits, we assumed no sample overlap between datasets, as we had insufficient data to estimate actual overlap.

### 2.2.7 - Latent Causal Variable Analysis: Understanding the causal relationship between ALS and secondary traits

We ran Latent Causal Variable Analysis (O'Connor and Price 2018) on summary statistics for ALS (van Rheenen et al. 2016; Nicolas et al. 2018) with genetically correlated traits cognitive performance (Davies et al. 2018; J. J. Lee, Wedow, et al. 2018), schizophrenia and bipolar disorder (Ruderfer et al. 2018) to identify whether there was a causal genetic relationship between the traits and, if so, determine the direction of causality. Summary statistics were formatted using the "munge\_sumstats.py" script provided with the LDSC software prior to running the LCV software (<https://github.com/lukejoconnor/LCV>) with default settings. We report results for the analysis of the 2018 ALS GWAS with secondary traits as the 2016 ALS GWAS produced a heritability Z score below the recommended value ( $Z > 7$ ) for use with this software.

### 2.2.8 - Functional annotation and analysis

We used the SNP2GENE function from Functional Mapping and Annotation of Genome-Wide association Statistics (FUMA) (Watanabe et al. 2017) to test for gene based associations, gene ontology enrichments and tissue enrichments for i.) the ALS sex differentiation scan and ii.) the multi-trait analyses (MTAG only). Both analyses were run using default settings, outputting p-values for gene analysis, gene-set analysis and tissue expression enrichment (GTEx v8, 54 tissues) run in MAGMA (de Leeuw et al. 2015).

As the cFDR multi-trait analysis accepts variants passing an FDR threshold of 0.01, which is too high for FUMA SNP2GENE analysis (Maximum  $P < 1e-05$ ), we instead input genes closest to the independent SNP hits for these scans (Table 2.3). Similarly the method for our male- and female-specific scans complicated SNP2GENE analysis, given each sex-specific analysis required exclusion of variants which are significant at two thresholds in the other sex, hence we also input genes closest to independent hits in these scans (Table 2.7) to FUMA's GENE2FUNC function. We used default settings for this analysis and identified enrichment in tissue-specific expression and gene ontologies to characterise functional features of these genesets.

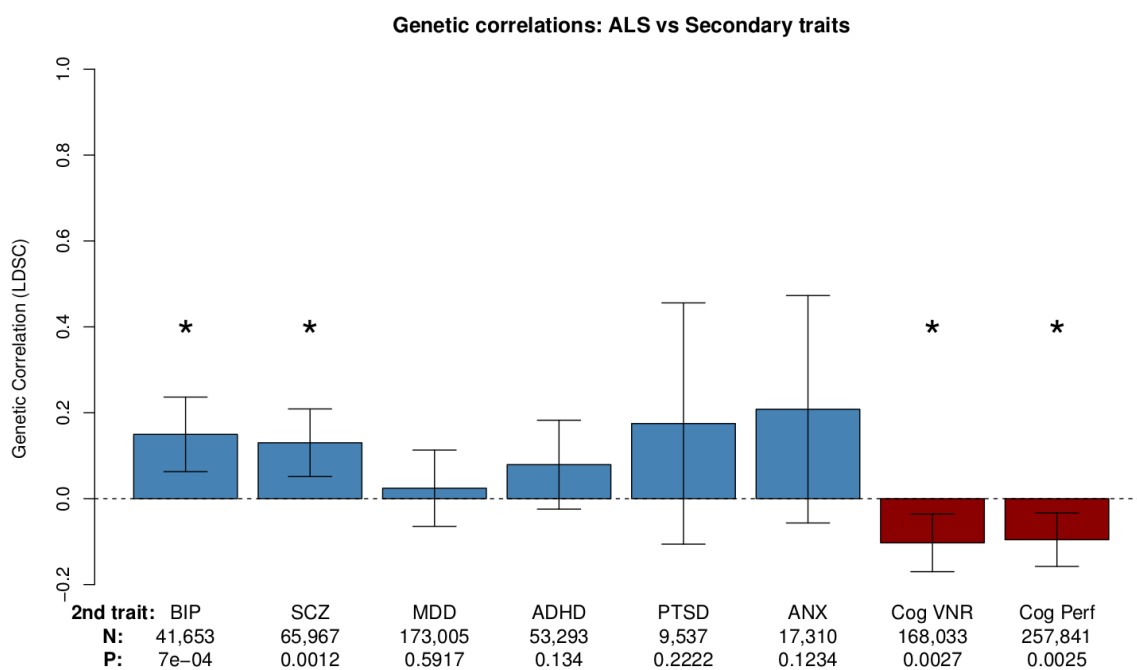
## 2.3 - Results

### 2.3.1 - Characterising the genetic overlap between ALS and an extended set of neuropsychiatric and cognitive traits

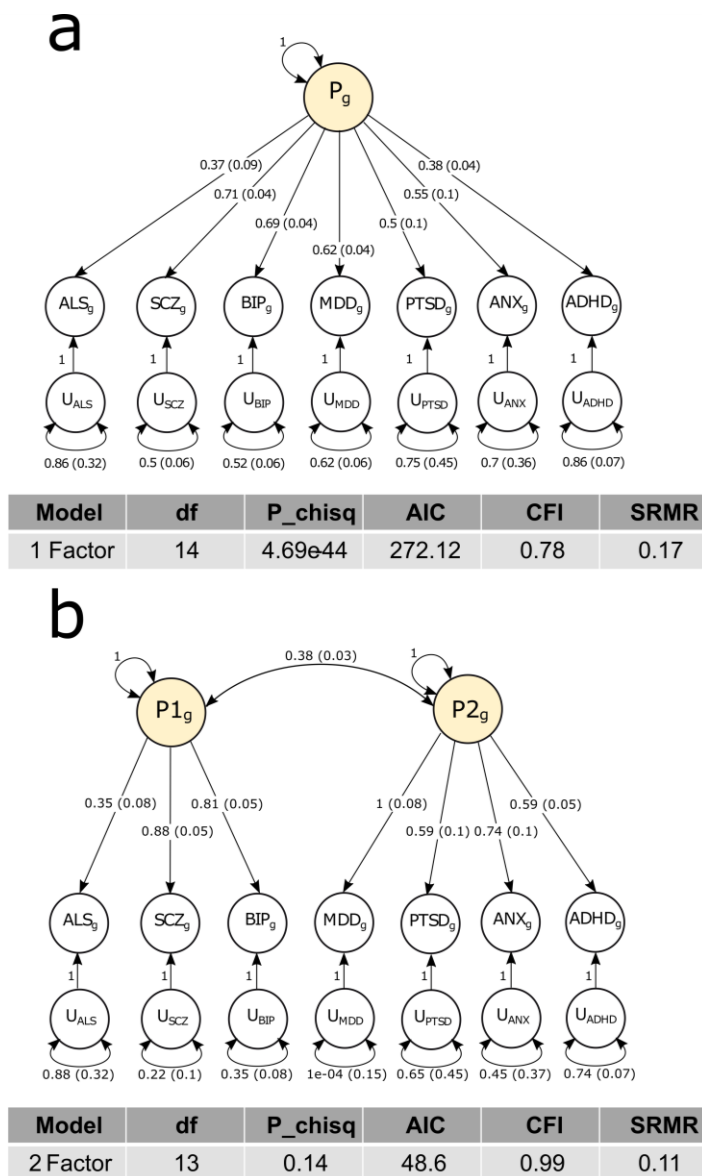
We revisited the genetic correlation analysis carried out by McLaughlin et al. (R. L. McLaughlin et al. 2017) to update the picture of genetic correlation between ALS and neuropsychiatric traits, by incorporating recent larger GWAS for secondary traits. We additionally included summary statistics from two robust studies of cognition to replicate and reinforce previous observations of negative genetic correlation with cognition. To this end we ran bivariate LD score regression on ALS and schizophrenia (Ruderfer et al. 2018), bipolar disorder (Ruderfer et al. 2018), major depressive disorder (Wray et al. 2018), ADHD (Demontis et al. 2019), anxiety (Otowa et al. 2016), PTSD (Duncan et al. 2018) and two measures of cognitive performance (Davies et al. 2018; J. J. Lee, Wedow, et al. 2018). This analysis replicated the previously reported genetic correlation between ALS and schizophrenia ( $r_g=0.13$ ;  $p=0.0012$ ) with an independent schizophrenia GWAS dataset and additionally identified a novel positive correlation with bipolar disorder ( $r_g=0.15$ ;  $p=0.0007$ ) (Figure 2.3). Additionally we replicated previously borderline negative genetic correlation with cognition using two measures of cognitive performance (verbal numeric reasoning:  $r_g=-0.1$ ,  $p=0.0027$ ; cognitive performance:  $r_g=-0.1$ ;  $p=0.0025$ ) (Figure 2.3). We noted that while the remaining psychiatric traits did not show significant correlation, they showed a consistent positive trend, motivating us to test the hypothesis of whether ALS and psychiatric traits might be driven by a shared latent factor.

We ran GenomicSEM (Grotzinger et al. 2019) on the genetic covariance matrix between ALS and neuropsychiatric traits to test whether the traits were mediated by the same or distinct latent factors. We tested the appropriate number of latent factors using exploratory factor analysis on the genetic covariance matrix in R using “factanal” function, followed by fitting these models in confirmatory factor analysis in genomic SEM. Both a single common factor model (as used in recent work exploring genetic evidence for a shared psychiatric “p-factor”) and a two correlated genetic factors model fit the data well based on the Comparative factor index (CFI) and Akaike’s information criteria (AIC) (Figure 2.4). The common factor model fit the data well ( AIC= 272.12; CFI=0.78; SRMR=0.17) suggesting that ALS may fit within the “p-factor” model for psychiatric traits, or that a shared genetic module influences both psychiatric traits and ALS, however the correlated two factor model had a slightly improved fit (AIC=48.6; CFI=0.99; SRMR=0.11) meaning this is likely the best model. The implication of this finding is that emerging larger GWAS

of psychiatric traits and ALS are likely to yield more genetic correlations between ALS and psychiatric traits as a single or two correlated latent factors best describe their covariance structure. Additionally these models hints that it might be useful to test for pleiotropic variants across these traits using a multi-trait GWAS method.



**Figure 2.3: Genetic correlations of ALS with psychiatric traits and cognition.** Genetic correlations between ALS and secondary psychiatric (blue) and cognitive (red) traits as estimated using LDSC. Error bars represent the 95% confidence interval for the genetic correlation estimate. Bonferroni corrected significant traits are highlighted with an asterisk. BIP, Bipolar disorder; SCZ, Schizophrenia; MDD, Major depressive disorder; ADHD, Attention deficit hyperactivity disorder; PTSD, Post traumatic stress disorder; ANX, Anxiety disorder; Cog VNR, Verbal numeric reasoning; Cog Perf, General cognitive performance.



**Figure 2.4: Common and two genetic factor model for ALS and psychiatric traits.** Path diagrams and model fit statistics describing the best two models from the genomicSEM analysis of ALS and psychiatric traits. a.) Displays a model where a single shared latent genetic factor describes the observed genetic relationships between traits, and b.) displays a model where two correlated latent genetic factors describe the observed genetic relationships between traits. The modelled loadings of the latent factors (i.e. how much variance the factor explains for each trait) are labelled on single headed arrows from the factors ( $P_g$ ,  $P1_g$ , and  $P2_g$ ) to the traits ( $ALS_g$ ,  $PTSD_g$  etc.). All trait loadings are positive indicating that traits are well described by these latent factors (estimate standard errors in brackets). Double headed arrows represent the residual variances of the indicators (traits) and the correlation between modelled factors. Both models are well behaved as indicated by their model fit statistics, however the two correlated factor model shows better overall fit. ALS, Amyotrophic lateral sclerosis; BIP, Bipolar disorder; SCZ, Schizophrenia; MDD, Major depressive disorder; PTSD, Post traumatic stress disorder; ANX, Anxiety disorder; ADHD, Attention deficit hyperactivity disorder.

To identify pleiotropic loci that may partially explain the genetic overlap between ALS and secondary traits we ran multi-trait GWAS using MTAG (Turley et al. 2018), which leverages correlated summary statistics from multiple traits to enhance power to detect causal SNPs. MTAG identified a number of novel loci that were jointly associated with ALS and: i.) bipolar disorder and schizophrenia; ii.) low cognitive performance in this multi-trait analysis at genome wide significance which may represent signals of shared risk for these traits (Table 2.1). While some known ALS loci were implicated as pleiotropic (e.g. *C9orf72*), many loci only reach genome-wide significance in the GWAS for the secondary trait, and are only nominally associated for ALS, hence may be false positives. MTAG is known to produce false positives due to its assumption that SNPs share the same variance-covariance matrix of effect sizes across traits, meaning the MTAG effect will be biased away from zero for SNPs with a null effect in one trait if they have non-null effects in the others. This is expected to happen often for lower-powered GWAS.

**Table 2.1: MTAG hits for ALS and secondary traits.**

				ALS source GWAS P-value		Secondary trait GWAS P-value			MTAG ALS GWAS P-value	
CHR	SNP	Gene	2nd trait	(2016)	(2018)	Cog	BIP	SCZ	(2016)	(2018)
1	rs7542974	<i>NEGR1</i>	Cog	1.28E-01	1.93E-02	6.00E-20	-	-	2.10E-08	1.02E-10
3	rs9848497	<i>MON1A</i>	Cog	2.98E-01	1.27E-01	4.77E-25	-	-	4.49E-09	7.36E-11
6	rs9384679	<i>FOXO3</i>	Cog	9.55E-02	5.25E-01	1.14E-21	-	-	3.36E-09	4.35E-08
7	rs12707087	<i>EXOC4*</i>	Cog	8.70E-05	1.41E-04	2.22E-08	-	-	3.76E-08	1.71E-08
7	rs6950324	<i>EXOC4*</i>	Cog	1.92E-03	4.92E-03	2.64E-11	-	-	4.62E-08	3.87E-08
7	rs6956399	<i>EXOC4*</i>	Cog	3.30E-04	5.35E-04	1.22E-10	-	-	1.27E-08	5.30E-09
7	rs12532950	<i>EXOC4*</i>	Cog	2.30E-03	2.04E-02	4.86E-14	-	-	4.60E-09	2.10E-08
8	rs4976976	<i>TSNARE 1</i>	Cog	1.34E-03	5.64E-04	1.24E-11	-	-	2.24E-08	1.91E-09
9	rs3849943	<b><i>C9orf72, IFNK, MOB3B*</i></b>	Cog	1.71E-24	3.77E-30	5.55E-01	-	-	1.72E-13	4.51E-17
9	rs117204439	<b><i>C9orf72*</i></b>	Cog	9.05E-16	3.08E-14	4.46E-01	-	-	2.72E-09	7.25E-09
15	rs12439619	<i>EFTUD1*</i>	Cog	3.52E-04	3.01E-05	7.94E-11	-	-	1.13E-08	1.83E-10
9	rs117204439	<b><i>C9orf72*</i></b>	BIP+SCZ	9.05E-16	3.08E-14	-	9.88E-02	8.58E-01	1.25E-08	1.71E-12
10	rs12218148	<i>AS3MT, CNNM2</i>	BIP+SCZ	2.32E-01	3.98E-03	-	4.32E-04	4.38E-13	9.43E-09	2.71E-08
10	rs34747231	<i>CNNM2</i>	BIP+SCZ	2.75E-01	6.53E-03	-	2.45E-03	1.14E-14	5.31E-09	2.54E-08

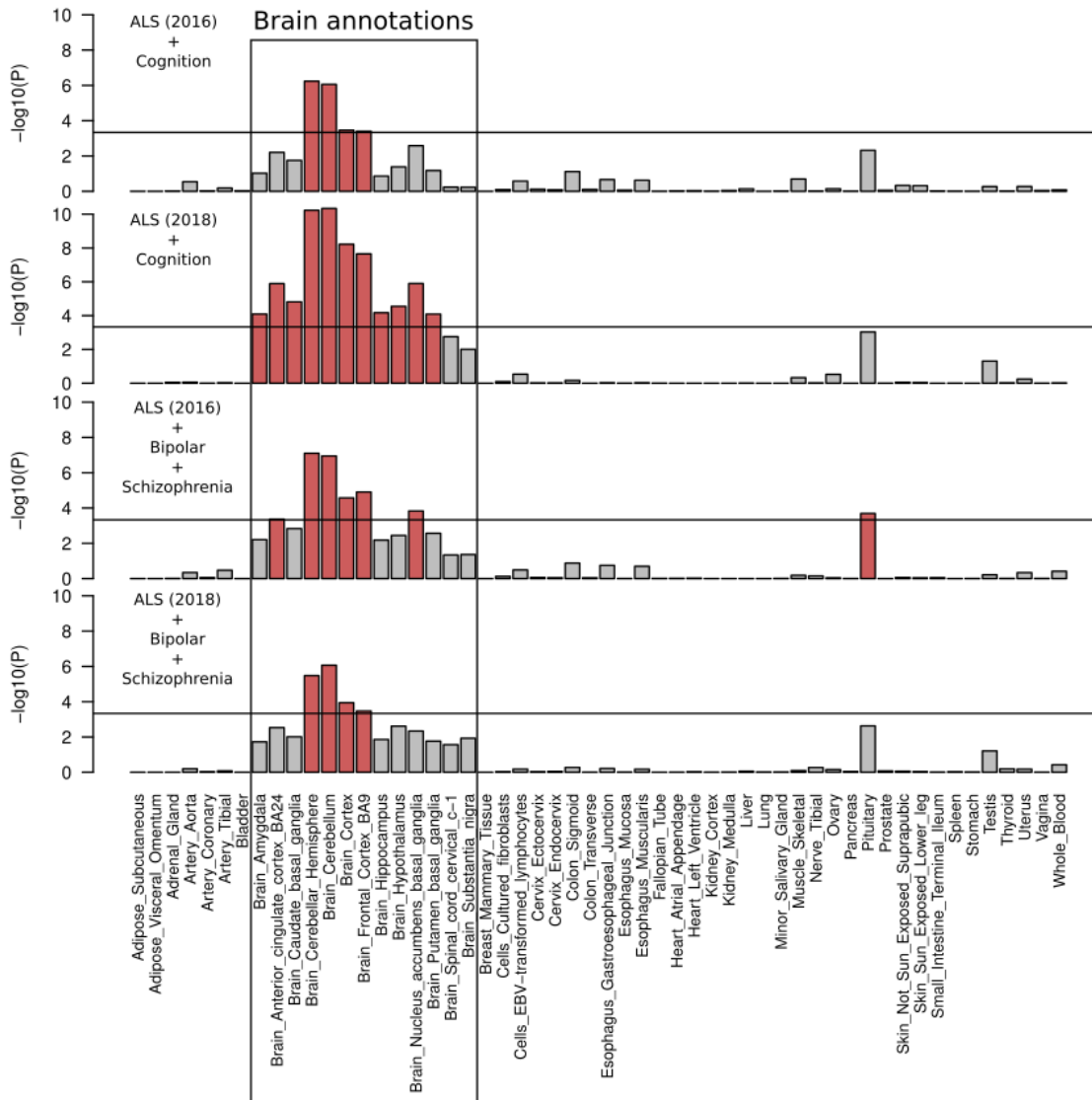
Lists independent SNPs (clumped) passing genome-wide significance threshold in MTAG in both ALS GWAS when run with secondary traits (Cog, cognition; BIP, Bipolar; SCZ, Schizophrenia). P-values are listed for the source GWAS for the primary and secondary traits and for the MTAG joint analysis. The closest gene is also listed for reference, however this has not been colocalised. Genes highlighted in bold are significant in the source single trait analysis for ALS. We expect a number of hits to be false positives as the maximum FDR analysis of MTAG suggests the false discovery rate could be high with such underpowered GWAS. \*Genes also identified in cFDR multi-trait analysis.



To estimate the upper bound for the rate of false positives we ran the “maximum FDR” (maxFDR) analysis from the MTAG paper, which maximises the estimate of false discovery rate using a grid search over a range of mixing weights. Maximum FDR was high for this analysis for both ALS GWAS (van Rheenen et al. 2016; Nicolas et al. 2018) suggesting many of these hits could potentially be false positives, meaning we should interpret them with caution (ALS 2016:  $\text{maxFDR}_{\text{bipolar/schizophrenia}} = 0.65$ ,  $\text{maxFDR}_{\text{cognition}} = 0.4$ ; ALS 2018:  $\text{maxFDR}_{\text{bipolar/schizophrenia}}=0.33$ ,  $\text{maxFDR}_{\text{cognition}}=0.23$ ). This indicates that ALS GWAS are likely still too small and underpowered to confidently perform analysis using the MTAG model with these traits, particularly for bipolar disorder and schizophrenia. To reduce the false discovery rate, we only accepted variants achieving genome-wide significance when analysed with secondary traits in both the 2016 and 2018 ALS GWAS (van Rheenen et al. 2016; Nicolas et al. 2018) (Table 2.1).

Functional analysis of tissue enrichment using the MAGMA gene set analysis tool (de Leeuw et al. 2015) for our MTAG runs showed significant enrichment of signal in genes expressed in several brain regions (Figure 2.5). In particular genes expressed in the frontal cortex and cerebellum showed enrichment across all ALS datasets with all secondary traits. This suggests that pleiotropic loci shared between ALS and cognitive and psychiatric phenotypes are typically expressed in the brain. Moreover while there was no significant enrichment in functional annotations for the multi-trait analysis of ALS and psychiatric traits, the analysis of ALS and cognition showed enrichment in genes associated with several biological processes (Table 2.2). In particular, genes with shared signal for ALS and cognition showed enrichment in processes related to neurone development and differentiation. Additionally, ALS and cognition showed enrichment in an annotation for the *MECP2* reactome, which regulates many transcription factors involved in the functioning of the nervous system (Chahrour et al. 2008). These results may elucidate the genetic root of cognitive symptoms in ALS, as well as the known familial overlap between ALS and neuropsychiatric traits if validated, motivating follow-up analysis in larger datasets. To further address the biases potentially introduced by MTAG, we also ran an analysis using another pleiotropy detection method (conditional false discovery rate (Andreassen et al. 2013)).

## MAGMA: Tissue enrichment (MTAG)



**Figure 2.5: ALS multi-trait analyses are enriched for genes expressed in brain.** MAGMA tissue enrichment analysis of MTAG summary statistics for ALS (2016 and 2018) and correlated secondary traits (cognition, bipolar disorder and schizophrenia) shows an enrichment of signal in genes highly expressed in brain tissues, and the pituitary gland. Notably both frontal cortex and cerebellum annotations were significantly enriched across all MTAG runs.

**Table 2.2: Functional annotations enriched in ALS and cognition MTAG analysis.**

Annotation	ALS 2016 + Cognition			ALS 2018 + Cognition		
	GENES	BETA	p-val	GENES	BETA	p-val
Curated gene sets: Reactome mecp2 regulates transcription factors	4	2.7428	5.3E-07*	3	3.7842	2.6E-10*
GO_bp: Neuron development	1031	0.093111	3.8E-04	1007	0.13527	1.7E-06*
GO_bp: Cell part morphogenesis	636	0.11281	7.4E-04	615	0.17298	1.7E-06*
GO_bp: Neurogenesis	1507	0.076112	5.2E-04	1465	0.12003	4.6E-07*
GO_bp: Neuron differentiation	1268	0.089337	2.0E-04	1234	0.1266	8.8E-07*

Summarises gene ontologies and curated gene sets with enriched signal from the FUMA analysis of MTAG summary statistics for multi-trait analysis run on ALS and cognition. P-values below the Bonferroni-adjusted significance threshold are denoted with an asterisk.

Given the likelihood of false positives in the MTAG analysis we decided to apply the conditional false discovery rate (cFDR) method (Andreassen et al. 2013) to further explore pleiotropic loci, with the aim of validating our MTAG results. cFDR conditions the association of a given SNP in a principal trait on the association in a second conditional trait, testing for the probability that association at a variant is null for the principal trait given p values of both, hence leveraging pleiotropy to gain power. Our cFDR results (Table 2.3) identified several putative loci shared between ALS and secondary traits. Notably these loci tend to have lower p-values in the base ALS GWAS than those identified by MTAG, indicating this method does not bias estimates away from zero as severely as the MTAG analysis. The cFDR analysis of ALS conditioned on cognition replicated signals from the MTAG analysis at the *EXOC4*, *EFTUD1* and *C9orf72* locus, suggesting these are likely real pleiotropic loci or true positives. This method also found hits in other known ALS loci (*SARM1*, *UNC13A*, *SCFD1*, *TBK1*, *MOBP*) and novel genes (*NCKAP5L*, *KRT18P55*, *GGNBP2*, *CRHR1*, *MAPT*, *KANSL1*, *NSF*, *SRGAP1*, *SPIRE1*), which despite their absence in the MTAG analysis, may also be pleiotropic loci shared between ALS and the secondary traits. Some of these novel loci are promising candidates for involvement in ALS, for example *MAPT* (microtubule-associated protein tau), which shows pleiotropic overlap with cognition in this analysis is well established as a monogenic causal gene for FTD (Takada 2015) was recently associated through cFDR with ALS when analysed with FTD as a secondary trait (Karch et al. 2018). Additionally,

*NSF*, *KANSL1* and *GGNBP2* were all identified as associated with ALS when conditioned on other diseases of the frontotemporal temporal dementia spectrum (Karch et al. 2018). Hence these variants appear to be robustly pleiotropically linked to ALS and cognitive and behavioural phenotypes.

**Table 2.3: cFDR hits for ALS and secondary traits.**

SNP	CHR	bp	Trait 2	Closest gene	p ALS	p trait 2	cFDR
rs6765697	3	39493239	Cog	-	1.11E-07	4.97E-02	9.89E-03
rs1768208	3	39523003	Cog	<b>MOBP</b>	1.77E-08	3.97E-01	1.12E-03
rs6956399	7	133305729	Cog	<b>EXOC4*</b>	3.30E-04	1.22E-10	8.43E-03
rs7813314	8	2415366	Cog	-	3.14E-08	2.51E-02	4.88E-03
rs147211831	9	27436084	Cog	<b>MOB3B*</b>	3.38E-14	2.00E-01	6.96E-08
rs10511816	9	27468461	Cog	<b>MOB3B*</b>	1.24E-11	1.66E-01	2.26E-05
rs4879515	9	27482235	Cog	<b>MOB3B*</b>	5.13E-10	3.75E-01	4.70E-05
rs139185008	9	27491942	Cog	<b>MOB3B*</b>	1.38E-13	8.85E-01	1.14E-08
rs4879524	9	27495362	Cog	<b>IFNK,MOB3B*</b>	4.84E-10	2.81E-01	6.54E-05
rs700786	9	27522925	Cog	<b>IFNK,MOB3B*</b>	2.26E-10	2.11E-02	1.29E-04
rs4879554	9	27546275	Cog	<b>C9orf728</b>	4.98E-09	2.11E-01	8.15E-04
rs774357	9	27559835	Cog	<b>C9orf72,IFNK,MOB3B*</b>	4.76E-24	6.99E-01	3.85E-18
rs2492816	9	27565105	Cog	<b>C9orf72,IFNK,MOB3B*</b>	2.03E-11	3.91E-01	2.24E-06
rs2120721	9	27566141	Cog	<b>C9orf72*</b>	3.81E-10	7.21E-02	2.10E-04
rs11795154	9	27577611	Cog	<b>C9orf72*</b>	1.13E-09	4.04E-01	8.42E-05
rs7864502	9	27583128	Cog	<b>C9orf72*</b>	2.49E-12	2.70E-01	7.78E-07
rs117204439	9	27607973	Cog	-	9.05E-16	4.46E-01	9.33E-11
rs79676202	12	50180558	Cog	<b>NCKAP5L</b>	9.18E-08	3.16E-01	5.66E-03
rs73124200	12	64949248	Cog	-	7.78E-08	5.83E-01	3.55E-03
rs74654358	12	64881967	Cog	<b>TBK1</b>	6.65E-08	6.92E-01	3.14E-03

rs10139154	14	31147498	Cog	<b>SCFD1</b>	4.95E-08	3.56E-01	3.10E-03
rs12439619	15	82546946	Cog	<i>EFTUD1*</i>	3.52E-04	7.94E-11	8.79E-03
rs34517613	17	26610252	Cog	<i>KRT18P55</i>	8.62E-08	8.66E-01	3.56E-03
rs35714695	17	26719788	Cog	<b>SARM1</b>	8.96E-11	9.52E-01	6.70E-06
rs3736166	17	34900836	Cog	<i>GGNBP2</i>	1.24E-06	8.01E-10	1.71E-03
rs62057158	17	43907143	Cog	<i>CRHR1,MAPT</i>	2.19E-04	3.08E-11	6.07E-03
rs112073200	17	44201791	Cog	<i>KANSL1</i>	2.29E-04	2.96E-11	6.11E-03
rs7224296	17	44800046	Cog	<i>NSF</i>	3.26E-04	6.06E-10	9.92E-03
rs12608932	19	17752689	Cog	<b>UNC13A</b>	2.69E-10	1.53E-01	1.17E-04
rs4676496	3	39498005	BIP+SCZ	<b>MOBP</b>	8.86E-08	8.18E-01	3.57E-03
rs13067055	3	39510517	BIP+SCZ	<b>MOBP</b>	2.15E-07	4.53E-01	5.28E-03
rs616147	3	39534481	BIP+SCZ	<b>MOBP</b>	1.43E-08	4.87E-01	4.79E-04
rs7813314	8	2415366	BIP+SCZ	<b>LOC101927815</b>	3.14E-08	9.36E-03	3.41E-04
rs10511816	9	27468461	BIP+SCZ	<b>MOB3B*</b>	1.24E-11	2.14E-01	3.33E-07
rs4879515	9	27482235	BIP+SCZ	<b>MOB3B*</b>	5.13E-10	6.74E-02	5.72E-06
rs139185008	9	27491942	BIP+SCZ	<b>MOB3B*</b>	1.38E-13	8.51E-03	7.38E-09
rs4879524	9	27495362	BIP+SCZ	<b>IFNK,MOB3B*</b>	4.84E-10	2.97E-03	4.59E-05
rs700786	9	27522925	BIP+SCZ	<b>IFNK,MOB3B*</b>	2.26E-10	3.48E-01	7.56E-06
rs4879541	9	27533452	BIP+SCZ	<b>C9orf72,IFNK,MOB3B*</b>	8.48E-10	8.80E-03	1.46E-05
rs4879554	9	27546275	BIP+SCZ	<b>C9orf72*</b>	4.98E-09	8.18E-01	2.58E-04
rs17696570	9	27558437	BIP+SCZ	<b>C9orf72*</b>	4.67E-07	9.24E-02	4.50E-03
rs2484319	9	27563755	BIP+SCZ	<b>C9orf72,IFNK,MOB3B8</b>	3.15E-24	1.71E-02	2.94E-19
rs2120721	9	27566141	BIP+SCZ	<b>C9orf72*</b>	3.81E-10	2.48E-01	9.83E-06
rs7864502	9	27583128	BIP+SCZ	<b>C9orf72*</b>	2.49E-12	5.27E-03	3.12E-07
rs2453556	9	27586162	BIP+SCZ	<b>C9orf72*</b>	3.18E-11	5.15E-02	3.57E-07

rs117204439	9	27607973	BIP+SCZ	-	9.05E-16	2.39E-01	3.36E-11
rs79676202	12	50180558	BIP+SCZ	<i>NCKAP5L</i>	9.18E-08	3.82E-01	2.22E-03
rs116900480	12	58656105	BIP+SCZ	-	2.00E-07	7.64E-02	1.73E-03
rs76805704	12	64532377	BIP+SCZ	<i>SRGAP1</i>	5.65E-07	2.28E-01	8.92E-03
rs74654358	12	64881967	BIP+SCZ	<b><i>TBK1</i></b>	6.65E-08	4.00E-02	4.72E-04
rs73124200	12	64949248	BIP+SCZ	-	7.78E-08	3.86E-01	1.90E-03
rs10139154	14	31147498	BIP+SCZ	<b><i>SCFD1</i></b>	4.95E-08	4.46E-01	1.40E-03
rs34517613	17	26610252	BIP+SCZ	<i>KRT18P55</i>	8.62E-08	7.68E-01	3.33E-03
rs35714695	17	26719788	BIP+SCZ	<b><i>SARM1</i></b>	8.96E-11	3.32E-01	3.06E-06
rs12967284	18	12532098	BIP+SCZ	<i>SPIRE1</i>	7.33E-07	1.52E-01	8.79E-03
rs12608932	19	17752689	BIP+SCZ	<b><i>UNC13A</i></b>	2.69E-10	1.15E-01	4.33E-06
rs117635456	21	43460912	BIP+SCZ	-	2.52E-07	6.97E-01	7.02E-03

Loci passing the cFDR threshold (cFDR<0.01) for ALS conditioned on cognitive performance (Cog) and a combined bipolar/schizophrenia (BIP+SCZ) GWAS. Genes in bold have been previously identified in the source ALS GWAS using standard single trait analysis.

\*Gene also identified in MTAG analysis.

Genes closest to hits identified using cFDR on ALS conditional on cognition showed enrichment in sets of genes highly expressed in several brain regions (Table 2.4), as seen in the MTAG analysis of ALS and cognition, lending some support for their validity. These genes were also enriched in a large number of gene ontology annotations (Table 2.5), many of which have feasible links to ALS biology. Notably these ontologies indicate that several of these pleiotropic genes are functional in cell components of neurones including dendrites, axons and synapses, and fulfil biological processes relating to neurone death and synaptic connectivity, which may highlight shared biological features between ALS and cognition. In contrast loci identified as pleiotropic with psychiatric traits via cFDR do not show significant enrichment in tissue expression sets or gene ontologies.

To further explore whether this overlap with psychiatric traits and cognition is partially driven by a shared tissue of expression of risk genes we performed partitioned LD score regression on ALS, bipolar disorder, schizophrenia and cognitive performance using annotations for cell type-specific gene expression. This analysis estimates whether

heritability is enriched in each annotation (i.e. does the per SNP heritability of that annotation exceed the per SNP heritability genome wide). These traits all showed a significant enrichment of heritability in genes which are highly expressed in the central nervous system (CNS) (Figure 2.6), as expected from our MTAG and cFDR analysis, suggesting the genetic correlation may be due to a tissue overlap. Regression of coefficients of enrichment across tissues between ALS and these traits showed significant linear relationships (Figure 2.7), supporting the hypothesis of shared tissues of effect across these correlated traits. While ALS showed a trend towards enrichment in the CNS, only the annotation for the frontal lobe passed the stringent Bonferroni-corrected multiple testing threshold. This could indicate that the CNS involvement in ALS is specific to the frontal lobe, or alternatively that the method used is insufficiently powered to identify other CNS enrichments. We note that cell types within a tissue are correlated for gene expression, meaning treating each test as independent and correcting in this manner is perhaps overly conservative and may lower power unnecessarily. In fact the source paper for this method instead employs a less stringent false discovery rate approach for identifying significant enrichments (Finucane et al. 2018), which may have improved resolution. To get a more nuanced picture of which regions of the brain may be involved in ALS pathology we also ran partitioned LD score regression on an annotation for genes expressed in different cell types in the brain (GTEx Brain dataset). This returned a significant enrichment for the frontal cortex ( $p=9\times 10^{-4}$ ), consistent with the multi-tissue result of frontal lobe involvement, and the MAGMA tissue enrichment analysis (Figure 2.5). Overall our results suggest that ALS and correlated traits are largely driven by genes expressed in the CNS, which may partially explain the genetic overlap between ALS and psychiatric and cognitive traits.

**Table 2.4: Tissue expression annotations enriched for cFDR(ALS|Cognition) hits.**

Gene Set	N genes in set	N genes overlap	p	FDR adj p	Genes
Brain Amygdala	9576	13	1.41E-04	7.61E-03	<i>MOBP, EXOC4, MOB3B, C9orf72, NCKAP5L, TBK1, SCFD1, EFTUD1, CRHR1, MAPT, KANSL1, NSF, UNC13A</i>
Brain Anterior cingulate cortex BA24	9101	13	7.87E-05	4.25E-03	<i>MOBP, EXOC4, MOB3B, C9orf72, NCKAP5L, TBK1, SCFD1, EFTUD1, CRHR1, MAPT, KANSL1, NSF, UNC13A</i>
Brain Caudate basal ganglia	9143	12	5.41E-04	2.92E-02	<i>MOBP, EXOC4, MOB3B, C9orf72, NCKAP5L, TBK1, SCFD1, EFTUD1, MAPT, KANSL1, NSF, UNC13A</i>
Brain Hippocampus	9541	12	8.32E-04	4.49E-02	<i>MOBP, EXOC4, C9orf72, NCKAP5L, TBK1, SCFD1, EFTUD1, CRHR1, MAPT, KANSL1, NSF, UNC13A</i>
Brain Putamen basal ganglia	9598	12	8.84E-04	4.77E-02	<i>MOBP, EXOC4, MOB3B, C9orf72, NCKAP5L, TBK1, SCFD1, EFTUD1, MAPT, KANSL1, NSF, UNC13A</i>

Tissues showing significant differential expression of genes identified by cFDR analysis of ALS and cognition (either up or down regulated). Notably several tissues here match those identified in MTAG analysis of ALS and cognition (Figure 2.5).

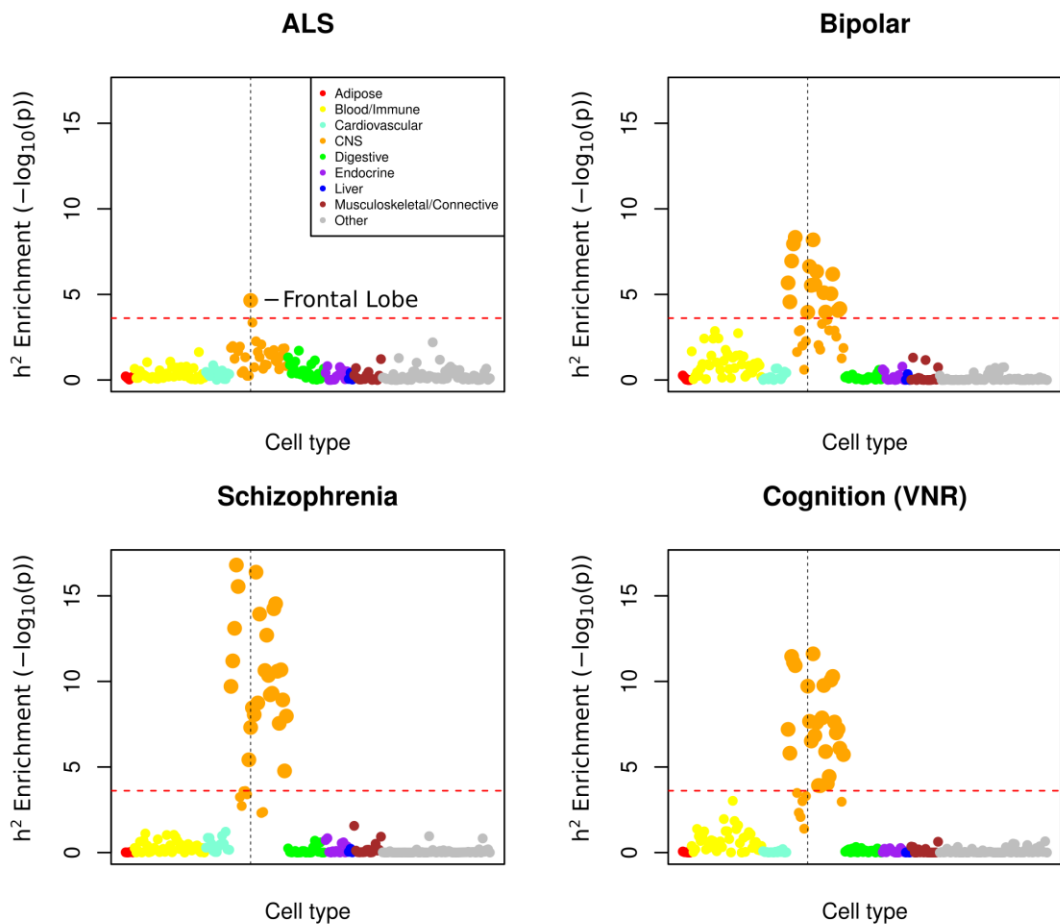


**Table 2.5: Gene ontology terms enriched for cFDR (ALS|Cognition) hits.**

Category	GeneSet	N genes set	N genes overlap	p	FDR adj p	Genes
GO bp	GO: VESICLE DOCKING	65	4	1.91E-08	1.40E-04	<i>SCFD1, NSF, UNC13A, EXOC4</i>
GO bp	GO: ESTABLISHMENT OF ORGANELLE LOCALIZATION	491	6	5.13E-08	1.89E-04	<i>SCFD1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>
GO bp	GO: ORGANELLE LOCALIZATION	685	6	3.64E-07	8.92E-04	<i>SCFD1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>
GO bp	GO: VESICLE DOCKING INVOLVED IN EXOCYTOSIS	44	3	1.01E-06	1.59E-03	<i>SCFD1, UNC13A, EXOC4</i>
GO bp	GO: MEMBRANE DOCKING	177	4	1.08E-06	1.59E-03	<i>SCFD1, NSF, UNC13A, EXOC4</i>
GO bp	GO: EXOCYTIC PROCESS	81	3	6.47E-06	7.92E-03	<i>SCFD1, UNC13A, EXOC4</i>
GO bp	GO: VESICLE LOCALIZATION	303	4	9.09E-06	8.72E-03	<i>SCFD1, NSF, UNC13A, EXOC4</i>
GO bp	GO: VESICLE TARGETING	92	3	9.49E-06	8.72E-03	<i>SCFD1, NSF, EXOC4</i>
GO bp	GO: REGULATION OF AUTOPHAGY	319	4	1.11E-05	9.09E-03	<i>TBK1, SCFD1, MAPT, EXOC4</i>
GO bp	GO: POST GOLGI VESICLE MEDIATED TRANSPORT	104	3	1.37E-05	1.01E-02	<i>SCFD1, NSF, EXOC4</i>
GO bp	GO: NEURON DEATH	345	4	1.51E-05	1.01E-02	<i>TBK1, SARM1, CRHR1, MAPT</i>
GO bp	GO: REGULATION OF PEPTIDYL SERINE PHOSPHORYLATION	139	3	3.27E-05	2.00E-02	<i>TBK1, GGNBP2, IFNK</i>
GO bp	GO: EXOCYTOSIS	897	5	3.70E-05	2.09E-02	<i>SCFD1, CRHR1, NSF, UNC13A, EXOC4</i>
GO bp	GO: REGULATION OF PEPTIDYL SERINE PHOSPHORYLATION OF STAT PROTEIN	23	2	4.89E-05	2.48E-02	<i>GGNBP2, IFNK</i>
GO bp	GO: REGULATION OF CATABOLIC PROCESS	959	5	5.09E-05	2.48E-02	<i>TBK1, SCFD1, MAPT, NSF, EXOC4</i>
GO bp	GO: PROCESS UTILIZING AUTOPHAGIC MECHANISM	485	4	5.72E-05	2.48E-02	<i>TBK1, SCFD1, MAPT, EXOC4</i>
GO bp	GO: REGULATION OF MACROAUTOPHAGY	168	3	5.74E-05	2.48E-02	<i>TBK1, SCFD1, EXOC4</i>
GO bp	GO: SERINE PHOSPHORYLATION OF STAT PROTEIN	27	2	6.78E-05	2.77E-02	<i>GGNBP2, IFNK</i>
GO bp	GO: REGULATION OF SYNAPTIC PLASTICITY	182	3	7.28E-05	2.82E-02	<i>CRHR1, MAPT, UNC13A</i>
GO bp	GO: MYD88 INDEPENDENT TOLL LIKE RECEPTOR SIGNALING PATHWAY	32	2	9.56E-05	3.37E-02	<i>TBK1, SARM1</i>
GO bp	GO: INTRACELLULAR TRANSPORT	1815	6	9.62E-05	3.37E-02	<i>SCFD1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>

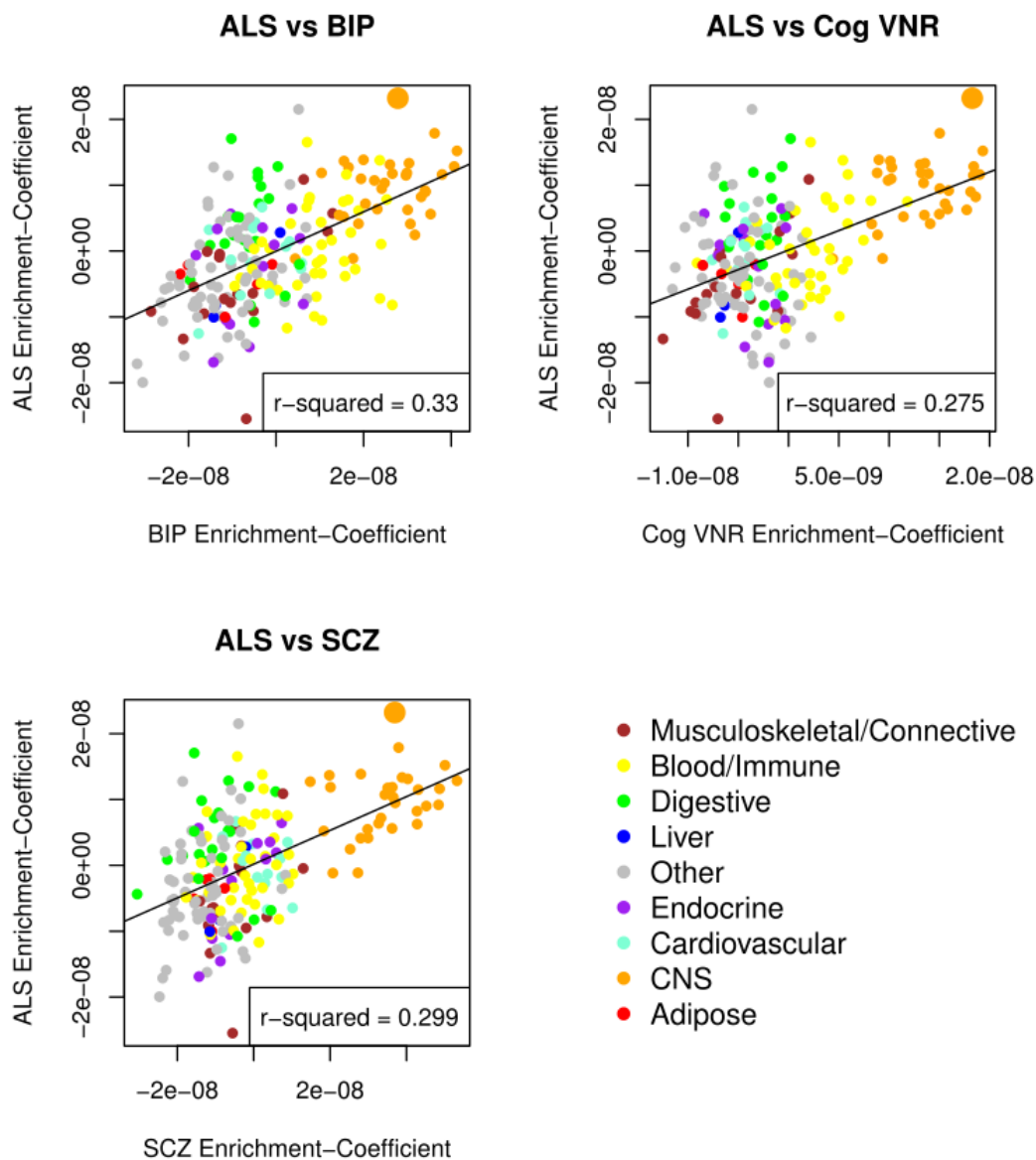
GO cc	GO: NEURON PROJECTION	1301	6	1.48E-05	1.31E-02	<i>SARM1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>
GO cc	GO: CELL PROJECTION PART	1438	6	2.61E-05	1.31E-02	<i>SARM1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>
GO cc	GO: CYTOPLASMIC REGION	488	4	5.86E-05	1.38E-02	<i>MAPT, UNC13A, MOBP, EXOC4</i>
GO cc	GO: CELL CORTEX PART	176	3	6.59E-05	1.38E-02	<i>UNC13A, MOBP, EXOC4</i>
GO cc	GO: NEURON PART	1709	6	6.89E-05	1.38E-02	<i>SARM1, CRHR1, MAPT, NSF, UNC13A, EXOC4</i>
GO cc	GO: DENDRITIC TREE	588	4	1.20E-04	1.63E-02	<i>SARM1, CRHR1, MAPT, NSF</i>
GO cc	GO: SYNAPSE	1169	5	1.30E-04	1.63E-02	<i>SARM1, MAPT, NSF, UNC13A, EXOC4</i>
GO cc	GO: AXON	600	4	1.30E-04	1.63E-02	<i>SARM1, MAPT, UNC13A, EXOC4</i>
GO cc	GO: DISTAL: AXON	280	3	2.60E-04	2.89E-02	<i>MAPT, UNC13A, EXOC4</i>
GO cc	GO: CELL CORTEX	302	3	3.24E-04	3.24E-02	<i>UNC13A, MOBP, EXOC4</i>
GO cc	GO: SOMATODENDRITIC COMPARTMENT	818	4	4.24E-04	3.86E-02	<i>SARM1, CRHR1, MAPT, NSF</i>
GO cc	GO: ORGANELLE SUBCOMPARTMENT	375	3	6.09E-04	4.24E-02	<i>SCFD1, CRHR1, NSF</i>
GO cc	GO: AXON PART	376	3	6.14E-04	4.24E-02	<i>MAPT, UNC13A, EXOC4</i>
GO cc	GO: INCLUSION BODY	81	2	6.17E-04	4.24E-02	<i>TBK1, MAPT</i>
GO cc	GO: WHOLE MEMBRANE	1647	5	6.36E-04	4.24E-02	<i>TBK1, SARM1, MAPT, NSF, UNC13A</i>
GO cc	GO: MICROTUBULE	410	3	7.89E-04	4.93E-02	<i>NCKAP5L, SARM1, MAPT</i>
GO mf	GO: SYNTAXIN BINDING	70	3	4.16E-06	4.57E-03	<i>SCFD1, NSF, UNC13A</i>
GO mf	GO: PROTEIN CONTAINING COMPLEX BINDING	1096	6	5.56E-06	4.57E-03	<i>SCFD1, EFTUD1, CRHR1, MAPT, NSF, IFNK</i>
GO mf	GO: SNARE BINDING	105	3	1.41E-05	7.74E-03	<i>SCFD1, NSF, UNC13A</i>
GO mf	GO: SYNTAXIN 1 BINDING	22	2	4.47E-05	1.84E-02	<i>NSF, UNC13A</i>

Gene ontologies showing significant enrichment (5% FDR) for genes identified in the cFDR analysis of ALS and cognition. Abbreviations: bp, Biological process; mf, Molecular function; cc, Cell component



**Figure 2.6: Cell type specific heritability enrichments.**

Panels display the probability that per SNP heritability is enriched (i.e. greater than expected genomewide; y-axis) for ALS and secondary traits in sets of genes (defined in Finucane et al. (Finucane et al. 2018)) highly-expressed in a range of cell types (points). The horizontal red line represents the Bonferroni-corrected p-value threshold for the analysis and the vertical dashed line highlights the Frontal lobe annotation which is significantly enriched in ALS (and other traits). Cell types passing Bonferroni-corrected p-value threshold are enlarged for emphasis. Points are coloured by the tissue group their respective cell type belongs to as described in the legend of panel 1. Notably there is strong evidence of enrichment in the CNS for ALS and all secondary traits studied.



**Figure 2.7: Cell type specific heritability enrichment correlations.**

Regressing the cell type specific coefficients of herability enrichment for ALS vs secondary traits shows a significant correlation between cell types involved in each trait ( $p_{\text{als:bip}}=2.1 \times 10^{-19}$ ;  $p_{\text{als:cog}}=7.1 \times 10^{-16}$ ;  $p_{\text{als:scz}}=2.1 \times 10^{-17}$ ). Points representing cell types are coloured by their tissue of origin, and cell types significant in both traits are enlarged (e.g. frontal lobe). The line represents a linear model fit in R. BIP, Bipolar Disorder; ALS, Amyotrophic Lateral Sclerosis; SCZ, Schizophrenia; Cog VNR, Cognition Verbal Numeric reasoning.

The negative genetic correlation between ALS and cognitive performance is consistent with presentation of cognitive decline in ALS patients. To answer whether this correlation was simply due to shared genetic loci or whether ALS has a causal impact on risk for lower cognitive performance we performed latent causal variable analysis (LCV) (O'Connor and Price 2018). This analysis tests a model where a latent causal variable mediates the genetic correlation between the two traits. It then tests if one trait is highly correlated with the latent causal variable, suggesting that it is partially causal for the second (i.e. a large portion of its genetic component is causal for the second trait.). The analysis returns an estimate for the genetic causality proportion (GCP) and a p value that the GCP value is non-zero. LCV analysis run using a GWAS for ALS (Nicolas et al. 2018) and two GWAS for cognitive performance returns a significant intermediate value GCP (Table 2.6) suggesting that ALS is partially causal for lower cognitive performance (Verbal Numeric Reasoning: GCP=0.61,  $p=5.65 \times 10^{-8}$ ; cognitive performance: GCP=0.56,  $p=8.39 \times 10^{-8}$ ). This is interesting as it indicates that a large portion of the genetic component for ALS has a causal effect on cognition, potentially explaining findings of cognitive decline in ALS patients. This suggests that some future interventions treating the root genetic cause of ALS are likely to also offset cognitive decline if successful. In contrast other genetically correlated traits such as bipolar disorder and schizophrenia do not share a detectable causal relationship with ALS (bipolar disorder: GCP=0.02,  $p=0.75$ ; schizophrenia: GCP=-0.15,  $p=0.45$ ), suggesting that ALS variants do not have a causal impact on schizophrenia or bipolar disorder (or vice versa) (Table 2.6).

**Table 2.6: Latent Causal Variable analysis of ALS and secondary traits.**

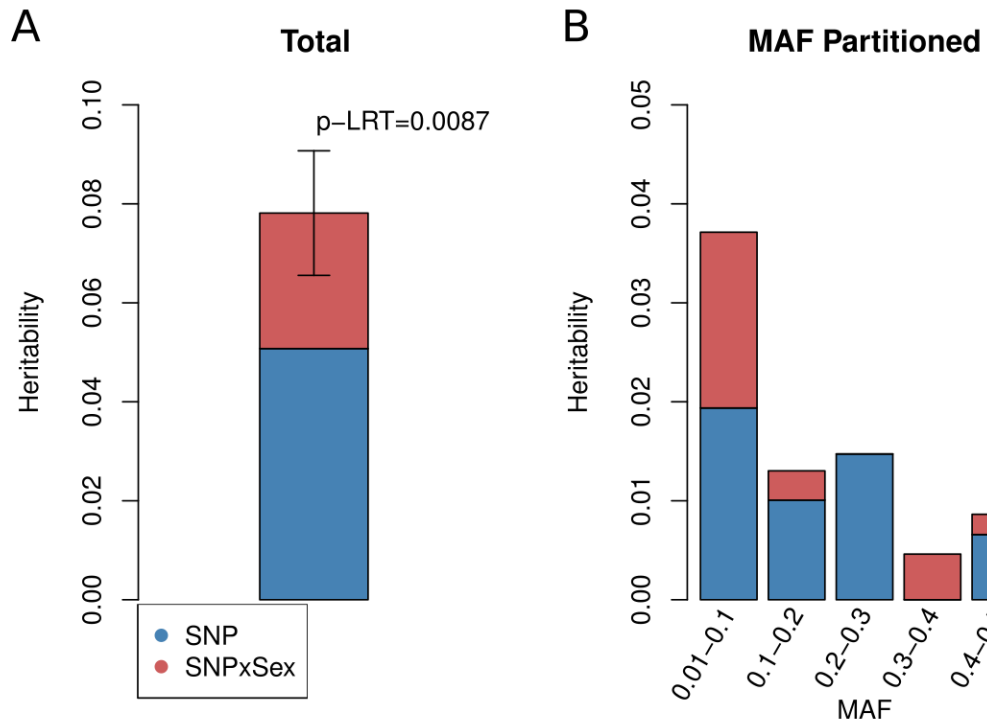
Trait 1	Trait 2	GCP	p GCP=0
ALS	Cognitive performance	0.56	$8.39 \times 10^{-8}$
ALS	Cognitive verbal numeric reasoning	0.61	$5.65 \times 10^{-8}$
ALS	Schizophrenia	-0.15	0.45
ALS	Bipolar	0.02	0.75

Summarises the estimated genetic causality proportion and probability that it is non-zero for ALS and correlated secondary traits.

### 2.3.2 - Sex specific architecture

ALS heritability estimated from mother to daughter pairings in Ireland has been shown to be higher than other parent offspring pairings (Ryan et al. 2019), suggesting a potential sex specific component to the genetics of ALS. Additionally population registers in Italy and the UK show that ALS prevalence is much higher in males pre-menopause age (Manjaly et al. 2010), but closer to parity after, suggesting that sex plays a biological role in disease risk. To explore this hypothesis further we tested whether a fraction of the heritability of ALS could be explained by a gene-sex interaction. We fit sex as an interaction term in GCI-GREML model run on individual-level data from the 2016 ALS GWAS (van Rheenen et al. 2016), and tested its relative likelihood versus a standard model without this interaction using a likelihood ratio test. We observed a significant gene-sex interaction in this ALS GWAS dataset ( $p=8.73 \times 10^{-3}$ ). The total model estimated that SNP based heritability was  $\sim 7.8\%$  (consistent with estimates in the source paper) with a gene-sex interaction accounting for about a third of this heritability (Figure 2.8 A). When partitioned by minor allele frequency the majority of heritability was in the lowest frequency bin (MAF 0.01-0.1), consistent with the source paper for the dataset (van Rheenen et al. 2016), with the gene by sex interaction accounting for close to 50% of this fraction. This result is particularly interesting as heritability estimates in this analysis are from the autosome, suggesting that the gene-sex interaction observed here is not explained by variants on the sex chromosomes interacting with ALS risk. However, it is possible hormones differentially affect expression of risk genes in males and females, resulting in a sex-gene interaction on the autosomes. Alternatively, female hormones may play a protective role against ALS, which may mean that a higher genetic load or a distinct set of risk variants is needed in females. Both endogenous female hormones (de Jong et al. 2013) and exogenous female hormones (e.g. hormone replacement treatment and oral contraceptives) (Rooney, Visser, et al. 2017) have been associated with lower risk of ALS in females, supporting the latter hypothesis. Together with the literature, our results suggest that the genetic architecture of ALS differs in males and females.

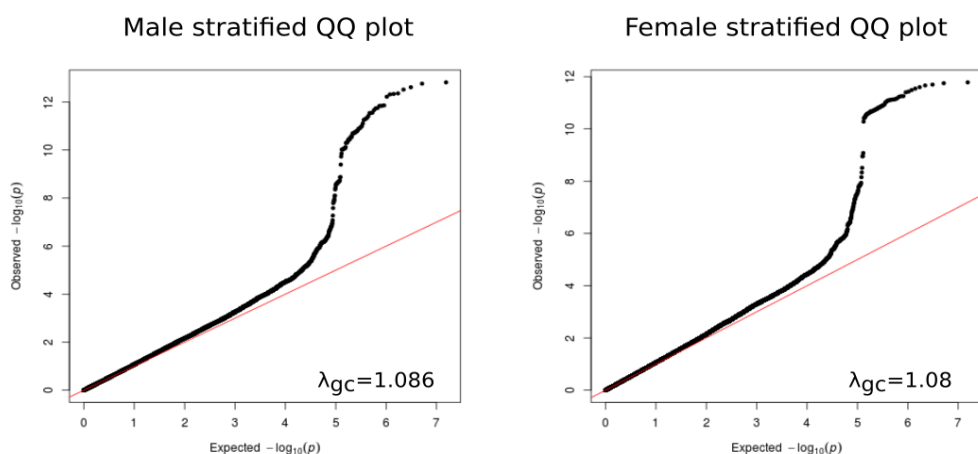
## ALS genotype by sex interaction



**Figure 2.8: Genotype by sex interaction in ALS.**

Barplots displaying the GREML heritability estimates under a genotype by sex interaction model for (A) all SNPs and (B) SNPs partitioned by minor allele frequency (MAF). Main SNP effects are displayed in blue while SNP by sex interactions are displayed in red. Error bars represent the 95% confidence interval for the total heritability estimate.  $p\text{-LRT} = p$  value produced by the likelihood ratio test for a gene by sex interaction versus a model with only genetic effects.

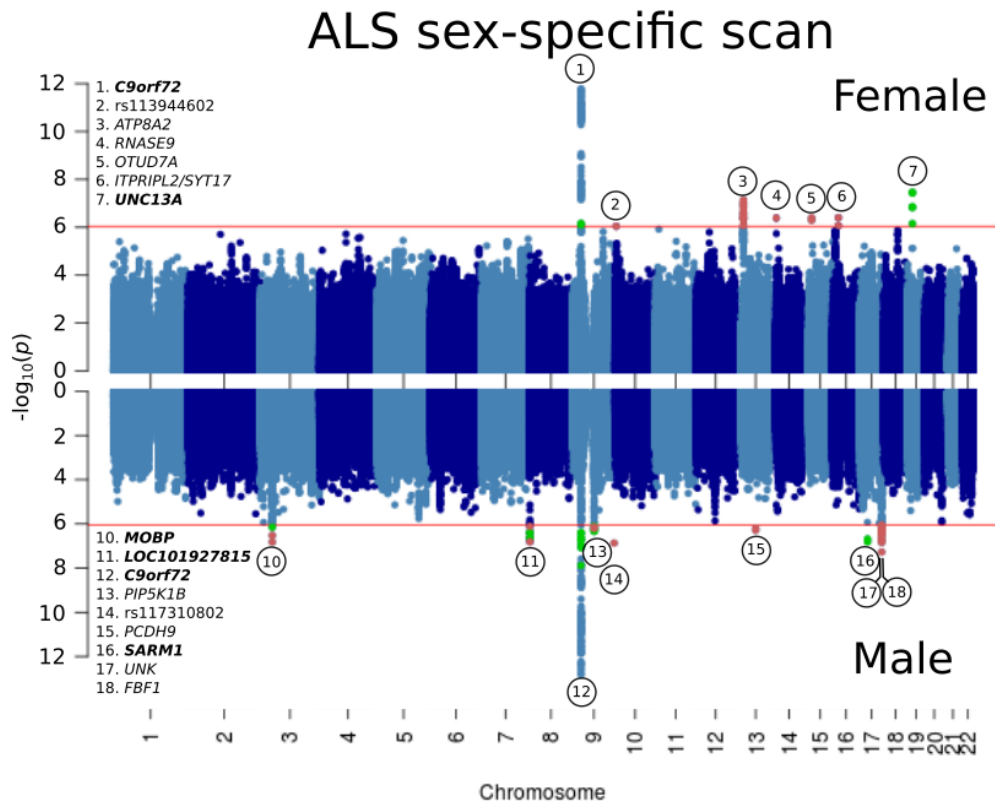
To further explore the nature of the gene-sex interaction we split the GWAS dataset into male-only ( $n=18,730$ ; 7,442 cases) and female-only ( $n=17,322$ ; 5,135 cases) subsets and ran a mixed linear model GWAS for ALS in each dataset. Our goal here was to investigate differences in association signals in ALS in males and females, and additionally to assess if there are global differences in the genetic architecture of the disease between males and females. The datasets produced well-controlled summary statistics as seen by inspection of the qq-plots (Figure 2.9) and Lambda gc estimates which are close to one (Male = 1.086, Female = 1.08). We performed a genome-wide scan using these sex-stratified summary statistics to identify sex-specific variants, which showed significant association with ALS in one sex but not the other. For our “sex-specific” scan we classified variants as “strongly” sex-specific if they were significantly associated with ALS at 5% FDR in one sex ( $p_{\text{als}|\text{sex}1} < 9.1 \times 10^{-7}$ ), and had no evidence of association in the other ( $p_{\text{als}|\text{sex}2} > 0.05$ ), and “suggestively” sex-specific if they were associated with ALS in one sex ( $p_{\text{als}|\text{sex}1} < 9.1 \times 10^{-7}$ ) and only nominally associated in the other ( $9.1 \times 10^{-7} < p_{\text{als}|\text{sex}2} < 0.05$ ) (Figure 2.10). Variants meeting these criteria were then clumped to remove signals coming from the same LD region, and the closest gene was reported (Table 2.7). FUMA analysis of genes adjacent to independent loci with specific association in males or females indicates they are significantly upregulated in several brain annotations from GTEx expression data (Table 2.8). Additionally genes identified using this scan show enrichment in gene ontologies for several biological processes, many of which involve neuronal growth or cell projection (Table 2.9).



**Figure 2.9: Inflation in sex stratified GWAS for ALS.**

QQ-plots for sex stratified GWAS for ALS. Male (left) and female (right) show similar levels of inflation ( $\lambda_{gc} \sim 1.08$ ) and appear well controlled (i.e. do not deviate excessively from the expected null distribution).





**Figure 2.10: Genome wide sex-specificity scan for ALS.**

A Miami plot of a female-only (top) and male-only (bottom) GWAS for ALS. The horizontal red lines represent the 5% FDR threshold for significance ( $p=9.1 \times 10^{-7}$ ). SNPs in green show a suggestive signal of sex specificity ( $p_{\text{sex1}} < 9.1 \times 10^{-7}$  and  $9.1 \times 10^{-7} < p_{\text{sex2}} < 0.05$ ), and SNPs in red show a strong signal of sex-specificity ( $p_{\text{sex1}} < 9.1 \times 10^{-7}$  and  $p_{\text{sex2}} > 0.05$ ). Loci passing the scan are numbered, and the closest gene is reported (or rsid of lead SNP if no genes are proximal). Known loci are highlighted in bold.

**Table 2.7 Sex-specific ALS GWAS scan results.**

CHR	SNP	BP	p fem	p male	p diff	Sex-specific	Closest Gene	Known
3	rs816487	39487490	3.44E-02	7.10E-07	3.30E-02	Suggestive - Male	<i>MOBP</i>	Known
3	rs28829975	39510671	1.38E-01	1.52E-07	5.04E-03	Strong - Male	<i>MOBP</i>	Known
8	rs6996532	2417678	6.50E-02	1.56E-07	1.10E-02	Strong - Male	<i>LOC101927815</i>	Known
9	rs10511816	27468461	1.56E-04	1.32E-08	1.37E-01	Suggestive - Male	<i>MOB3B, C9orf72</i>	Known
9	rs4879515	27482235	7.97E-07	8.91E-05	5.12E-01	Suggestive - Female	<i>MOB3B, C9orf72</i>	Known
9	rs139185008	27491942	1.53E-06	8.42E-08	5.32E-01	Suggestive - Male	<i>MOB3B, C9orf72</i>	Known
9	rs2492816	27565105	1.19E-04	1.18E-07	2.62E-01	Suggestive - Male	<i>C9orf72</i>	Known
9	rs7864502	27583128	1.03E-06	1.92E-07	7.57E-01	Suggestive - Male	<i>C9orf72</i>	Known
9	rs2453556	27586162	6.86E-07	6.19E-06	7.90E-01	Suggestive - Female	<i>C9orf72</i>	Known
9	rs10869323	71424040	1.68E-01	7.13E-07	1.42E-06	Strong - Male	<i>PIP5K1B, FXN</i>	Novel
9	rs117310802	138159175	6.74E-01	1.36E-07	2.38E-04	Strong - Male	-	-
10	rs113944602	4185278	9.03E-07	6.69E-01	1.09E-04	Strong - Female	-	-
13	rs112215101	26166022	7.17E-08	9.67E-01	8.06E-05	Strong - Female	<i>ATP8A2</i>	Novel
13	rs2419324	26254841	8.24E-07	1.96E-01	7.61E-03	Strong - Female	<i>ATP8A2</i>	Novel
13	rs9599108	66957056	7.50E-01	5.20E-07	5.02E-04	Strong - Male	<i>PCDH9</i>	Novel
14	rs1268411	21018970	4.09E-07	5.62E-02	2.54E-02	Strong - Female	<i>RNASE9</i>	Novel
15	rs183786557	31913997	4.00E-07	8.23E-01	3.33E-04	Strong - Female	<i>OTUD7A</i>	Novel
16	rs58808799	19154719	3.97E-07	4.68E-01	1.66E-03	Strong - Female	<i>ITPRIPL2, SYT17</i>	Novel
17	rs35714695	26719788	2.23E-04	1.56E-07	2.13E-01	Suggestive - Male	<i>SARM1</i>	Known
17	rs11652539	73797775	5.30E-01	4.21E-07	2.54E-05	Strong - Male	<i>UNK</i>	Novel
17	rs56174511	73903482	3.03E-01	5.36E-08	1.79E-06	Strong - Male	<i>FBF1</i>	Novel
19	rs12608932	17752689	3.44E-08	1.28E-04	2.67E-01	Suggestive - Female	<i>UNC13A</i>	Known

Lists independent SNPs (clumped) passing 5% FDR threshold for association with ALS in one sex but not the other (suggestive evidence), or passing the 5% FDR threshold in one sex but not even reaching nominal significance in the other (strong evidence). The closest gene or genes are listed for each variant. P values are provided for the male and female only GWAS and the p value for the difference in their effects (p diff) is also listed. Rows are coloured to represent stretches of loci in or adjacent to the same gene.

**Table 2.8: Tissue specific upregulation for sex-specific genes.**

GeneSet	N genes set	N genes overlap	p	FDR adj p	Female-specific genes	Male-specific genes
Brain Amygdala	1825	6	1.29E-04	6.97E-03	<i>ATP8A2, OTUD7A, SYT17, UNC13A</i>	<i>MOBP, PCDH9</i>
Brain Anterior cingulate cortex BA24	2376	6	5.45E-04	2.94E-02	<i>ATP8A2, OTUD7A, SYT17, UNC13A</i>	<i>MOBP, PCDH9</i>
Brain Frontal Cortex BA9	3142	7	3.33E-04	1.80E-02	<i>ATP8A2, OTUD7A, SYT17, UNC13A</i>	<i>MOBP, PCDH9, PIP5K1B</i>
Brain Hippocampus	1970	6	1.97E-04	1.06E-02	<i>ATP8A2, OTUD7A, SYT17, UNC13A</i>	<i>MOBP, PCDH9</i>
Brain Hypothalamus	2539	6	7.78E-04	4.20E-02	<i>ATP8A2, OTUD7A, SYT17, UNC13A</i>	<i>MOBP, PCDH9</i>
Brain Substantia nigra	1650	6	7.37E-05	3.98E-03	<i>ATP8A2, OTUD7A, SYT17, MOB3B</i>	<i>MOBP, PCDH9, MOB3B</i>

Tissue specific enrichments for genes closest to sex-specific loci (Table 2.7). Male- and female-specific genes are enriched in genes highly expressed in the brain (significant at 5%FDR)

**Table 2.9: Gene ontology annotations enriched for sex-specific genes.**

GeneSet	N genes set	N genes overlap	p	FDR adj p	Female-specific genes	Male-specific genes
GO: POSITIVE REGULATION OF GROWTH	266	4	3.03E-06	2.22E-02	<i>ATP8A2, SYT17, UNC13A</i>	<i>FXN</i>
GO: REGULATION OF DEVELOPMENTAL GROWTH	338	4	7.8E-06	2.87E-02	<i>ATP8A2, SYT17, UNC13A</i>	<i>FXN</i>
GO: NEUROMUSCULAR PROCESS CONTROLLING POSTURE	16	2	1.76E-05	4.03E-02	<i>ATP8A2</i>	<i>FXN</i>
GO: POSITIVE REGULATION OF CELL GROWTH	160	3	3.25E-05	4.03E-02	<i>SYT17, UNC13A</i>	<i>FXN</i>
GO: REGULATION OF NEURON PROJECTION DEVELOPMENT	487	4	3.27E-05	4.03E-02	<i>ATP8A2, SYT17, UNC13A</i>	<i>SARM1</i>
GO: REGULATION OF DENDRITE EXTENSION	23	2	3.71E-05	4.03E-02	<i>SYT17, UNC13A</i>	-
GO: NEURON DEVELOPMENT	1094	5	4.59E-05	4.03E-02	<i>ATP8A2, SYT17, UNC13A</i>	<i>SARM1, UNK</i>
GO: POSITIVE REGULATION OF DEVELOPMENTAL GROWTH	181	3	4.69E-05	4.03E-02	<i>ATP8A2, SYT17, UNC13A</i>	-
GO: CELLULAR COMPONENT MORPHOGENESIS	1111	5	4.94E-05	4.03E-02	<i>ATP8A2, SYT17, UNC13A</i>	<i>SARM1, UNK</i>

Gene ontology enrichments for genes identified to be closest to sex-specific loci (Table 2.7). All annotations listed are significant at 5% FDR.

Sex-specific SNPs were identified in several known ALS loci in this scan (Table 2.7; e.g. *MOBP, UNC13A, SARM1, C9orf72, LOC101927815*). A key example of strong sex-specificity at a known ALS locus can be seen at the *MOBP* locus, which shows strong association with ALS in males ( $p_{\text{male}}=1.5 \times 10^{-7}$ ), but almost no signal in females ( $p_{\text{fem}}=0.14$ ). In contrast, the *UNC13A* locus reaches genome wide significance ( $p_{\text{fem}}=3.3 \times 10^{-8}$ ) in females, but has a nominal signal in males ( $p_{\text{male}}=1.3 \times 10^{-4}$ ), hence we would classify it as suggestively female-specific locus (Table 2.7). Of the known ALS loci identified with sex-specific association, *MOBP* and *SARM1* have an interesting history in

ALS GWAS in that they show strong association in some GWAS, but not others, which could be attributed to sex-specific effects like those seen here confounding standard unstratified GWAS. For example the *MOBP* locus was only identified in the source GWAS for this dataset (van Rheenen et al. 2016) through application of the LMM model, and missed using a standard meta-analysis in the same dataset and a later meta-analysis with more samples (Nicolas et al. 2018). This could be due to model misspecification as meta analyses assume homogeneous effects across all sub cohorts, whereas the association between *MOBP* and ALS appears to be sex-dependent, and hence may be heterogeneous across cohorts depending on their sex balance. Similarly *SARM1* was only identified in the 2016 ALS GWAS (van Rheenen et al. 2016) and not the 2018 ALS GWAS (Nicolas et al. 2018), and appears to have also had a male-specific effect. It is possible that the sex-biased sampling in the 2018 ALS GWAS (~60% female controls), may have reduced power to detect these loci in the 2018 ALS GWAS, despite its greater sample size.

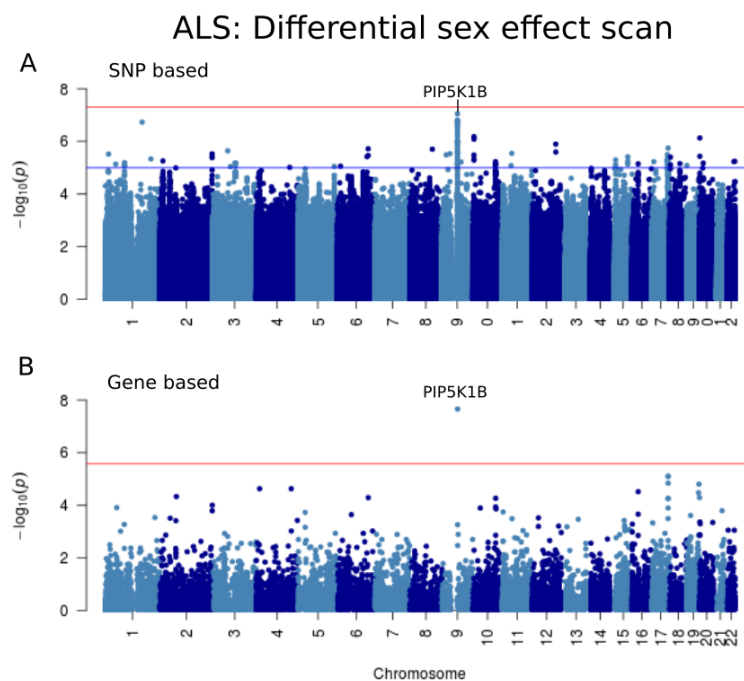
Of the known ALS loci identified in the scan, *C9orf72* is interesting in that it seemingly harbours SNPs with sex-specific effects for both males and females. We note that sex-specificity is only ever suggestive in each of these variants however, with quite low p-values for both males and females indicating that these hits may be false positives for sex-specific association resulting from false negative association with ALS in either males or females. This may be due to the lower power of these stratified GWAS (caused by splitting sample size), which means we should be cautious of “suggestively sex-specific” variants which may sometimes be real associations in both males and females, but simply undetected in one sex due to power issues. In spite of this caveat, sex has been shown to have modifying effects on the impact of *C9orf72* in ALS (Rooney, Fogh, et al. 2017), so there is potential for these effects to be real.

Our sex-specific scan also found strong evidence of sex-specific associations in several novel loci which may have been missed by unstratified GWAS, many of which are near or in feasible candidate ALS genes. For example in our female-only GWAS we found several variants in the *ATP8A2* gene (e.g. rs112215101,  $p_{\text{fem}}=7.17 \times 10^{-8}$ ) with no signal in males ( $p_{\text{male}}=0.97$ ). This gene is widely expressed in the brain and spinal cord, and mutations in it have been shown to cause axonal degeneration in wobbler-lethal mice (X. Zhu et al. 2012), making it a sensible candidate given the involvement in ALS. We also identify strong evidence of sex-specificity near *PCDH9* in males ( $p_{\text{male}}=5.2 \times 10^{-7}$ ) with no signal in females ( $p_{\text{fem}}=0.75$ ), which shows high expression in the brain and spinal cord (GTEx

portal 2/6/2020) alongside involvement in major depressive disorder risk (Xiao et al. 2018), meaning it may be involved in the genetic overlap between ALS and psychiatric traits discussed above. Notably *PCDH9* was also identified in a cFDR scan for ALS and FTD (Karch et al. 2018), suggesting that it is a plausible candidate ALS gene. While these and several genes near other novel sex-specific ALS hits show heightened expression in brain tissue (e.g. *OUTD7A*, *SYT17* and *PIP5K1B*; GTEx portal 2/6/2020 and Table 2.9), and hence may have some feasible mechanistic tie to ALS pathology, some of our novel hits are near genes most abundantly expressed in sex-specific tissues such as the testis (*FBF1*, *RNASE9*; GTEx portal 2/6/2020) or the cervix, ovary and uterus (*UNK*; GTEx portal 2/6/2020), which are more difficult to explain as these tissues are not directly linked to tissues of disease onset. However the latter genes are likely differentially expressed in males and females due to their presence in these tissues, meaning variants within them may differentially expose males and females to disease risk due to the relative abundance of their transcripts. This scan is quite permissive and results should be interpreted as suggestive hits until replicated in an independent sex-stratified GWAS, ideally with standard GWAS multiple testing thresholds.

We also performed a “sex-differentiated” scan to identify variants with significantly different effect size in males and females (Figure 2.11). This scan is best powered for identifying variants with opposite effects in each sex, but could theoretically identify variants with majorly different effects in the same direction in males and females. As in the source paper for this method (Randall et al. 2013), this scan showed no significant hits at a 5% FDR, despite showing clear statistical inflation ( $\lambda=1.073$ ), and a borderline genome wide significant peak on chromosome 9 (Figure 2.11 A,  $p=8.9\times 10^{-8}$ ), which is suggestive evidence of the variant having differing effects in males and females. This lack of genome wide significance here may simply be a matter of power which will be overcome by sample size increases in future sex-stratified ALS GWAS, as seen for BMI which returned no sex differentiated loci in the initial study applying this method (Randall et al. 2013), but several in a follow up study with more samples (Winkler et al. 2015). Notably the peak on chromosome 9 in this scan does not correspond to *C9orf72*, but instead falls within *PIP5K1B*, a gene which also harboured hits in our male-only ALS sex-specific scan. Gene-based analysis using p-values from this scan in MAGMA (de Leeuw et al. 2015) shows SNPs in this gene are enriched for differential effects in males and females, suggesting this might be a real signal. This gene, and its neighbour *FXN* are implicated in Friedreich's Ataxia (Bayot et al. 2013), where it is likely silenced by the long *FXN* repeat expansion flanking it (Bayot and Rustin 2013). As both ALS and Friedreich's

Ataxia are neurodegenerative disorders mediated by large repeat expansions this seems like a feasible candidate ALS gene. It is also of note that this locus falls in 9q21-22 region of the genome, which was identified as associated with ALS-FTD in early linkage studies (Hosler et al. 2000), but has not been replicated in following years, which may be due to sex composition of the cohorts studied. Aside from the *PIP5K1B* locus, genes from our sex differentiated scan did not show clear enrichment in tissues or gene ontologies when analysed with FUMA, meaning further work will need to be done to characterise which functional regions are affected differently in ALS across sexes.



**Figure 2.11: Genome wide sex-differentiation scan for ALS.**

Manhattan plots for (A) SNPs and (B) aggregate p-values across genes (MAGMA) for our genome-wide sex-differentiation scan for ALS which tests for loci with divergent effects in males and females on ALS risk. The genome-wide significance line is plotted in red and the suggestive significance line is plotted in blue. For the MAGMA analysis genome wide significance is corrected for the number of genes tested (0.05/18900). Notably only one clear peak stands out (*PIP5K1B* on chromosome 9), which does not quite reach genome wide significance in the SNP based scan. The *PIP5K1B* gene in which this SNP is located does reach significance in our MAGMA aggregated gene based test, and was also identified by our sex-specific scan (Figure 2.10, Table 2.7).

To further explore the differences in genetic architecture of ALS between males and females we ran univariate LD score regression on the male and female GWAS to estimate the SNP-based heritability from each. Notably the male-only sample had an extremely low SNP-based heritability estimate (LDSC Male:  $h^2=0.001$ ;  $SE=0.0124$ , Table 2.10) which was not significantly non-zero. In contrast the female-only GWAS had a non-zero SNP-

based heritability estimate on par with the estimate from the meta-analysis of the full dataset (LDSC Female:  $h^2=0.0434$ ;  $SE=0.0144$ , Table 2.10), with substantially larger SE than the full analysis, probably driven by lower sample size. This higher SNP-based heritability estimate in females could indicate that ALS is more polygenic in females than in males, especially when noting the LDSC intercept, which was much higher in males than females (Male = 1.0896; Female = 1.0533, Table 2.10), indicating that a higher proportion of inflation in the male-only GWAS can be attributed to confounding. Considering that the male- and female-only GWAS show similar levels of inflation overall, this hypothesis seems likely.

**Table 2.10: Sex-partitioned SNP-based heritability and inflation.**

Analysis	Cases	Controls	LDSC-intercept (SE)	LGC	Ratio	$h^2$ SNP (SE)
Male	7,442	11,288	1.0896 (0.0073)	1.086	0.9903 (0.0803)	0.001 (0.0124)
Female	5,135	12,187	1.0533 (0.0068)	1.08	0.6393 (0.0814)	0.0434 (0.0144)

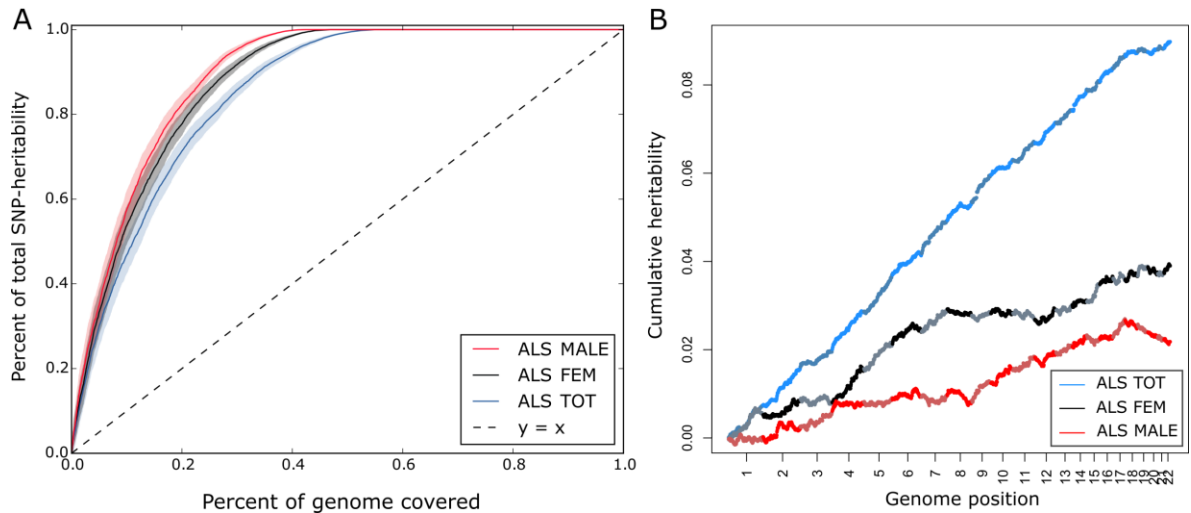
Summaries of measures of inflation (LDSC-intercept and LGC) and SNP-based heritability ( $h^2_{SNP}$ ) from LD score regression on male-only and female-only GWAS. Ratio here refers to  $(LDSC\ intercept - 1) / (\text{mean}(\chi^2) - 1)$  and should reflect the proportion of inflation contributed by confounding. While the ratio in the female-only GWAS is high, indicating it is quite heavily confounded, the male-only GWAS ratio is near one indicating inflation is almost entirely confounding.

However, the LDSC model assumes even spread of heritability across the genome, which is at odds with our MAF partitioned GREML result (Figure 2.8), which shows higher heritability in lower frequency variants, suggesting that the LDSC model may be misspecified here. To test this hypothesis further within another framework, we ran HESS (Heritability Estimation from Summary Statistics) on the male-only and female-only ALS GWAS to re-evaluate sex-partitioned heritability and test for patterns of enrichment of heritability across the genome. The output from HESS can be ordered by contribution to total heritability to create a plot of fraction of heritability explained by cumulative fraction of the genome, which can be thought of as a visualisation of the polygenicity of the trait (Figure 2.12 a). We noted that the male-specific polygenicity curve was marginally steeper than the female-specific polygenicity curve, suggesting that its heritability was explained by fewer variants and indicating that it may be less polygenic than female-specific ALS. However as both male and female curves are steeper than the total ALS curve, there is a possibility that this discrepancy is simply a power issue. Notably local male-specific heritability estimates for ALS were lower than local female-specific estimates using HESS (Figure 2.12 b), supporting our observations from LD score regression above. Additionally this analysis showed differing patterns of local heritability across the 1,700 independent



regions tested (Figure 2.12 b), suggesting a different local architecture. However due to large standard errors in the sex specific genome-wide heritability estimates produced by HESS, which encompass zero, it may be unwise to draw strong conclusions from this analysis (Total  $h^2_{\text{male}}=0.022$ ,  $SE=0.07$ ;  $h^2_{\text{female}}=0.039$ ,  $SE=0.08$ ).

## ALS: Sex stratified HESS



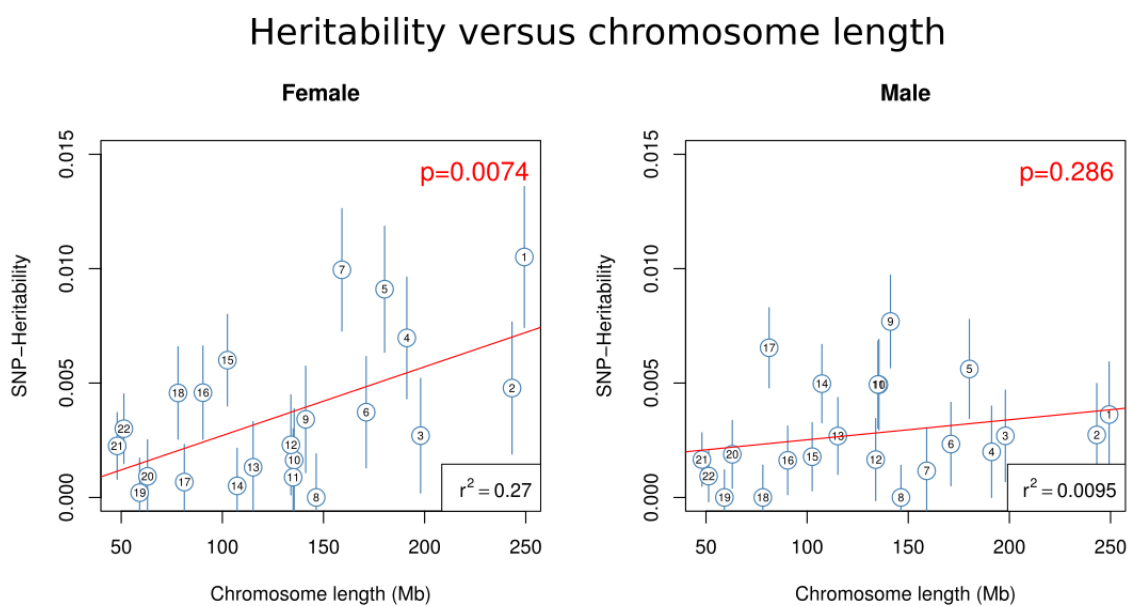
**Figure 2.12: Sex stratified local heritability compared.**

A.) HESS contrast polygenicity curves for the female-only (ALS FEM), male-only (ALS MALE) and full GWAS (ALS TOT). The curves describe the cumulative percentage of total SNP-based heritability attributable to a fraction of the genome (i.e. the top 10%, top 20% etc). Where heritability is completely uniform we expect this to produce a 1:1 line. The full analysis is closest to the 1:1 line, followed by the female specific, and finally the male specific analysis in line with our LDSC intercept estimate suggesting that the male specific GWAS has the lowest polygenicity.

B.) HESS cumulative genome-wide heritability plots show the relative patterns of heritability for the full GWAS (ALS TOT), the female-only GWAS (ALS FEM) and the male-only GWAS (ALS MALE).

Due to the statistically zero heritability of ALS in males using univariate LDSC and HESS, it was not possible to estimate genetic correlation between males and females based on summary statistics. We therefore instead ran univariate GREML analysis on individual-level male- and female-specific data to improve power in estimating sex-specific heritability. The female-specific analysis showed a SNP-heritability on par with the full analysis from the source paper (GREML Female:  $h^2=0.08$ ;  $SE=0.01$ ), and a linear relationship between SNP-heritability and chromosome length (Figure 2.13;  $r^2=0.27$ ;  $p=0.0074$ ), suggesting that ALS is polygenic in females. In the male-only GREML analysis SNP heritability was significantly non-zero (unlike in LDSC), and only slightly lower than the female heritability (GREML Male:  $h^2=0.06$ ;  $SE=0.008$ ). However there was

no significant linear relationship between SNP-heritability and chromosome length in males (Figure 2.13,  $r^2=0.009$ ,  $p=0.286$ ), with heritability instead appearing to be concentrated disproportionately on chromosomes 5,9,10,11,15 and 17 providing further evidence that ALS might not be as polygenic in males. While male and female total heritability estimates overlap statistically, it appears from HESS (Figure 2.12b) and partitioned univariate GREML analysis (Figure 2.13) that they are driven by quite different genetic loci across the genome. For example a linear regression of heritability per chromosome in males and females shows zero correlation ( $r^2=-0.045$ ,  $p=0.99$ ), suggesting variants contributing to the heritability of ALS in males may be distributed differently to those in females.



**Figure 2.13: Heritability vs chromosome length partitioned by sex.** Male and female heritability estimates per chromosome (GREML) plotted against chromosome length demonstrate sex differences in the spread of genetic risk for ALS. Females (left) show a positive linear relationship between chromosome length and ALS heritability, suggesting that variants contributing to heritability are spread genome-wide, indicating polygenicity. In contrast Male ALS heritability estimates do not appear to have a linear relationship with chromosome length (right), and instead are concentrated in a number of key chromosomes likely representing a more oligogenic architecture.

Finally we ran a bivariate GREML analysis to estimate genetic correlation between ALS in males and females, given that this was impossible using summary statistics due to the statistically zero estimates of ALS heritability in males. Contrasting our aforementioned evidence of apparent differences in distribution of heritability in males and females,

bivariate GREML analysis reveals a high genetic correlation between ALS in males and females ( $r_g=1$  (0.75-1.25)). While the point estimate for the genetic correlation between males and females for ALS is one, it is important to note the wide confidence interval which suggests correlation may not be perfect, indicating that males and females share most but not necessarily all genetic variation for ALS (i.e. risk variants might be highly overlapping but not necessarily identical). Larger sample sizes should improve the precision of this estimate to allow us to better understand how similar ALS genetics are in males and females.

## 2.4 - Discussion

In this chapter we have expanded the insights into the genetics of ALS yielded from GWAS by exploring (i.) the shared genetic components with secondary neuropsychiatric and cognitive traits and (ii.) the potential differential genetic architecture of ALS in males and females. These analyses enable us to better understand the genetic root of clinically observed extra motor symptoms seen in patients, and genetically contextualize the difference in observed risk and heritability differences between males and females.

### 2.4.1 - Genetic overlap with psychiatric and cognitive traits

Towards our first aim (i) our analysis yielded a number of interesting results regarding the shared genetic architecture between ALS and secondary neuropsychiatric and cognitive traits. We uncovered a novel genetic overlap between ALS and bipolar disorder using LD score regression, alongside replicating known correlations with schizophrenia and lower cognitive performance. We observed that these traits all had enriched SNP heritability in genes highly expressed in the CNS, suggesting that all share a common tissue of effect. Multi-trait analyses using these correlated traits yielded several novel loci potentially associated with ALS and these secondary traits, and showed pleiotropic loci are enriched in genes highly expressed in the brain. ALS and cognition-related genes showed functional enrichment in several gene ontologies relating to neuronal function and development, solidifying the role of shared biological pathways in a shared tissue of effect. Beyond directly observed genetic correlations, structural equation model analysis revealed evidence that ALS shares a common genetic factor with an extended profile of psychiatric traits, pointing towards a genetic basis for the enrichment of psychiatric traits in ALS kindreds. However, latent causal variable analysis indicated that ALS does not have a causal genetic relationship with bipolar or schizophrenia, suggesting that while they share genetic features, exposure to one trait does not causally lead to the other. In contrast, we found evidence that ALS has a causal genetic impact on lower cognitive performance, suggesting that the exposure of having ALS may directly lead to the cognitive phenotypes observed in patients. This result highlights the differences between simple genetic pleiotropy across phenotypes and causal genetic links between phenotypes, and potentially explains the abundance of enriched functional annotations when analysing ALS and cognition in tandem, but not when analysing ALS and psychiatric traits.

The novel genetic correlation we observed between ALS and bipolar disorder, and

replication of the correlation with schizophrenia strengthens the emerging narrative that ALS has a shared genetic component with psychiatric traits (R. L. McLaughlin et al. 2017), suggesting that possibly overlapping biological pathways or mechanisms drive ALS and psychiatric traits. Supporting this we found that ALS, bipolar disorder and schizophrenia all have enriched heritability in genes expressed in the central nervous system, indicating that their shared genetic features and coincidence in families may partially reflect perturbations of genes highly expressed in the same tissue. In our attempts to identify pleiotropic loci shared between ALS and psychiatric traits we identified putative signals in a number of novel putative ALS genes (*CNNM2*, *SRGAP1*, *KRT18P55*, *SRGAP1*, *NCKAP5L*, *SPIRE1*, *AS3MT*) and one known ALS gene (*C9orf72*). Functional analysis showed that pleiotropic genes appear to be particularly highly expressed in the cerebellum and frontal cortex, but show no significant functional enrichment across all gene ontologies, indicating that a shared burden in genes highly expressed in the brain may be the major feature responsible for their genetic overlap. However, given the relatively small sample sizes of both the ALS and secondary psychiatric trait datasets, we cannot conclusively rule out the existence of shared pathways between these traits. Future studies may be better powered to identify the functional overlap between ALS and psychiatric traits.

Interestingly, despite a near 10-fold increase in sample size for MDD and ADHD GWAS compared to the initial study of genetic correlation between ALS and psychiatric traits (R. L. McLaughlin et al. 2017), bivariate LD score regression analysis of our data suggests that neither of these traits shares significant genetic correlations with ALS, implying secondary trait sample size is unlikely to be driving these null results. Moreover neither anxiety nor PTSD (which were absent from the aforementioned study) showed significant genetic correlation with ALS, despite evidence of increased diagnosis of stress related disorders and anxiety both before and after diagnosis with ALS (Longinetti et al. 2017). It is possible that the increased rates of these disorders in ALS patients is instead driven by a non-genetic component, such as the burden caused by ALS symptoms, though this does not explain the increased rate of diagnosis in the years preceding ALS diagnosis. While these results could be true negatives, it is important to note our study focused solely on common variation, while ALS is expected to have a rare variant architecture (van Rheenen et al. 2016), suggesting that the genetic overlap with these traits may be substantially underestimated. The use of large whole-genome sequencing datasets for ALS (e.g. those being developed by the Project MinE sequencing consortium (van Rheenen et al. 2018)) and secondary psychiatric traits will allow closer examination of overlapping rare variant burden in coming years. Nevertheless, when modelling for a

shared genetic component across all studied psychiatric traits and ALS using genomicSEM (Grotzinger et al. 2019) we found that either a single shared latent genetic factor or two correlated latent genetic factors best described the genetic covariance structure between these traits, indicating that there may in fact be a common genetic factor shared between ALS and psychiatric traits, even for traits that appeared non-correlated in the bivariate analysis.

Our bivariate LD-score regression analysis also identified negative genetic correlations with two measures of cognition (verbal numeric reasoning and cognitive performance). This result adds to prior evidence that polygenic risk for ALS is associated with lower measures of cognition in the UK Biobank and the previously observed nominal genetic correlation with a proxy measure of cognitive performance (Hagenaars et al. 2018; Bandres-Ciga et al. 2019). Given the high rates of cognitive impairment in ALS patients, the potential existence of shared genetic pathways between ALS and cognition has many implications for understanding the root of these extra-motor symptoms and potentially treating them. Notably we observed that ALS not only shares a tissue of effect with cognition (CNS), but appears to be partially genetically causal for lower cognition, implying ALS doesn't just share pathways with decreased cognitive function, but that the genetic exposure of ALS may result in these symptoms. Hence identifying treatments for ALS may not only alleviate motor symptoms, but could also improve negative cognitive symptoms in patients, greatly improving quality of life for both patients and caregivers. Notably while we have identified a number of putative pleiotropic loci affecting both ALS and cognition, whole-genome or exome sequencing patients with cognitive symptoms may further elucidate the overlap between these traits, providing a more direct assessment of the shared genetic architecture of cognition and ALS.

Our multi-trait analysis between ALS and cognition showed significant enrichment in gene ontologies associated with neurone development, differentiation and general cell morphogenesis, all of which may play a role in neuronal vulnerability in ALS. On top of this we also saw enrichment in the *MECP2* reactome, which affects transcription in the CNS and is a known contributor to many neurological diseases (Chahrour et al. 2008). *MECP2* expression is affected by the known ALS gene *FUS*, and has been observed to colocalise in *FUS* aggregates (Coady and Manley 2015), hence it is feasible that ALS pathology involves or leads to gene expression dysregulation via *MECP2*. It is possible that the disruption of this pathway plays a role in the extra-motor cognitive symptoms seen in ALS patients, given its widespread effects on gene expression central to the nervous system. *MECP2* is associated with a number of neurodevelopmental disorders including autism

and Rett syndrome whose core shared phenotypic features include deficits in cognition and motor function, both of which are apparent in ALS (Gonzales and LaSalle 2010), lending some support to the possible involvement of the *MECP2* reactome in ALS. Interestingly *MECP2* is located on the X-chromosome, which is typically excluded in GWAS analysis due to dosage imbalance in males and females, hence it could not be detected directly in datasets used in this study. It would be worthwhile investigating if the dosage of this gene has any role in the sex differences in ALS risk and heritability discussed in the second half of this chapter. The observed enrichment of these neuronal developmental biological processes and the *MECP2* reactome provide novel insights into the shared pathways and biological mechanisms between ALS and cognition which may inform future study into mechanism and treatment of extra-motor cognitive symptoms. However, careful replication of these observations may be needed given the limitations of our study, in particular the small sample size of the ALS GWAS used and the lack of a replication cohort.

While the novel loci identified in our multi-trait analysis may represent true pleiotropic loci shared between ALS and secondary traits, there is a high chance that many of them are false positives. One noted cause of this is potential misspecification of the MTAG (Turley et al. 2018) model which assumes that all traits share the same variance-covariance matrix of effect sizes across traits, meaning a SNP which has a null effect in one trait but a non-null effect in the others will have an effect biased away from zero. Given many variants identified as pleiotropic with this method showed extremely weak association ( $p > 0.1$ ) in the base ALS GWASes, implying they may simply be loci associated with the secondary trait assayed, our results fit this profile of being biased from null effects to non-null effects. We also observed high maxFDR values for all pairings of traits, confirming that MTAG analysis is likely to be biased. To address this bias we applied cFDR analysis as an alternative method to detect pleiotropic loci by conditioning association signals in the primary trait on association with a secondary trait. While this method identified some shared loci with our MTAG analysis (e.g. *EXOC4* and *EFTUD1* for cognition), it did not replicate a large number of our MTAG hits, in particular those with weak association in the main ALS GWAS indicating that these are likely false positives. cFDR also identified a number of novel loci missed by MTAG which have more clear evidence of association in the single trait ALS GWAS, including a number of known ALS loci, indicating that it is likely more robust in this setting. Some of these loci identified using cFDR on ALS conditioned on cognition have previously been identified in a cross-trait analysis with disease of FTD spectrum (*GGNBP2*, *MAPT*, *KANSL1*, *NSF*) (Karch et al. 2018), suggesting they are likely associated with the cognitive and behavioural profile of ALS.

However this analysis of pleiotropy between ALS and traits from the FTD spectrum was also run on the 2016 ALS GWAS, so it is not a fully independent replication of these loci. These genesets are enriched in the CNS, meaning they are feasible candidate genes for involvement in ALS pathology. However, replication of these signals in larger ALS GWAS is needed before any concrete action is taken with regards these loci.

Our multi-trait analysis may also be biased by sample overlap in control panels across GWAS summary statistics investigated. Shared controls lead to a positive correlation between the principal and secondary phenotype studied even at null SNPs, leading to an inflation of false positives in cFDR analysis. While adjusted cFDR methods have been developed to accommodate for shared controls between principal and conditional traits (Liley and Wallace 2015), they require knowledge of the number of overlapping samples, which we unfortunately could not estimate or correct due to lack of access to the raw individual level genotype data for secondary traits. In the absence of this correction we would expect an increase in the number of false positives emerging from this analysis. A future direction for this analysis would be to acquire access to individual level data for the secondary datasets to fully assess this issue. However given the current power issues in the ALS dataset, it may be more profitable to wait for the release of larger ALS GWAS which are currently underway. For example an ongoing GWAS effort by the Project MinE GWAS consortium with large numbers of samples (N~150,000) is expected to be released in the next year or so (van Rheenen et al., personal communication), which may solve power issues seen in the MTAG analysis, and improve the return on the cFDR and MTAG analysis. Additionally this improved sample size may boost our power to detect genetic correlations with an extended set of psychiatric traits, which combined with other multi-trait analysis as seen in this chapter should yield further insights into the extra motor symptoms of ALS.

#### 2.4.2 - Differential genetic architecture of ALS in males and females

Our analysis also uncovered a number of interesting results regarding the differential architecture of ALS in males and females (aim (ii) ). In line with observations of differential risk and heritability in males and females, we observed that a significant portion of ALS SNP-based heritability is explained by interaction with sex. Scanning for variants significant in only one sex, and those with differential effects across sexes, we identified a number of novel and known loci which may have a sex-specific effect on ALS risk. Many of these sex-specific loci are highly expressed in the CNS, suggesting they have a feasible role in ALS pathology. We also found suggestive evidence that ALS may be more polygenic in females than males, supported by LD score regression, HESS analysis and



GREML analysis. In spite of these observations however, genetic correlation for ALS between males and females as estimated using bivariate GREML was high, suggesting that these differences between sexes may be the exception to the rule, and the majority of genetic effects are likely shared across sexes. These results provide evidence of the heterogeneity of ALS genetics across sexes, with potential implications for clinical trial design, however there are also a number of caveats that must be considered for the above analyses.

The interaction between sex and SNP-based heritability we observed mirrors recent work in pedigree studies showing increased heritability between mother and daughter pairings (Ryan et al. 2019) and suggests that the genetic component of ALS partially differs between males and females. As ALS is a complex disease, this gene by sex interaction could reflect a great number of influences from biological differences, to lifestyle and behavioural differences or differential environmental exposures related to these lifestyle and behavioural differences, all of which could modify the risk conferred by a given genetic variant. At base however a gene by sex interaction implies males and females have differential risk for developing ALS when exposed to a given set of genetic variants. This may boil down to different thresholds of the same genetic variants needed to develop the disease, or indeed distinct sets. To dissect this we scanned for variants that appear to a.) have a different degree of association with ALS in males and females or b.) have differential effects on males and females. We identified a number of novel and known variants with association signals in only males or only females, which may partially explain the differences in heritability seen. However, of these variants, only one also showed evidence of differential effect sizes in males and females (*PIP5K1B*), indicating that few of these variants have opposite effects on biology in males and females (i.e. protective vs risk). Instead variants likely confer risk in the same direction, but with different magnitudes of effect, or simply no effect in one sex. Genes proximal to sex specific loci were enriched for expression in the brain and several biological processes involving cell morphogenesis and growth, including neural and dendrite projection, which have feasible roles in ALS given that its defining feature is death of upper and lower motor neurones. However due to power issues, these variants are unlikely to fully describe the extent of genetic difference between males and females for ALS, motivating replication in a larger sample size.

The existence of variants with differential association with ALS in males and females may have important implications for the design of clinical trials and treatments, given the possibility of pharmacogenetic interactions disproportionately affecting each sex. For

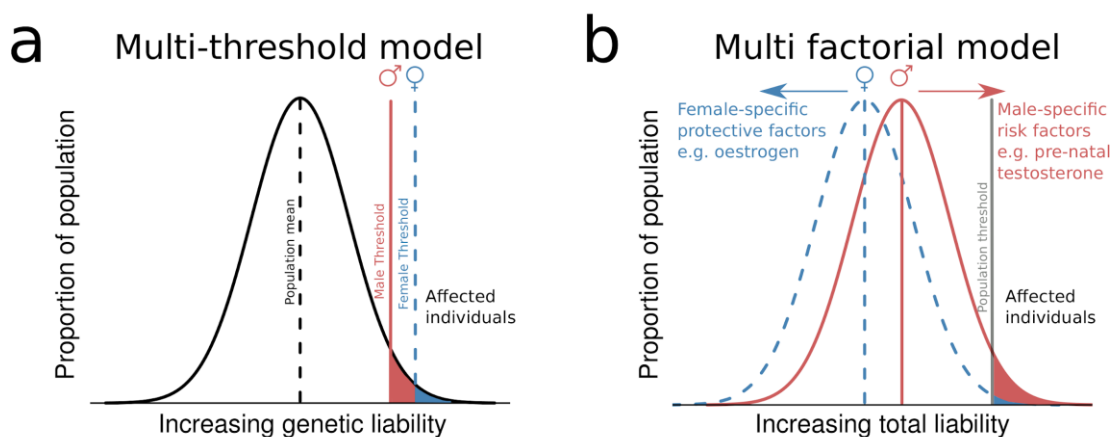
example, a recent meta-analysis of the effects of lithium carbonate on survival in ALS patients showed no response in the unstratified cohort, but significantly improved survival in carriers of the *UNC13A* C/C genotype (van Eijk et al. 2017), suggesting that lithium carbonate modifies survival specifically in carriers of this mutation. Given that *UNC13A* shows evidence of stronger association with ALS in females in our data it might be important to consider this putative sex imbalance both for future trials and treatments with this drug. Moreover loci that have a differential effect on risk in males and females such as *PIP5K1B*, and not simply a difference in association, may further confound clinical trials and treatments as variants in such loci could feasibly interact with treatment differently in males and females. This highlights the potential importance of these sex differential loci to finding suitable treatments for ALS on an individual basis and motivates further efforts to identify and replicate sex dimorphic loci.

To capture the full magnitude of sex differences globally we turned to methods combining the effects across the genome. Estimates of SNP-heritability were slightly lower in males in our data across three heritability methods (LDSC, HESS and GREML), consistent with the direction seen in pedigree studies, which also show lower heritability in male-male parent offspring pairs (Ryan et al. 2019). However rates of ALS are also puzzlingly higher in males than females (Johnston et al. 2006), meaning that while genetics appears to explain more of the variance in females, males are more likely to develop the disease. Together this suggests that affected females likely carry more genetic risk (hence the increased heritability among females), potentially due to a higher liability threshold for developing the disease. Alternatively male risk likely may have larger contributions from other non-genetic factors (e.g. environmental and lifestyle risks). If true, a higher non-genetic risk left unaccounted for might have the potential to act as a confounder when studying ALS in males, which may bias our male-only GWAS estimates. Consistent with this hypothesis we noted that a larger portion of the inflation in GWAS in males was attributable to confounding (LD score intercept and ratio). It is possible that some of this inflation is caused by the spread of non-genetic environmental risk factors in our male dataset tracking with population structure, which due to drift would lead to falsely associated variants and bias our sex-specific estimates (although this would not track with LD in the LDSC model). While we have attempted to correct for population structure here using standard methods such as principal components, the existence of subtle residual confounding in GWAS due to subtle population structure is pervasive (Sohail et al. 2019) and may affect this study, which is a motivation for work in future chapters (Chapters 3-5) characterising finescale population structure.

Heritability per chromosome in males does not scale with chromosome length (Figure 2.13), suggesting ALS in males may be explained by a much smaller set of variants than in females. Genetic signal for ALS in males also appears to be explained by a smaller fraction of the genome (HESS), suggesting the trait may be more polygenic in females. Combined with the higher heritability in females this could be indicative of a sex-dependent genetic liability for ALS which is higher for females (i.e. females require more risk alleles to develop ALS, Figure 2.14 a). This higher liability may be coupled with a multifactorial model whereby protective effects lower the mean total liability in females (Figure 2.14 b), requiring greater genetic burden in female ALS patients on average to overcome this protective effect and develop ALS. This model is consistent with the observed protective effect of oestrogen (Manjaly et al. 2010; de Jong et al. 2013; Rooney, Visser, et al. 2017) and risk effect of testosterone (Vivekananda et al. 2011) in ALS. A higher genetic burden needed in one sex resulting in lower rates in that sex and higher transmission from affected individuals of that sex is termed the “Carter effect” (Khrantsova, Davis, and Stranger 2019) in reference to its proposal by Cedric Carter as an explanation for the lower rates of pyloric stenosis in females, but strikingly higher chance of a woman with the disease passing it on to their offspring (Carter and Evans 1969). More recently, protective effects in females leading to a higher mutational burden and maternal transmission have been molecularly characterised for neurodevelopmental disorders (Jacquemont et al. 2014), lending credibility to the Carter effect. In further support of this model in ALS, children of female probands have been observed to have a higher relative risk of developing ALS than children of a male proband, suggesting a higher genetic load in females (Fang et al. 2009). However the same study of familial aggregation saw no difference in the relative risk of siblings from male and female probands, which would also be expected under this model. Should this model hold, it may have important implications for genetic counselling of ALS patients as it could further inform the relative risk of a given patient’s children or relatives developing the disease.

Despite significant evidence for sex-by-gene interaction effects, and lack of correlation of heritability estimates across chromosomes in males and females (linear regression:  $r^2=-0.049, p=0.99$ , Appendix Figure 2.1), when we estimated between-sex genetic correlation using the genome wide relatedness of samples in bivariate GREML, we observed a high correlation estimate with a confidence interval bounding one (0.75-1.25), denoting near perfect shared genetic variance in ALS between males and females. This estimate is considerably higher than the pedigree-based genetic correlation estimate of 0.628 (0.48-0.73). This high genetic correlation result seemingly conflicts our model of a sex-specific architecture in males and females, however the large standard error on this

estimate leaves room for the possibility that this correlation is not absolute. Combined with our other evidence of sex being a significant interaction term in ALS GWAS, it is likely that while sex-independent main effects make up the majority of the genetic component of ALS, the remaining proportion likely is composed of multiple small sex-dependent genetic effects. Similar statistically perfect correlations (interval overlapping one) between sexes in spite of evidence for widespread sex interactions have also been observed for numerous behavioural traits (J. Martin et al. 2020), meaning genetic correlation analysis may be currently underpowered to detect deviations from perfect genetic correlation between subgroups within a trait where main effects are large. Larger sample sizes and use of a wider set of variants (including rare variants) should enable us to construct more powerful GRMs allowing better characterisation of the true magnitude of genetic correlation between ALS across sexes.



**Figure 2.14: Illustration of sex differentiated liability threshold models of disease.**

a.) A multi threshold model for disease liability. Where genetic liability is normally distributed in the population, but the thresholds for genetic liability differ for males (lower) and females (higher) we would expect that heritability is higher for females as a greater genetic load is needed to cause the disease. Equally rates of the disease are expected to be higher in males under this model as a greater proportion are expected to pass the lower liability threshold. NB: This model may also be reversed for a higher liability in males.

b.) Multifactorial model: Multiple genetic and environmental factors etc contribute to total liability for males and females. Here multiple protective factors shift mean female liability away from the total liability threshold for the disease, leading to a lower rate of the disease in females, while risk factors shift males towards the disease liability threshold. In this case the heritability may be the same despite different rates of disease in males and females as other factors contribute to total liability.

Figure adapted from (Khramtsova, Davis, and Stranger 2019) for illustrative purposes (<https://www.nature.com/articles/s41576-018-0083-1#Sec7>)

Sex has recently been observed to appear “heritable” on the autosome (falsely) in large biobank scale GWAS due to sex-mediated participation bias in studies that had active participation (Pirastu et al. 2020). This could impact analysis of sex differences genetic architectures of traits in GWAS datasets with similar recruitment schemes by altering the baseline frequencies of alleles associated with this participation bias in males and females. Sex-biased participation is likely to have less of an impact on recruitment of ALS patients as they are passively recruited for GWAS in the clinic at the time of ALS diagnosis (provided they consent), meaning they do not have to actively seek participation in the study. Due to the fatal nature of the disease and lack of a cure patients are typically keen to be involved in research, leading to high rates of participation. However this phenomenon may bias participation of controls, who may in some cohorts have to actively seek recruitment to the study. Hence at the moment it is not wise to fully rule out the possibility that some associated variants observed in our sex-specific analysis are driven partially or fully by differential sex participation bias. Unfortunately, testing for replication of sex-specific variants in an independent replication cohort may not fully address this if control participation bias remains as associated variants appear to show some (though not complete) consistency across studies (Pirastu et al. 2020). At present it is not clear how to account for this bias as all control data features it to some degree, but perhaps development of population level passive population control recruitment schemes will alleviate this problem in future studies. Establishing ethical practices to organise these are not trivial, though the authors suggest some feasible methods, such as genotyping neonatal bloodspots and releasing only the population level allele frequencies (Pirastu et al. 2020).

## Chapter 3 - Finescale Irish Population Structure and Migration

*Now published in Byrne et al. Insular Celtic population structure and genomic footprints of migration. PLoS Genet. 2018;14: e1007152.*

### 3.1 - Introduction

#### 3.1.1 - Background

Situated on the northwestern frontier of Europe, the Island of Ireland has a population of approximately 6.4 million. The island is politically partitioned into the Republic of Ireland and Northern Ireland, with the latter forming part of the United Kingdom (UK) alongside the neighbouring island of Britain. Alternative divisions separate Ireland into four provinces reflecting early historical divisions: Ulster to the north, including Northern Ireland; Leinster (east); Munster (south) and Connacht (west). The island has been inhabited by humans for over 10,000 years (Bayliss and Woodman 2009), and there is genetic evidence of strong genetic continuity from Early Bronze Age Irish (~2200 BCE) and modern individuals (Cassidy et al. 2016). This continuity may in part be explained by the relative isolation of the island from mainland populations, preserving the genetic signature of the Irish people.

Despite this isolation, numerous settlements and invasions of Ireland from the neighbouring island of Britain and continental Europe have been recorded. This includes Norse-Vikings (9th-12th century), especially in east Leinster, and Anglo-Normans (12th-14th century), who invaded through Wexford in the southeast and established English rule mainly from an area later called the Pale in northeast Leinster (Duffy 2000). There has also been continuous movement of people from Britain, in particular during the 16-17th century Plantation periods during which Gaelic and Norman lands were systematically colonized by English and Scottish settlers. These events had a particularly enduring impact in Ulster in comparison with other planted regions such as Munster. As with the previous Norman invasion, the less fertile west of the country (Connacht) remained largely untouched during this period.

The genetic contributions of these migratory events cannot be considered mutually independent, given that they derive from either related Germanic populations (such as the Vikings and their purported Norman descendants) or from other Celtic populations inhabiting Britain, which had themselves been subjected to mass Germanic influx from Anglo-Saxon migrations and later Viking and Norman invasions (Leslie et al. 2015). Moreover, each movement of people originated from northern Europe, a region which had

witnessed a mass homogenizing of genetic variation during the migrations of the Early Bronze Age, possibly linked to Indo-European language spread (Haak et al. 2015; Lazaridis et al. 2016). However, each event had a geographic and temporal focal point on the island, which may be detectable in local population structure.

While initial studies of (unlinked) genome-wide autosomal variation in Ireland have shown that the island has lower genetic diversity than mainland Europe (Cronin et al. 2008; O'Dushlaine et al. 2010), and detected little evidence of population structure, numerous studies using alternative measures of genetic similarity suggest that population structure is pervasive in Ireland. The earliest evidence of genetic structure in Ireland comes from analysis of the frequency of ABO blood group and Rhesus factor markers, which show a notable cline across Ireland from east to west (Relethford 1983). The authors of this study proposed this may reflect waves of east to west migration both during the peopling of Ireland and in more recent years (i.e. Anglo-Norman invasion) (Relethford 1983), suggesting migration has led to population structure in Ireland. Moreover analysis of patrilineal markers including several Y-haplogroups (Hill, Jobling, and Bradley 2000; Moore et al. 2006; McEvoy, Simms, and Bradley 2008) and the distribution of Irish family surnames (Smith and Macrauld 2009) from the 19th century show clear geographic patterning in Ireland, suggesting that patrilineal ancestry is structured across the Irish population. Of note the haplogroup R1b3 reaches its European maximum in the west of Ireland (Hill, Jobling, and Bradley 2000), showing an east to west cline across the island similar to blood group frequencies. A sub haplotype of this group called the Irish Modal Haplotype (IMH) has been shown to be present almost exclusively in the northwest of the island, and was observed to be enriched in a cohort of individuals with surnames derived from the Uí Néill root (Moore et al. 2006), serving as a genetic signature of the dominance of the Uí Néill ruling clan in this region in the early medieval period. In addition long runs of homozygosity in Ireland correlate negatively with population density and diversity of grandparental origin (R. L. McLaughlin et al. 2015), suggesting that low ancestral mobility may play a role in preserving such regional genetic legacies in Ireland. Subsequent work on large autosomal genome-wide datasets described in this chapter (R. P. Byrne et al. 2018) (n=911) and in an independent parallel study (n=194) (Gilbert et al. 2017) have enhanced our understanding of the genetic structure in Ireland in a manner unlimited to patrilineal ancestry through the use of haplotype sharing methods in place of the unlinked methods above.

To investigate the potential of subtle population structure in Ireland, we applied the haplotype-based methods ChromoPainter and fineSTRUCTURE (Lawson et al. 2012),

which exploit rich haplotype sharing information across a sample population to partition it into distinct clusters of genetically similar individuals at a resolution not achievable by unlinked methods such as PCA or ADMIXTURE (Alexander, Novembre, and Lange 2009). This was motivated by previous work using these methods to reveal the hidden genetic structure in Britain (Leslie et al. 2015), which identified discrete genetic clusters of individuals that strongly segregated with geographical regions within Britain, even distinguishing the neighbouring regions of Cornwall and Devon. In doing so we identified subtle structure across the island which correlates strikingly with geography. We further explored patterns of haplotype sharing with both Britain and Europe to identify and date signals of admixture into Ireland in recent history using the GLOBETROTTER method (Hellenthal et al. 2014). The identified events correlated well with the historical record, revealing signals overlapping the Viking and Norman settlements across Ireland, and even the plantations in a restricted geographic range.

### 3.1.2 - Research aims

The chapter presents a study characterising Irish population structure using an Irish ALS case-control dataset and has been published in PLoS Genetics

(<https://doi.org/10.1371/journal.pgen.1007152>). This research had three major aims:

- i.) To identify whether subtle population structure is present in Ireland in spite of reported homogeneity;
- ii.) To characterise the extent and geographic patterning of this population structure;
- iii.) To contextualise this structure in terms of the historical record and in relation to recent migrations from neighbouring populations, namely Britain and mainland Europe.

While these questions have historical and cultural importance, the potential existence of systematic genetic differences in a supposedly isolated and homogeneous population also has relevance to medical genetics, in particular association studies which aim to identify disease related variants and must correct for confounding due to non-disease related variance in the data. As stated in the discussion of chapter 2, properly characterising subtle population structure may be of critical importance for disentangling correlated non-genetic effects mediating disease. The goal of exploring the effects local population structure has on medical genetics is an ongoing project which will be explored further in a later chapter (Chapter 5), however our preliminary results in Ireland highlight an important consideration with regards the design of sound association studies.



## 3.2 - Methods

### 3.2.1 - Datasets

Analyses in this chapter were carried out using three population level genotype datasets sampled from Ireland (n=991; EGA accession ID EGAS00001002769) (van Rheenen et al. 2016; R. P. Byrne et al. 2018), The UK (n=2,020; EGA accession ID EGAD00010000632) (Leslie et al. 2015) and across Europe (n=4514; EGA accession ID EGAD00000000120) (Sawcer et al. 2011), and merges of these datasets. These datasets are described in detail below:

#### 1.) Irish dataset:

Our Irish dataset is a population based ALS case control dataset composed of newly genotyped (R. P. Byrne et al. 2018) and previously published Irish samples (van Rheenen et al. 2016). The newly genotyped samples consist of 407 Irish samples (271 cases; 136 controls) from the Irish ALS DNA bank which were genotyped at 2.5 million single nucleotide polymorphisms (SNPs) using the Infinium HumanOmni2.5-8 SNP array v1.2 as part of the Project MinE Sequencing Consortium effort (van Rheenen et al. 2018). This dataset was merged with published Irish data from an ALS GWAS dataset (van Rheenen et al. 2016) containing 713 individuals (265 cases; 448 controls) genotyped on the Infinium OmniExpress-24 kit. The final merge of these datasets following quality control (QC) contained a total of 991 Irish individuals genotyped at 407,750 SNPs.

#### 2.) UK/British dataset (PoBI dataset):

The UK dataset from the People of the British Isles (PoBI) project (Leslie et al. 2015) was downloaded from the EGA (EGA accession ID EGAD00010000632). We retained only samples with a recorded geographic origin (n=2,039) prior to QC. Following QC 2,020 individuals, genotyped at a total of 521,833 SNP sites remained. These were included in the initial UK-only fineSTRUCTURE run to define homogeneous clusters for use as surrogate/donor populations in GLOBETROTTER.

#### 3.) European dataset:

We subsetting European individuals (n=6,670) from a Multiple Sclerosis (MS) GWAS dataset (Sawcer et al. 2011) for use as a reference dataset. Our initial QC reduced the European dataset to 4,737 individuals genotyped at 363,396 SNPs. A further 223 individuals were removed prior to GLOBETROTTER analysis recommended by the

WTCCC quality control left 4,514 individuals.

#### 4.) UK-Irish Merge:

Following the merging of Irish and UK datasets and an additional round of QC, three additional UK individuals were removed. The final UK-Irish merge contained 3,008 individuals and 214,632 SNPs. This dataset was used for the Ireland/Britain fineSTRUCTURE analysis (Figure 3.2) and the GLOBETROTTER analysis of British admixture into Ireland. We also used a subset of this dataset containing Irish samples (n=991) and individuals from Northern Ireland from the UK dataset (n=44) amounting to a complete dataset of 1,035 individuals from the island of Ireland genotyped at 214,632 SNPs (Figure 3.1) .

#### 5.) European-Irish merge:

Following merging and additional QC our European-Irish merge contained 5,506 individuals at 166,139 SNPs. This dataset was used for the GLOBETROTTER analysis of European admixture into Ireland.

### 3.2.2 - Quality control

We applied the following quality control to each single population and merged multi-population dataset using PLINK 1.9 (Chang et al. 2015): First uncommon SNPs (--maf 0.05) and those with missingness greater than 2% across samples (--geno 0.02) were filtered. Individuals with high missingness (--mind 0.1) or heterozygosity (--het; exceeding three median absolute deviations from the median) were then removed. Related individuals were next identified using an identity-by-descent matrix (--genome) and one individual from each pair exceeding 12.5% relatedness was removed from downstream analysis (retaining the sample with lower missingness). Principal component outliers were removed from the Irish population dataset (>4 standard deviations from the mean for principal components 1-2 for the Irish and European/Irish merge). Finally individuals who had been removed during the QC carried out in the source papers for the datasets (Sawcer et al. 2011; Leslie et al. 2015) were removed.

As the European dataset included patients from a GWAS for MS, we additionally removed SNPs in a 15 Mb region surrounding the strongly associated HLA locus on chromosome 6 (GRCh37 position chr6:22,915,594–37,945,593) for all datasets including the European dataset, to prevent haplotypic bias arising from this locus.

Finally, prior to ChromoPainter analysis all SNPs with missingness in each single

population and merged dataset were removed to prevent bias arising from differentially missing data (--geno 0). We noted the SNP missingness filter (--geno 0) led to a significant loss of SNPs in the European dataset, particularly after merging with the Irish due to poor overlap of SNP panels with the Irish dataset. Given that the dataset was solely intended for use in the merge with the Irish dataset, we resolved to balance this SNP loss by removing individuals from the European dataset above a threshold of missingness prior to applying the --geno 0 filter. We assayed a range of thresholds to maximise SNP count while minimising individual loss using the --mind command in PLINK. We settled on the threshold --mind 0.0005 which maintained sufficient SNPs and individuals for a meaningful analysis.

The post-QC Irish (n=991), British (n=2,020) and European (n=4,514) datasets retained 407,750 SNPs, 521,883 SNPs and 363,396 SNPs at zero missingness, respectively. The merged British and Irish dataset (n=3,008) and the merged European and Irish dataset (n=5,506) retained 214,632 and 166,139 SNPs at zero missingness respectively.

### 3.2.3 - Phasing

We phased the autosomal chromosomes of each single population dataset and merged dataset using SHAPEIT V2 (Delaneau, Marchini, and Zagury 2011) with the 1000 Genomes (Auton et al. 2015)(Phase 3) panel as a reference. A pre-phasing step was carried out (--check) to remove any SNPs missing in the reference panel or with alleles incompatible with the reference panel. Samples were split by chromosome and phased simultaneously, using the GRCh37 genetic map to estimate recombination rate.

### 3.2.4 - fineSTRUCTURE analysis

We ran the ChromoPainter/fineSTRUCTURE (Lawson et al. 2012) pipeline on each phased population dataset individually (Irish, British and pan-European), as well as on the merge of the Irish and British datasets to detect subtle population structure and partition the data into homogeneous clusters. We first ran ChromoPainter using the built in version in fineSTRUCTURE (fs-2.0.8) painting each individual with all others as donors (-a 0 0). We ran 10 expectation maximisation iterations to estimate  $N_e$  and  $\mu$  (switch rate and mutation rate) for each chromosome (for the analysis of the combined Irish and NIR dataset, which contained data from Northern Ireland and Ireland, we used the weighted average for  $N_e$  and  $\mu$  across chromosomes 1,8,15,20), and used these parameters for the final painting. Default settings were used for all other parameters with the exception of the “chunks” per region switch which was set to 50 (-k 50) for analyses including British or Irish individuals to account for the long haplotypes observed in these populations.

Individual paintings were then combined into a single “coancestry” matrix using the chromocombine function in fineSTRUCTURE.

For the Irish and British datasets we ran the fineSTRUCTURE MCMC model on each “coancestry” matrix using 2,000,000 burn-in and 2,000,000 sampling iterations, sampling every 10,000 iterations. For reasons of computational tractability, for the much larger European dataset we ran 1,000,000 burn-in and 1,000,000 sampling iterations sampling every 1,000 iterations. fineSTRUCTURE samples the possible partitions of the “coancestry” matrix and estimates the posterior probability of each partition to search for population assignments where all individuals in a group share haplotypes consistently within their group and with other groups. Two MCMC chains were run to assess convergence of the final cluster membership. We extracted the state with the maximum posterior probability and ran an additional 10,000 hillclimbing iterations before building trees. Trees were built with both the default climbtree method and the maximum concordance method described in the PoBI study (Leslie et al. 2015). For all GLOBETROTTER analyses the maximum concordance method was used to define donor populations as it provides more confident cluster assignments.

fineSTRUCTURE trees were visualised using the R scripts provided at <http://www.paintmychromosomes.com>. Cluster names were devised based on the geographic spread of a given cluster.

### 3.2.5 - Cluster robustness

We evaluated the robustness of the Irish clusters using the total variation distance (TVD) metric defined in the PoBI study (Leslie et al. 2015). This metric compares the “copying vectors” of pairs of clusters where the copying vector for a given cluster A is a vector of the average length (cM) of DNA donated by each cluster to individuals in cluster A, as estimated by ChromoPainter. The magnitude of the sum of absolute differences between the copying vectors (or TVD) of a pair of clusters is thus a measure of their distance in terms of haplotypic sharing with other clusters, which directly relates to ancestry.

We used permutation testing to interrogate whether the fineSTRUCTURE clustering performed better than chance by permuting the individuals in each of our cluster pairings into pseudo clusters of the same size 1,000 times, and calculating the number of permutations exceeding our original TVD score. Where 1,000 permutations were not possible due to cluster size, the maximum number of unique permutations was used. P-values were calculated based on the proportion of permutations greater than or equal to our original clustering. All p-values for Irish clusters were less than or equal to 0.001.

We also calculated mean genome wide  $F_{ST}$  between cluster groups to demonstrate the degree of genetic differentiation between clusters with a non-haplotype based metric. Variants were first pruned using PLINK 1.9 (Chang et al. 2015) (--indep-pairwise 1000 50 0.25) and  $F_{ST}$  was calculated (--fst) between pairs of clusters groups.

### 3.2.6 - Estimating admixture dates

We used the GLOBETROTTER (Hellenthal et al. 2014) method to infer admixture events from Europe and Britain into Ireland separately. GLOBETROTTER uses ChromoPainter output to fit a mixture model for a target population's haplotypic makeup as a combination of modern donor populations, which can be thought of as a proxy for the true ancestral groups. The better the ancestral groups are represented the cleaner the model fit will be. GLOBETROTTER then estimates the pairwise likelihood of being painted by two donor populations at a variety of genetic distances to generate coancestry curves. Under a simple single admixture event model the rate of decay of these exponential curves should equal the number of generations since the event.

We performed two separate GLOBETROTTER analyses modelling admixture events from 1.) British and 2.) European sources separately. First we defined homogenous subgroups in the British and European populations using the Maximum concordance trees generated for Europe and Britain above. Note that for the British analysis PoBI samples from Northern Ireland were excluded as donors from their clusters as they might confound the model by being overly similar to the target group. Using these populations as donor groups for each analysis, we painted target Irish individuals, and each respective donor population with ChromoPainter v2 to generate a copying matrix (chunk lengths). This matrix represents the average length of genetic material received from each donor population by each Irish individual and each individual in the donor populations. We also generated 10 painting samples for each Irish individual under the same model. We ran GLOBETROTTER twice for 5 mixing iterations: first using the null.ind:1 setting to test for evidence of admixture, and then the null.ind:0 setting to infer dates and sources. We generated 100 bootstraps for the admixture date and calculated the probability of a null model of no admixture as the proportion of nonsensical dates (<1 or >400 generations) produced by the null.ind:1 model, as in the original paper (Hellenthal et al. 2014). Confidence intervals for the date were calculated from the bootstraps of the null.ind:0 model using the empirical bootstrap method. For all conversions of generation to year we assumed a generation time of 28 years as was done in previous studies. (Hellenthal et al. 2014; Leslie et al. 2015).

### 3.2.7 - Ancestry proportion estimation

We modelled each Irish cluster's makeup in terms of European and British sources using GLOBETROTTER (Hellenthal et al. 2014) in order to investigate if there was ancestral variation across Ireland. GLOBETROTTER uses the ChromoPainter "Chunklength" output to model the average genome of each cluster as a linear mixture of the donor populations, and estimates the proportion of DNA which most closely matches individuals from each of the sources. The model also takes sharing between donor populations into account to correct for noise caused by population splits that may have occurred after the coalescence of these donor populations. A multiple linear regression is fitted of the form:

$$Y_p = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_g X_g, \quad (5)$$

where  $Y_p$  is a vector of the average length in centimorgan of DNA that individuals of cluster  $p$  copy from each donor group ( $g$ ) as a proportion of the total genome, and  $X_g$  is the vector describing the average proportion of DNA genome wide that a donor group  $G$  copies from each donor group including their own. This regression is solved for each  $\beta_g$  using a non-negative-least squares function implemented in GLOBETROTTER. Each coefficient  $B_g$  thus represents the average proportion genome-wide for which individual from cluster  $p$  is most closely related to donor group  $g$ . We ran GLOBETROTTER using `num.mixing.iterations:0` to carry out this regression.

### 3.2.8 - ADMIXTURE analysis

We ran ADMIXTURE (Alexander, Novembre, and Lange 2009) on the combined British and Irish datasets, alongside eighteen ancient individuals from the Iron age, Roman and Anglo-Saxon periods (Martiniano et al. 2016; Schiffels et al. 2016) to assess differential British admixture in Irish clusters. Pseudo-haploid genotypes were generated for the ancient genomes at the variant sites called in the British and Irish, as is standard for low coverage data. Data were merged and pruned for linkage disequilibrium using PLINK 1.9 ( $r^2 > 0.25$  in a sliding window of 1000 SNPs advancing 50 SNPs at a time). No missingness was allowed in modern individuals while ancient individuals ranged from 33,643-85,553 sites out of 86,481 sites. Cross validation error was used to determine the  $k$  value for which the model has the best predictive accuracy (`--cv`), for a default of five iterations. Additionally 200 bootstraps were run to estimate the standard error of parameters (`--B`). The British/Anglo-Saxon ADMIXTURE component was regressed against ChromoPainter

PC2 to determine the role of British ancestry in differentiation in Ireland. ANOVA was also applied to identify between cluster differences of this ancestry component.

### 3.2.9 - PCA and t-SNE analysis

We projected the ChromoPainter coancestry matrix in lower-dimensional space using both principal components analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE). PCA was run using the R tools provided on the fineSTRUCTURE page (<http://www.paintmychromosomes.com>), while t-SNE was generated using the Rtsne R package (<https://github.com/jkrijthe/Rtsne>). We ran t-SNE on the coancestry matrix for 5,000 iterations, using a perplexity of 30, a learning rate of 200 and 100 PCA dimensions as an input.

### 3.2.10 - Mapping samples

Of the 991 Irish samples in this dataset after QC, geographic information was available for 544 in the form of home address. For the purposes of preserving anonymity this was jittered in all maps containing patients (Figures 3.1 and 3.8). Additionally, 44 PoBI individuals from Northern Ireland were used, but precise sampling location was not available so these are plotted as a circle in Figure 3.1. The maps were generated using Global Administrative areas from GADM version 2.8 (November 2015; <http://www.gadm.org>).

For UK data sample location was described in terms of membership of 35 sampling regions in supplement to the PoBI data (Leslie et al. 2015). To plot these regions in Figure 3.2 and Figure 3.6 we used the UK map and administrative boundary data from GADM (<http://www.gadm.org>) to approximate regions defined in NUTS 2010 (Commission and Others 2011) (Nomenclature of Territorial Units for Statistics). We then combined sets of these NUTS 2010 regions to best approximate the 35 sampling regions. The 35 sampling regions were then allotted to the cluster group containing the majority of samples from the respective sampling region and labeled accordingly. Where a cluster or cluster group had the majority of samples from multiple adjacent sampling regions (as in the case of southeast England), we subsumed these regions into one for visualisation. For consistency with the UK map we divided Ireland into regions using the NUTS 2010 definitions (Commission and Others 2011), with each region assigned to the cluster with the majority of samples in that region as above (Figures 3.2 and 3.6).

### 3.2.11 - Statistical analyses

All ANOVA, linear regression and other statistical tests were carried out in R version 3.2.3 (CoreTeam 2015)



### 3.3 - Results

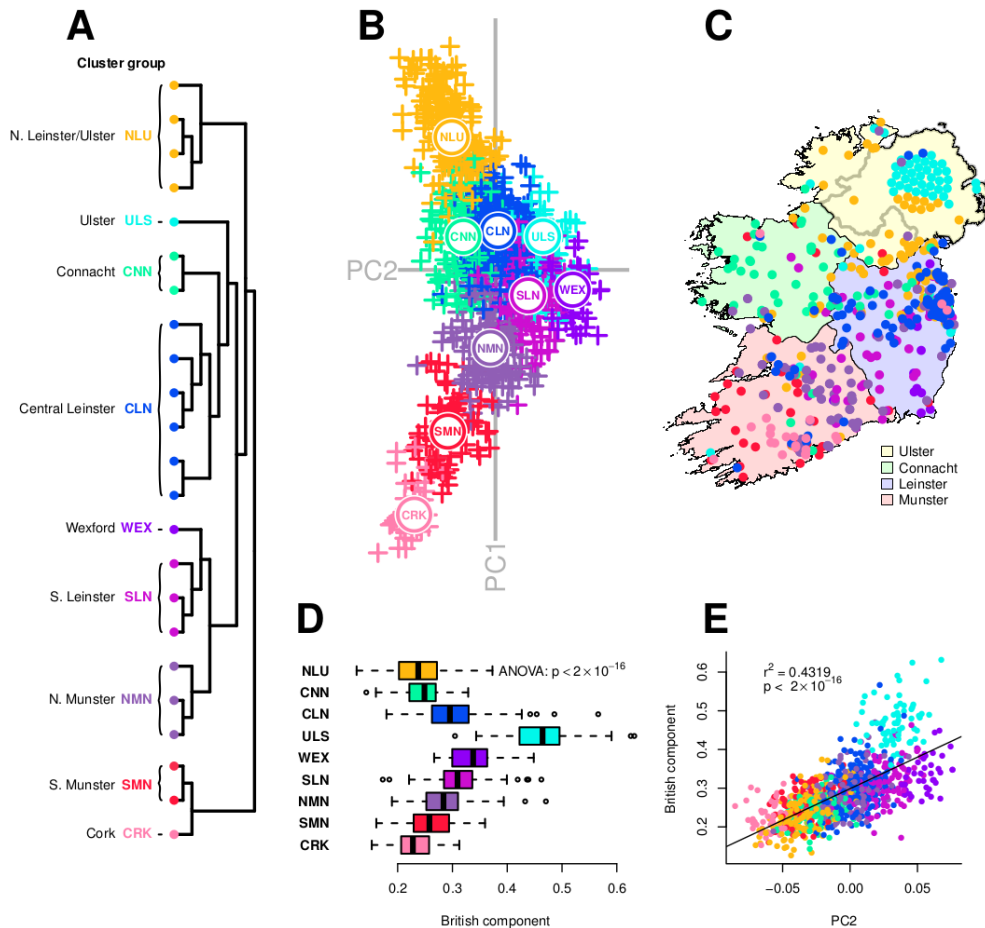
#### 3.3.1 - Finescale population structure in Ireland

We painted the phased genome of each individual in our combined Irish dataset (n=1,035 individuals, including 44 Irish samples from PoBI (Leslie et al. 2015)) in terms of all remaining individuals using the ChromoPainter algorithm (Lawson et al. 2012) to generate a pairwise coancestry matrix of haplotypic sharing between samples. This coancestry matrix summarises genome wide ancestral “chunk” sharing between unrelated individuals, and thus can be used to determine subtle relationships between individuals not captured by independent marker summaries, which discard linkage information. To explore the structure in this matrix we first clustered it into homogeneous groups of individuals using fineSTRUCTURE (Lawson et al. 2012), a software which samples possible partitions of the coancestry matrix using a Markov Chain Monte Carlo (MCMC) algorithm with the goal of finding a split where all individuals within a cluster share similar numbers of “chunks” with individuals within the cluster, and with individuals from other inferred clusters. Upon convergence of the MCMC chains fineSTRUCTURE identified 23 genetic clusters, forming 9 cluster groups (Figure 3.1 A; cluster groups described below), which were shown to be robustly defined by TVD analysis ( $p < 0.001$ ) and demonstrated clear but subtle differences using the unlinked  $F_{ST}$  statistic (Tables 3.1 and 3.2).

**Table 3.1: Mean pairwise  $F_{ST}$  between Irish cluster groups.**

$F_{ST}$	NLU	SMN	NMN	CLN	CNN	SLN	WEX	CRK	ULS
NLU	0	4.63E-04	3.36E-04	1.32E-04	2.01E-04	3.06E-04	5.27E-04	7.71E-04	4.74E-04
SMN	-	0	1.30E-04	2.56E-04	3.39E-04	2.46E-04	4.94E-04	1.49E-04	5.07E-04
NMN	-	-	0	1.19E-04	1.97E-04	7.84E-05	2.84E-04	4.92E-04	3.14E-04
CLN	-	-	-	0	1.20E-04	2.02E-05	1.90E-04	5.79E-04	2.47E-04
CNN	-	-	-	-	0	2.44E-04	5.05E-04	6.41E-04	4.80E-04
SLN	-	-	-	-	-	0	1.98E-04	6.07E-04	2.76E-04
WEX	-	-	-	-	-	-	0	8.41E-04	3.69E-04
CRK	-	-	-	-	-	-	-	0	8.97E-04
ULS	-	-	-	-	-	-	-	-	0

Average Wright’s  $F_{ST}$  between pairs of cluster groups (Figure 3.1) calculated using PLINK 1.9 (Chang et al. 2015). Broad patterns of  $F_{ST}$  between groups match the major differences described by the fineSTRUCTURE tree.



**Figure 3.1: Fine-grained population structure in Ireland.**

(A) fineSTRUCTURE clustering dendrogram for 1,035 Irish individuals. Twenty-three clusters are defined, which are combined into cluster groups for clusters that are neighbouring in the dendrogram, overlapping in principal component space (B) and sampled from regions that are geographically contiguous. Details for each cluster in the dendrogram are provided in Appendix Figure 3.1 (B) Principal components analysis (PCA) of haplotypic similarity, based on ChromoPainter coancestry matrix for Irish individuals. Points are coloured according to cluster groups defined in (A); the median location of each cluster group is labeled with the cluster name. (C) Map of Ireland showing the sampling location for a subset of 588 individuals analysed in (A) and (B), coloured by cluster group. Points have been randomly jittered within a radius of 5 km to preserve anonymity. Precise sampling location for 44 Northern Irish individuals from the People of the British Isles dataset was unknown; these individuals are plotted geometrically in a circle. The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>). (D) “British admixture component” (ADMIXTURE estimates;  $k = 2$ ) for Irish cluster groups. This component has the largest contribution in ancient Anglo-Saxons and the SEE cluster. (E) Linear regression of principal component 2 (B) versus British admixture component ( $r^2 = 0.43$ ;  $p < 2 \times 10^{-16}$ ). Points are coloured by cluster group.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

**Table 3.2: Total Variation Distance between Irish cluster groups.**

TVD	NLU	SMN	NMN	CLN	CNN	SLN	WEX	CRK	ULS
NLU	0	1.05E-01	8.02E-02	4.36E-02	5.40E-02	7.11E-02	8.57E-02	1.34E-01	5.94E-02
SMN	-	0	4.17E-02	7.78E-02	7.80E-02	6.58E-02	8.60E-02	3.99E-02	7.60E-02
NMN	-	-	0	4.30E-02	5.47E-02	3.39E-02	6.05E-02	8.16E-02	4.81E-02
CLN	-	-	-	0	3.16E-02	3.15E-02	4.83E-02	1.15E-01	2.33E-02
CNN	-	-	-	-	0	5.15E-02	7.16E-02	1.11E-01	3.93E-02
SLN	-	-	-	-	-	0	2.90E-02	1.03E-01	3.64E-02
WEX	-	-	-	-	-	-	0	1.17E-01	5.58E-02
CRK	-	-	-	-	-	-	-	0	1.13E-01
ULS	-	-	-	-	-	-	-	-	0

TVD measures the distance between clusters based on haplotypic sharing profiles (See methods). TVD between cluster groups (Figure 3.1) captures much of the structure seen in the fineSTRUCTURE tree. All pairwise TVD values between clusters are significant at  $p < 0.001$  based on permutation testing.

To visualise the geographic patterning of our data we plotted individuals for whom we had geolocations ( $n=588$ ) onto a map of Ireland and coloured points by clusters assigned by fineSTRUCTURE (Figure 3.1 C). fineSTRUCTURE clusters demonstrated a clear relationship with geography, broadly segregating by province (Ulster, Leinster, Munster and Connacht), and even showing subtle within-province divisions. We noted that neighbouring clusters on the hierarchical fineSTRUCTURE tree localised to similar geographic regions displaying an isolation by distance-like effect (Figure 3.1). Deep cuts of the fineSTRUCTURE tree, which subsume closely related clusters, split samples approximately by province suggesting that genetic differences among individuals are deepest across provinces, while shallower splits exhibit clusters localised to neighbouring sets of counties, revealing structure even within Irish provinces.

To supplement our clustering, and provide a more continuous picture of the variation in our coancestry matrix we performed principal component analysis (PCA), and projected the data onto the axes explaining the greatest components of variation in the data. The

first two ChromoPainter principal components (cp-PCs) separated the clusters well and captured many of the trends described by the hierarchical clustering (i.e. clusters which were proximal/distant on the fineSTRUCTURE tree were proximal/distant in cp-PC space). Projection into cp-PC space also enabled us to test the geographic correlation of our genetic structure more explicitly by providing genetic coordinates for each sample which could be regressed against latitude and longitude. Regression analysis demonstrated a strong relationship between cp-PC1 and latitude ( $r^2 = 0.405$ ;  $p < 2 \times 10^{-16}$ ), and cp-PC2 with Longitude ( $r^2 = 0.126$ ;  $p = 5.9 \times 10^{-16}$ ) describing north-south and east-west genetic gradients in Ireland. When compared to PCA of independent markers calculated using a genetic relationship matrix (GRM) in GCTA (Yang et al. 2011), which is commonly used to correct population structure in GWAS, cp-PCs showed much stronger relationship to geography, with the first four components of the coancestry matrix explaining over 40% of the variance in Latitude and 33% of the variation in Longitude (Table 3.3). These observations suggest that the population structure we observe in Ireland is strongly linked to geography, following previous observations that genes mirror geography across Europe (Novembre et al. 2008) and more recent observations that local population structure in Britain segregates geographically (Leslie et al. 2015).

As some clusters contained very small numbers of individuals, overlapped with other clusters both in cp-PC space and geographically, and were proximal on the fineSTRUCTURE tree, we decided to subsume our 23 clusters into 9 cluster groups for further analysis. Considering the relationships between these super-clusters in more depth revealed some interesting trends. Firstly the North Leinster cluster group showed greater affinity (in cp-PC space) to the Connacht and Ulster cluster groups than to the South Leinster cluster group, which in turn appeared more related to the North Munster cluster group demonstrating a deviation from the relationships we would expect given modern political boundaries. This relatively deep split between North and South Leinster could, however, be explained in terms of pre-Norman territorial boundaries which divided Ireland into fifths (cuige), with North Leinster belonging to the kingdom of Meath (mide) (Duffy 2012). While this interpretation is open to debate, it is nonetheless interesting, regardless of cause, that such subtle differences and nuanced structure are detectable within the Irish population despite its relative homogeneity and isolation from mainland Europe.

**Table 3.3: Prediction of longitude and latitude using GCTA GRM and ChromoPainter.**

Predicted Variable	Formula	GCTA GRM		ChromoPainter Coancestry		Comparison
		p (F-test)	Adjusted R-squared	p (F-test)	Adjusted R-squared	R-squared diff# (CP - GCTA)
Latitude	PC1	0.93	0.00	<b>2.00E-16</b>	0.40	<b>0.41</b>
Latitude	PC2	0.12	0.00	0.99	0.00	0.00
Latitude	PC1 + PC2 + PC1*PC2	0.40	0.00	<b>2.20E-16</b>	0.42	<b>0.42</b>
Latitude	PC1 + PC2 + PC3 + PC4	0.60	0.00	<b>2.00E-16</b>	0.40	<b>0.41</b>
Latitude	PC1 + ... + PC20	<b>2.20E-16</b>	0.25	<b>2.20E-16</b>	0.38	<b>0.13</b>
Longitude	PC1	0.99	0.00	<b>1.99E-14</b>	0.14	<b>0.14</b>
Longitude	PC2	0.47	0.00	<b>5.90E-13</b>	0.12	<b>0.12</b>
Longitude	PC1 + PC2 + PC1*PC2	0.78	0.00	<b>2.00E-16</b>	0.27	<b>0.27</b>
Longitude	PC1 + PC2 + PC3 + PC4	0.94	-0.01	<b>2.00E-16</b>	0.34	<b>0.35</b>
Longitude	PC1 + ... + PC20	<b>1.80E-05</b>	0.10	<b>2.00E-16</b>	0.33	<b>0.23</b>

Linear regression of latitude and longitude of 544 geocoded Irish samples on principal components calculated from a GCTA GRM and the ChromoPainter coancestry matrix demonstrate that the ChromoPainter coancestry matrix explains a greater proportion of geographic variance. Here “R-squared diff” represents the difference in R-squared between the ChromoPainter PCs and the GCTA GRM PCs.

Another notable trend is the relatively high degree of differentiation between west Irish clusters along cp-PC1, compared to the more homogeneous eastern clusters. It is possible that this strong distinction between the western clusters may be due to the decreased impact of migration in these regions, allowing more ancient splits to be preserved. Indeed the South Munster cluster branches off from the fineSTRUCTURE tree first, demonstrating relatively strong differentiation from the neighbouring North Munster clusters. TVD analysis further supports that the Cork and South Munster clusters are the most distant from all other clusters (Table 3.2). It is possible that the isolating effects of mountain ranges surrounding Cork and Kerry may have restricted gene flow with the rest of Ireland, leading to the preservation of more ancient structure in Cork.

Eastern Ireland shows relative homogeneity, demonstrating large clusters of individuals which cannot be easily distinguished from each other by fineSTRUCTURE or cp-PCA. Notably the largest of the Central Leinster (CLN) clusters comprises about a fifth of our dataset. This homogeneity on the east coast was also observed when applying fineSTRUCTURE to England in the PoBI study (Leslie et al. 2015), where a single cluster in the southeast of England comprised almost half the dataset. Such homogeneity could suggest better mobility on the east coast of Ireland, leading to more random mixing and less structure, but may also be an imprint of migration from the neighbouring island of Britain erasing ancient structure. To explore this hypothesis we estimated the degree of similarity between each Irish individual across our dataset with samples from modern Britain from the PoBI dataset, and 18 ancient individuals from the Iron age, Roman and Anglo-Saxon periods in northeast and southeast England (Leslie et al. 2015; Martiniano et al. 2016; Schiffels et al. 2016). We performed an unsupervised ADMIXTURE analysis ( $k=2$ ) and identified a component that comprises the totality of several Anglo-Saxon individuals and forms the largest component in British samples, with lower but varying contribution to Irish samples (see section 3.3.2 below). This component showed significant variation across Irish fineSTRUCTURE clusters (ANOVA  $p < 2 \times 10^{-16}$ ), with lowest values in west coast clusters, and highest in the east coast clusters, lending support to the theory of differential impact of migrations from the Britain (Figure 3.1 D). Regression of PC2 of the coancestry matrix on this component was also highly significant, explaining 43% of the variance ( $p < 2 \times 10^{-16}$ ) suggesting that it may reflect an Anglo-Gaelic cline (Figure 3.1 E). Notably the Ulster cluster group harboured the greatest proportion of the Anglo-Saxon component, which may be explained by the strong impact of the Ulster Plantation in the 17th century on the genetic makeup of Ulster.

### 3.3.2 - The genetic structure of Ireland in the context of Britain

We investigated the relationship between Britain and Ireland using a pooled dataset containing samples from PoBI and our Irish samples ( $n=3,008$ ). Again we painted each individual in terms of all remaining individuals and generated a co-ancestry matrix representing pairwise haplotype sharing. Clustering with fineSTRUCTURE revealed 50 clusters which segregated geographically both on a cohort wide and local level. The deepest split in the fineSTRUCTURE tree ( $k=2$ ) separated most Irish and British data (Figure 3.2, with notable exceptions discussed below), while successive shallower splits showed more local regional clustering of the data. We noted our deepest split ( $k=50$ ) had a number of small clusters ( $n<10$ ), which may represent under sampled sub-populations, but are difficult to interpret meaningfully in the absence of more data, hence we decided to subsume them into their closest large cluster. However simply taking uniform tree cut to achieve this discarded subtle but meaningful splits such as the split between North and South Wales, where we had adequate sample size to distinguish with confidence between the clusters. Hence we decided to group clusters into cluster groups heuristically based on their proximity on the clustering dendrogram, geographic proximity and projection in cp-PC space to ease interpretation.

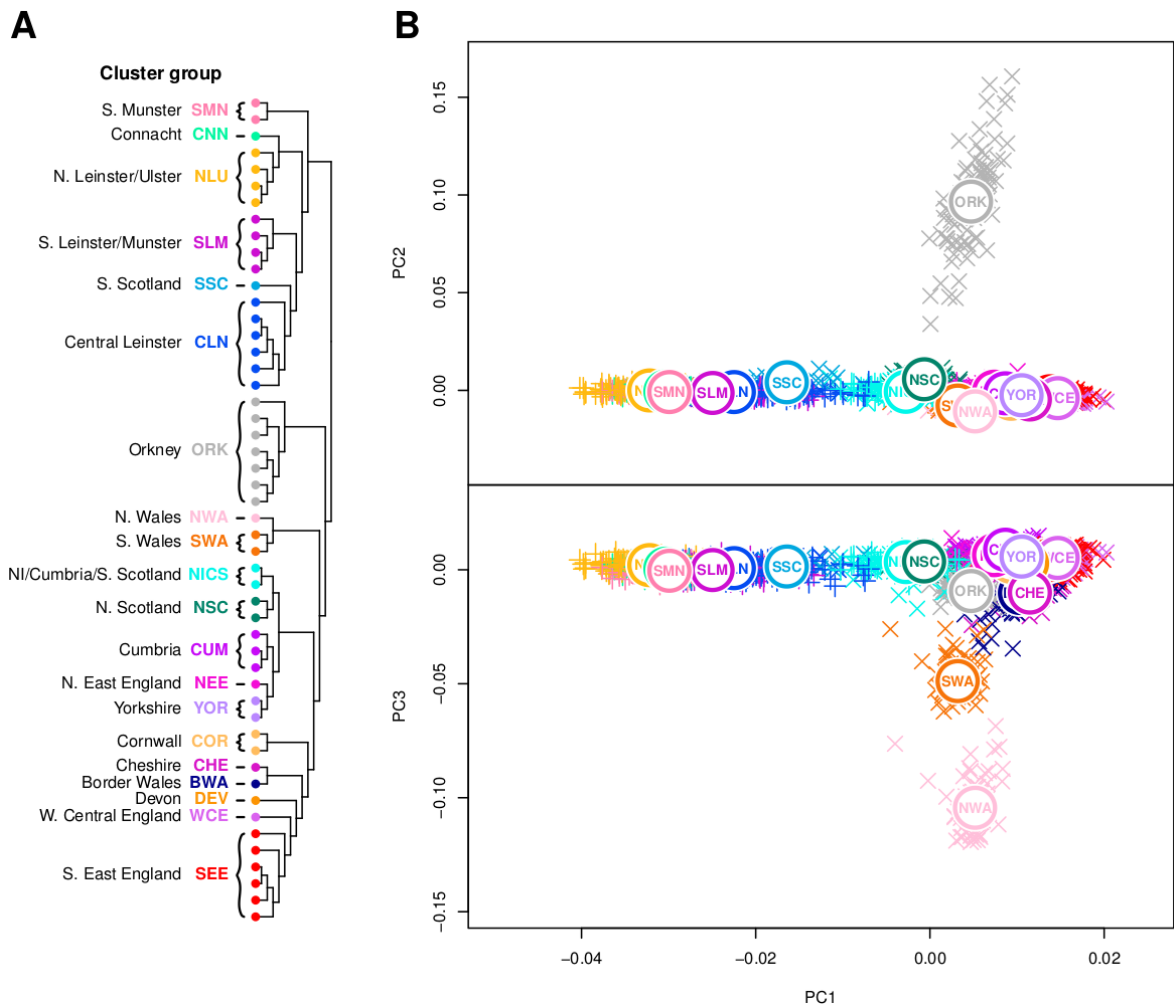
Projection of data into cp-PC space as above showed the principal split in the data was between Britain and Ireland (cp-PC1), (Figure 3.2) suggesting that the largest differences exist between Irish and British groups. Subsequently cp-PCs 2 and 3 separated out Orkney and Wales respectively (Figure 3.3) as expected from previous studies (Leslie et al. 2015); while cp-PC4 separated northern and southern regions in both Ireland and Britain. The projection of cp-PC1 vs cp-PC4 captures much of the geographic spread of samples, and clusters, suggesting that genes mirror geography across Britain and Ireland (Figure 3.2).



**Figure 3.2: Genes mirror geography in the British Isles.**

(A) fineSTRUCTURE clustering dendrogram for combined Irish and British data. Data principally split into Irish and British groups before subdividing into a total of 50 distinct clusters, which are combined into cluster groups for clusters that formed clades in the dendrogram, overlapped in principal component space (B) and were sampled from regions that are geographically contiguous. Names and labels follow the geographical provenance for the majority of data within the cluster group. Details for each cluster in the dendrogram are provided in Appendix Figure 3.3. (B) Principal component analysis (PCA) of haplotypic similarity based on the ChromoPainter coancestry matrix, coloured by cluster group with their median locations labelled. We have chosen to present PC1 versus PC4 here as these components capture new information regarding correlation between haplotypic variation across Britain and Ireland and geography, while PC2 and PC3 (Figure 3.3) capture previously reported splitting for Orkney and Wales, respectively, from Britain. The PCA plot has been rotated clockwise by 5 degrees. A map of Ireland and Britain is shown for comparison, coloured by sampling regions for cluster groups, the boundaries of which are defined based on the Nomenclature of Territorial Units for Statistics (Commission and Others 2011) (NUTS 2010), with some regions combined. Sampling regions are coloured by the cluster group with the majority presence in the sampling region. NI, Northern Ireland; PC, principal component. Cluster groups that share names with groups from Figure 3.1 (NLU; SMN; CLN; CNN) have an average of 80% of their samples shared with the initial cluster groups. Map produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))



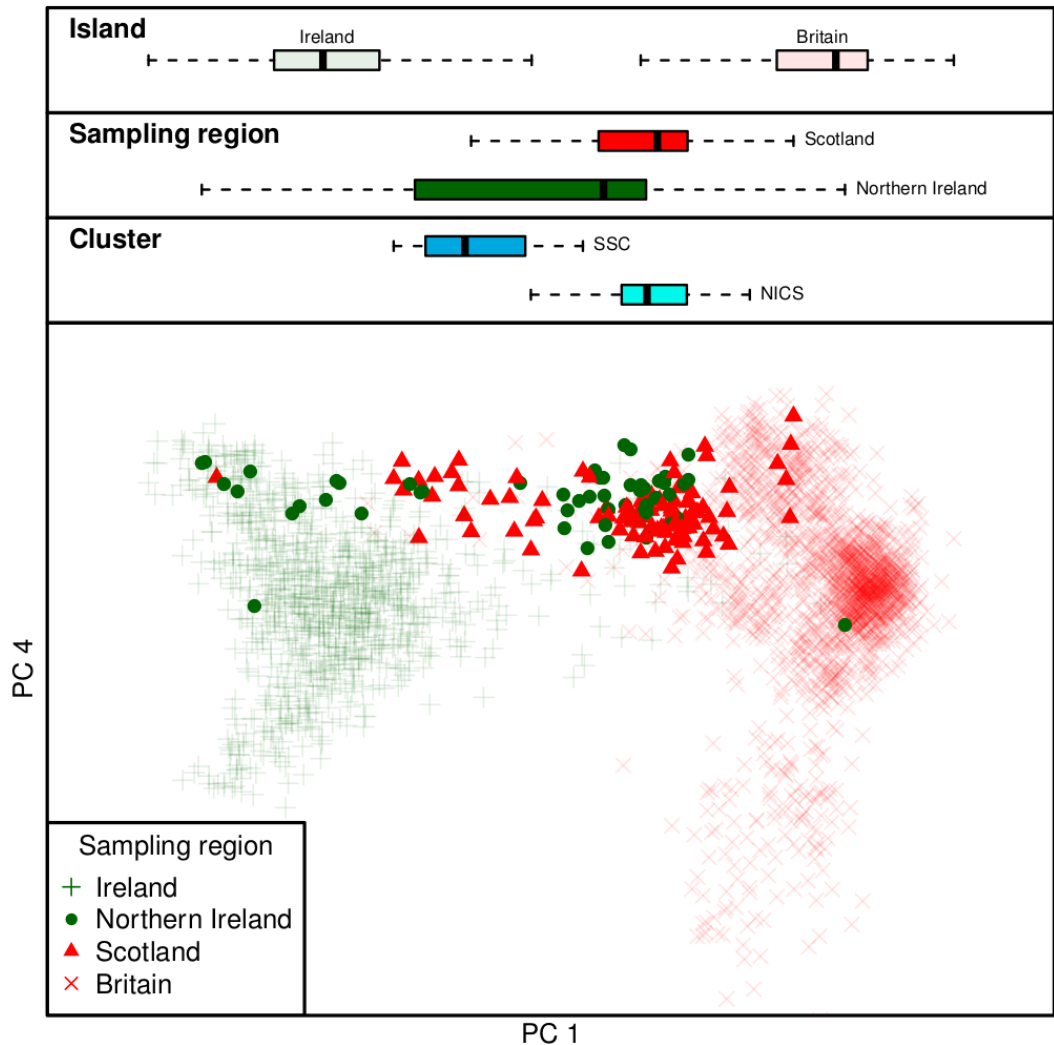


**Figure 3.3: Principal components 2 and 3 of combined Irish and British coancestry matrix.**

(A) fineSTRUCTURE clustering dendrogram for combined Irish and British data, with cluster groups defined as in Figure 3.2. Immediately following the principal inter-island split, Orkney and Wales branch in sequence, consistent with previous observations. (B) Principal component analysis (PCA) of haplotypic similarity based on the ChromoPainter coancestry matrix, coloured by cluster group with their median locations labelled. PC2 captures an Orkney split, while PC3 captures a Welsh split. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

One major advantage of cp-PCA over the hierarchical clustering dendrogram is that it displays data in a continuous space, and allows the relationships of clusters to be considered with greater nuance. For example while we can see from the fineSTRUCTURE tree that our South Scotland (SSC) cluster falls deep in the Irish branch of the tree, and the Northern Ireland/Cumbrian/South Scotland cluster (NICS) falls on the British branch of the tree, we can, however, see in cp-PC space that these clusters both span the gap between the British and Irish “genetic islands”, with sizable spread across cp-PC1 (Figure 3.4), suggesting that this genetic exchange is more continuous than discrete. This genetic exchange likely has its roots in three major historical contacts between Northern Ireland and Scotland. Firstly the third-generation PoBI Scottish samples (SSC) which look Irish are likely a result of major economic migrations from Ireland in the 19th and 20th century. Secondly the Northern Irish samples (NICS) who resemble Scottish individuals are likely so due to the impact of the major settlement of Scottish farmers in the 16th Century Ulster Plantation, which led to Scottish individuals forming a majority of the population of Ulster. Finally the suspected Irish colonisation of Scotland through the Dál Riata maritime kingdom in the 6th and 7th centuries may have established earlier genetic continuity between the two regions.

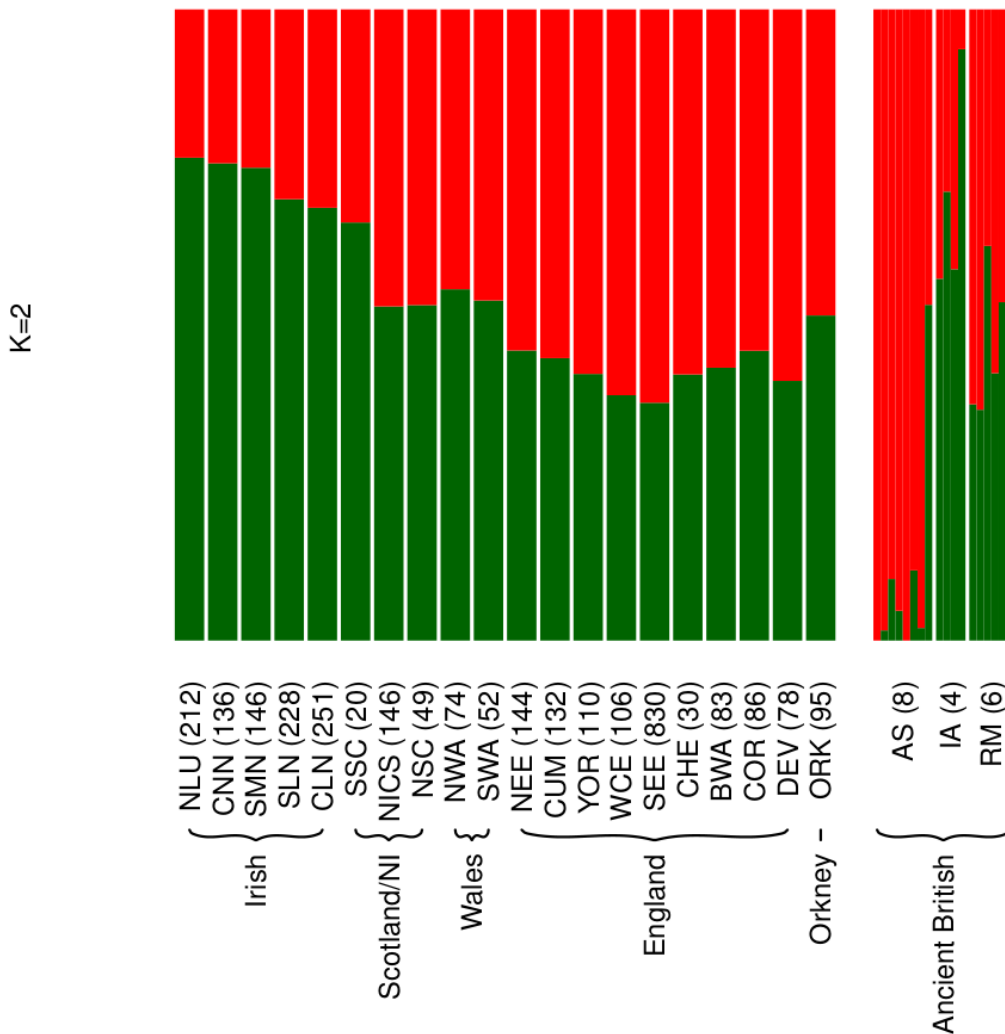
As noted for Ireland (Figure 3.1), ChromoPainter PCA demonstrates that Britain also shows eastern homogeneity and western diversity. We proposed that this may be due to the homogenising effects of migration erasing ancient structure in Ireland, a narrative which may also apply to England. The South East England cluster group (SEE), whose largest constituent cluster has over 818 individuals, and is indivisible by fineSTRUCTURE, shows the greatest homogeneity. This cluster shows the greatest affinity to the ADMIXTURE cluster ( $k=2$ ) best describing the ancient Anglo-Saxon individuals in our ADMIXTURE analysis (Figure 3.5), and much of its geographic spread is in known Anglo-Saxon territory. In contrast clusters originating from regions harbouring less Anglo-Saxon influence (Figure 3.5) separate out above and below SEE on cp-PC4, including Brittonic/Celtic speaking populations of England (Cornwall, Wales, Cumbria); the Gaelic/Celtic speaking populations of Scotland and Ireland, alongside Devon which resisted “Anglo-Saxonisation” for many years (Deacon 2007). Thus it is feasible that the Anglo-Saxon invasion is responsible for the loss of structure in the southeast of England.



**Figure 3.4: Inter-island exchange of haplotypes between the north of Ireland and northern Britain.**

The boxplots show the distribution of individuals on ChromoPainter principal component (PC) 1 for each island and for specific sampling regions (Scotland/Northern Ireland) and cluster groups (SSC and NICS; see Figure 3.2). A substantial proportion of Northern Irish individuals fall within the expected range for Scottish individuals in ChromoPainter PC space and *vice versa*. This exchange is particularly pronounced for Northern Irish and Scottish individuals that fall within the NICS and SSC cluster groups (Figure 3.2), respectively.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

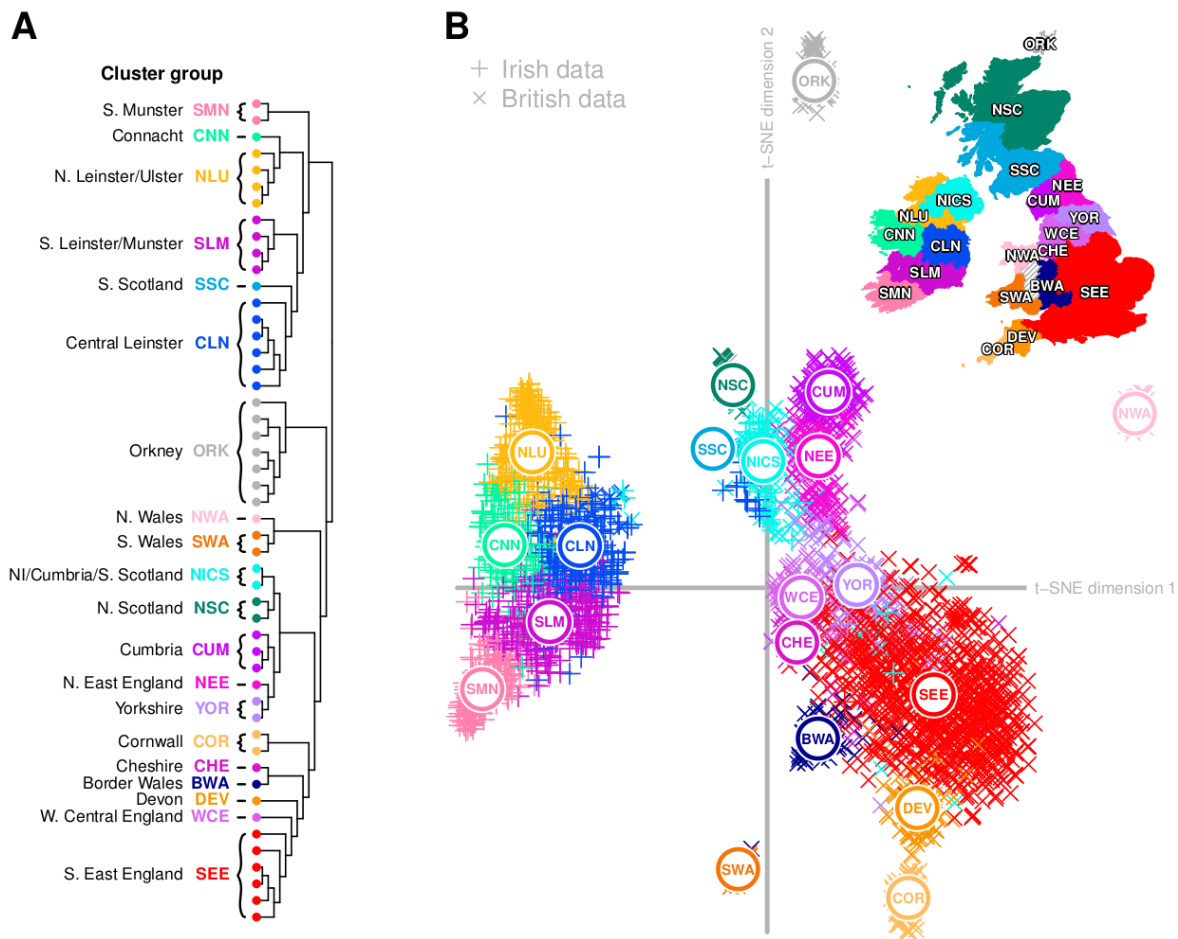


**Figure 3.5: ADMIXTURE analysis for PoBI/Irish cluster groups with ancient British samples.**

Mean ADMIXTURE component ( $k = 2$ ) for each cluster group in the PoBI/Irish fineSTRUCTURE tree (Appendix Figure 3.3) and 18 Ancient British Samples from the Iron age (IA;  $n = 4$ ), Anglo-Saxon (AS;  $n = 8$ ) and Roman (RM;  $n = 6$ ) periods. Admixture proportions are averaged across each cluster group (left) for brevity of display, while individual proportions are plotted for ancient samples. The Anglo-Saxon individuals are best described by the red component. This component is high in British cluster groups from areas affected by the Anglo-Saxon invasion such as the large SEE cluster, while relatively low in Celtic populations such as Ireland, Scotland and Wales. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

As many important cp-PCs describe important trends in the variation of the Irish-British dataset, we applied t-distributed neighbour embedding (t-SNE) (Maaten 2009) to the British/Irish coancestry matrix to find the optimal low-dimensional embedding of the data in such a way that captured both local and global structure. By iteratively grouping points together in nodes based on degree of similarity or splitting them, t-SNE can capture the trends of many principal component axes in two dimensional space and thus may be more appropriate for summarising the differences between our cluster groups. Our t-SNE projection captured the global splits between Ireland, Britain, Orkney and North and South Wales (analogous to PCs 1-3) as well as the regional splits within the Britain and Ireland nodes (reflecting PC4), supporting that the most important local and global structure in the data is well explained by cp-PCs 1-4 (Figure 3.6). We note that while t-SNE has been proposed for use with SNP data as early as 2013 (Platzer 2013) and has recently seen excellent use in visualising structure in the the large scale UK biobank (Diaz-Papkovich et al. 2019), t-SNE performed much less effectively when applied to SNP data from our dataset compared to the ChromoPainter coancestry matrix (Appendix Figure 3.2). While ChromoPainter t-SNE captures more global and local structure in data than cp-PCA, more structure is contained as a whole in the linked ChromoPainter co-ancestry matrix than independent SNP based similarity matrixes. Our publication was the first example of use of t-SNE on a ChromoPainter Coancestry matrix (R. P. Byrne et al. 2018), and we believe this method captures important trends in the data, and enables us to evaluate how well our fineSTRUCTURE clustering captures both global and local trends in the data.

Both cp-PC4 and t-SNE dimension 2 show covariation of British and Irish regional clusters along a roughly north-south axis, with northern Irish clusters (NLU and ULS) occupying similar values on PC4 to northern English (NEE and CUM) and Scottish clusters (NICS, NSC and SSC), while southern Irish clusters (SMN and SLM) occupy similar values to southern English (SEE, BWA, DEV and COR) clusters. We hypothesised that this might indicate that Ireland and Britain share a haplotypic gradient north to south. To test this hypothesis we modelled our Irish genomes as a linear mixture of British genomes and regressed haplotype sharing from northern British groups and Scotland across the Irish north-south cp-PC (cp-PC1 from Figure 3.1). This resulted in a significant correlation between sharing from North-England/Scotland and Irish cp-PC1 (Linear regression:  $p < 2 \times 10^{-16}$ ,  $r^2 = 0.24$ ). However it is possible that this signal is largely driven by increased sharing between Northern Irish and Scottish samples as discussed above.



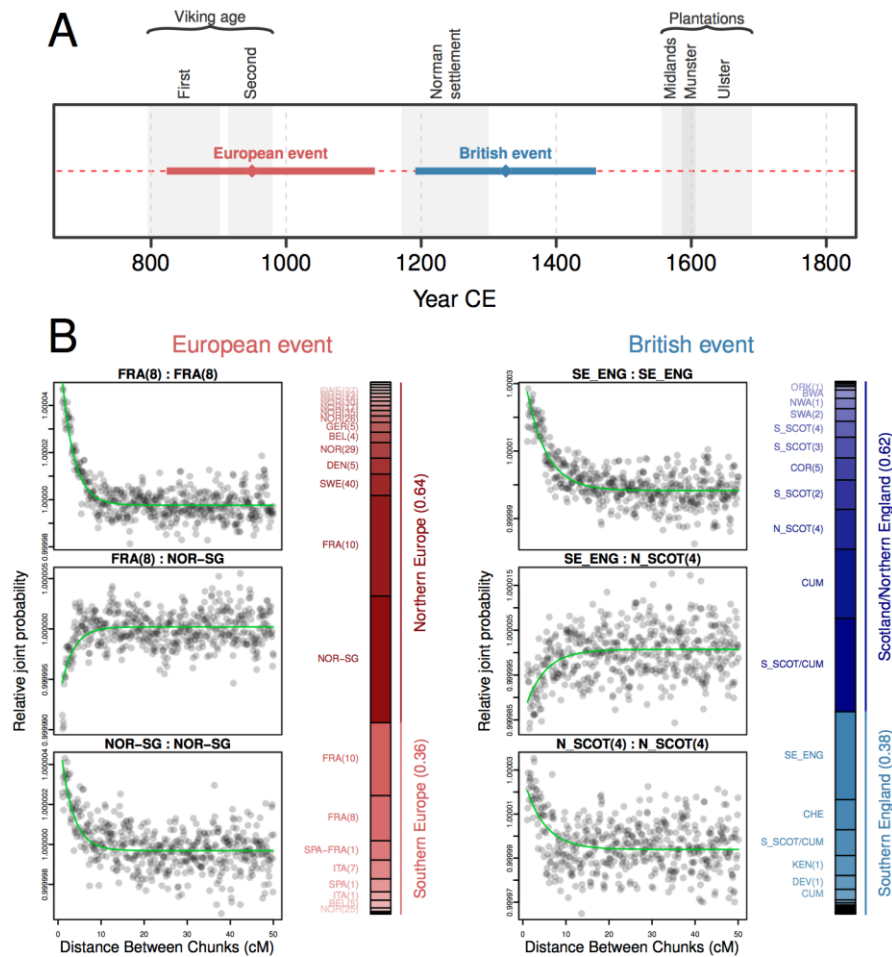
**Figure 3.6: t-distributed stochastic neighbour embedding (t-SNE) of Irish and British coancestry matrix.**

(A) fineSTRUCTURE dendrogram with clusters and cluster groups defined as in Figure 3.2. (B) Two-dimensional t-SNE embedding of ChromoPainter coancestry matrix, with median locations for cluster groups labeled. As t-SNE is a stochastic method, different runs produce different solutions to the 2-dimensional embedding; shown here is a typical result. t-SNE performed significantly better with the ChromoPainter coancestry matrix than with Hamming distances (identity-by-state) computed over single SNP markers (Appendix Figure 3.2). The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>), note some boundaries have been subsumed or modified to better reflect sampling regions. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

### 3.3.3 - Evidence of migration into Ireland

We next explored the genetic signatures left by migration events from Britain and Europe in samples from the Republic of Ireland (n=991) using GLOBETROTTER (Hellenthal et al. 2014). GLOBETROTTER uses the ChromoPainter output to model the genomes of a target population as a linear mixture of user specified donor populations, which together act as surrogates for historical admixing populations. The program works on the principle that for a simple single pulse admixture event, segments from each admixing source will decay across subsequent generations due to recombination, leaving an exponential distribution of segment lengths from those sources with rate equal to the number of generations passed. The advantage of GLOBETROTTER over softwares such as ROLLOFF (Moorjani et al. 2011) is that the actual admixing sources needn't be supplied, as the program models them as a mixture of modern surrogate populations.

GLOBETROTTER found significant evidence for an admixture event with a group best represented by modern Scandinavians and northern Europeans ( $P < 0.01$ ;  $FQ_B > 0.985$ ) with an estimated date overlapping the Norse-Viking invasion and settlement of Ireland (Figure 3.7). This signal was robust and replicated in 7/9 clusters defined using fineSTRUCTURE on the Republic of Ireland data, with the strongest signals in south and central Leinster (Figure 3.8). This strong signal in Leinster may have been bolstered by large sample sizes in Leinster clusters, leading to better statistical power, or alternatively due to the increased Viking influence in Leinster, particularly in Dublin (Formerly Dubh Linn, a Viking stronghold). This evidence for a significant contribution to the modern Irish genepool from Viking settlers contrasts the previous estimates of Viking ancestry in Ireland based on Y chromosome haplotypes which were low (McEvoy et al. 2006). Notably recent estimates of activity in Ireland from multiple archaeological proxies (radiocarbon date densities, dendrochronological date densities and entries in the Annals of Ulster) (R. McLaughlin, Hannah, and Coyle-McClung 2018) have demonstrated a decline in activity in Ireland beginning at roughly 823CE which could mark a population decline at this time. This population decline may have enabled the Viking genetic signal to persist by lowering baseline Irish diversity at the time.



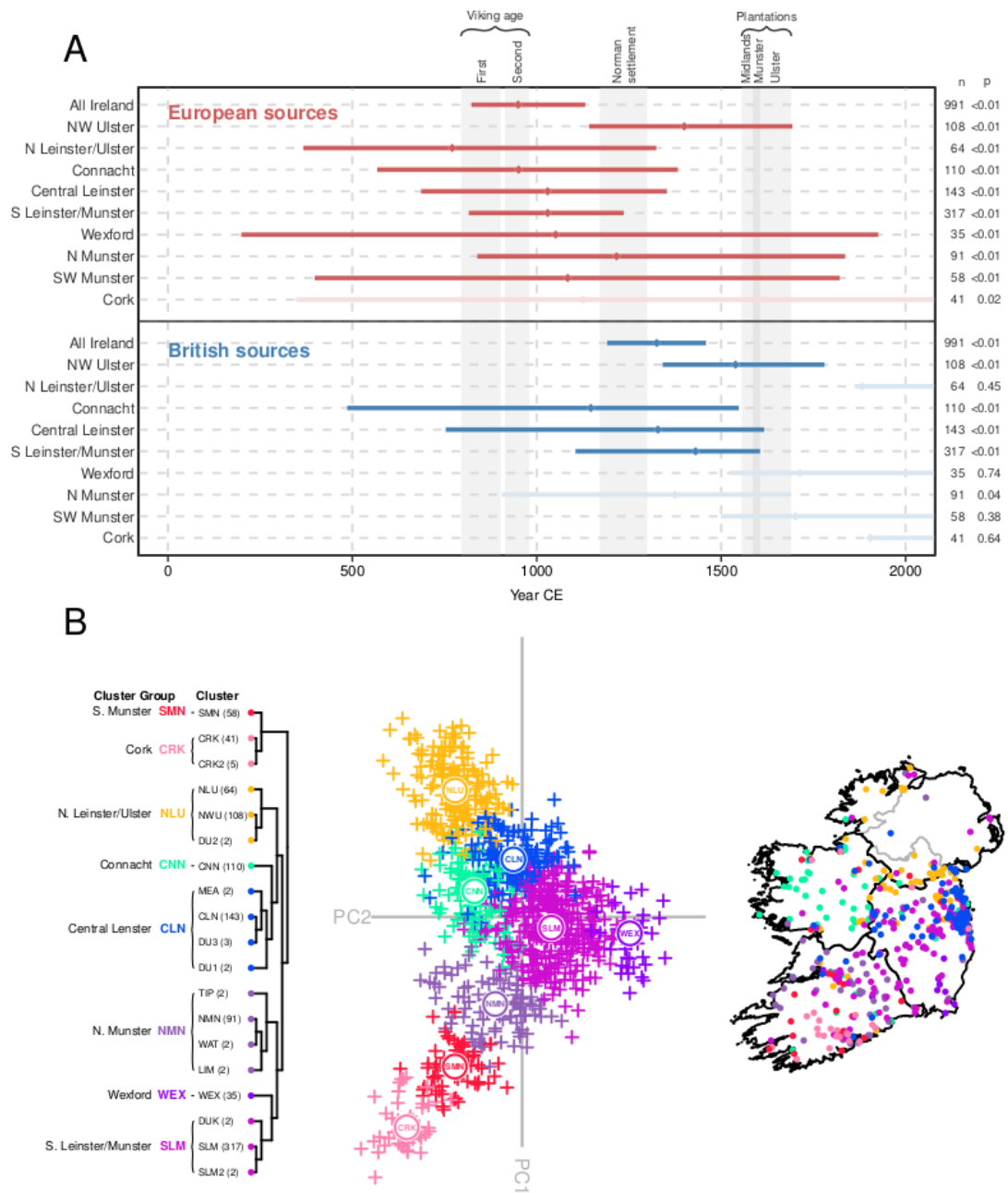
**Figure 3.7: All-Ireland GLOBETROTTER admixture date estimates for European and British surrogate admixing populations.**

A summary of (A) the date estimates and 95% confidence intervals for inferred admixture events into Ireland from European and British admixing sources (dating to the Norman and Viking invasions) with (B) ancestry proportion estimates for each historical source population for the two events and example coancestry curves.

In the coancestry curves *Relative joint probability* estimates the pairwise probability that two haplotype chunks separated by a given genetic distance come from the two modelled source populations respectively (i.e. FRA(8) and NOR-SG); if a single admixture event occurred, these curves are expected to decay exponentially at a rate corresponding to the number of generations since the event. The green fitted line describes this GLOBETROTTER fitted exponential decay for the coancestry curve. If the sources come from the same ancestral group the slope of this curve will be negative (as with FRA(8) vs FRA(8)), while a positive slope indicates that sources come from different admixing groups (as with FRA(8) vs NOR-SG). The adjacent bar plot shows the inferred genetic composition of the historical admixing sources modelled as a mixture of the sampled modern populations.

Cluster labels (for the European clustering dendrogram, see Appendix Figure 3.5; for the PoBI clustering dendrogram, see Appendix Figure 3.4): FRA(8), France cluster 8; NOR-SG, Norway, with significant minor representations from Sweden and Germany; SE\_ENG, southeast England; N\_SCOT(4) northern Scotland cluster 4.





**Figure 3.8: GLOBETROTTER breakdown for clusters in the Republic of Ireland.**

(A) Summary of the date estimates and 95% confidence intervals for inferred admixture events into Irish clusters from European (red) and British (blue) admixing sources. Faded lines highlight clusters in which there was no significant evidence of admixture ( $P > 0.01$ ).

(B) fineSTRUCTURE clustering dendrogram and PCA of the coancestry matrix describing the Republic of Ireland clusters considered in (A). Sampling locations are mapped for a subset of 544 individuals for which locational information was available. Points have been randomly jittered within a radius of 5 km to preserve anonymity. Map produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>).

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

When considering British admixing sources, GLOBETROTTER estimates from northwest Ulster overlapped the Plantation period, showing significant evidence of admixture ( $p < 0.01$ ), however due to the exclusion of Northern Irish samples from PoBI in this analysis the signal was largely diluted. The all Ireland ( $n=991$ ) estimates instead span the Norman period, although the signal is much noisier than the European event as evident in the weaker fit of the coancestry curves ( $p < 0.01$ ;  $FQ_B < 0.985$ ) (Figure 3.7), and the poor replication across clusters (4/9; Figure 3.8). It is possible that GLOBETROTTER could not adequately disentangle the complex migrational history between Britain and Ireland which has involved extensive and continuous gene flow both to and from Ireland between the major population movements outlined in Figure 3.7 and since. We also note that due to the minimal sampling from Northern Ireland in our data, the power to detect the most significant effects of the plantations is lacking in this dataset.

### 3.4 - Discussion

Our results demonstrate that significant regional population structure exists in Ireland in spite of its relative homogeneity compared to the European mainland (O'Dushlaine et al. 2010). This genetic structure shows substantial ties to geography, and hence is likely influenced both by simple isolation by distance and social and political structures which have encouraged increased haplotype sharing within restricted regions. For example much of the population structure correlates well with the provinces of Ireland, and in fact the pre-Norman kingdoms, lending to a narrative that such territories may have influenced the mobility of Irish individuals in previous generations. While much of the structure is likely deep, it is also probable that the differential effects of recent migration from neighbouring countries including Britain and mainland Europe have played a role in some of the structure observed, given that many large migration events into Ireland in the past have been geographically punctuated. For example the Plantations of the 17th century had a clear focal point in Ulster, while sites of the Viking and Norman invasions are concentrated mostly on the East coast of Ireland. Indeed we find greater affinity to British and ancient Anglo-Saxon samples on the East Coast, and strong Scottish affinity in Ulster, lending credence to the potential role of differential gene flow from neighbouring populations in creating some of the observed structure. We also note that the clearest signal of the Viking admixture we observed using GLOBETROTTER was located in eastern clusters within Leinster, while the only signal of admixture from Britain overlapping plantation is found in a Northern cluster, further supporting the possible role of differential migration from neighbouring countries in defining the structure we see in Ireland.

The structure described above is well defined, but also extremely subtle, with an average  $F_{ST}$  of 0.00036 between clusters observed. Notably a concurrent, yet independent, study of fine-scale structure carried out by Gilbert and colleagues in a smaller independent sample from the Irish population ( $n=194$ ) found similar population structure in Ireland (Gilbert et al. 2017), with comparable  $F_{ST}$  estimates between clusters (Mean  $F_{ST} = 0.0003$ ) indicating that the structure we are describing is not an artifact of our dataset, or biased by the inclusion of ALS cases in our sample. Aside from increased sample size, a major difference between our study design and the Gilbert et al study design is that we did not ascertain individuals with 4 grandparents from each region, but instead repurposed a medical genetics dataset, showing that such structure exists not only in studies optimised to detect it, but likely in all medical genetics datasets from this region. The implications of the existence of this structure for medical genetics are to date untested, however it is likely important as it highlights a source of confounding hitherto ignored. Given that this

structure was undetectable in studies relying on PCA of SNP data (Cronin et al. 2008; O'Dushlaine et al. 2010), the method most commonly used to correct for population structure in association studies, there is a strong case to be made that similar local genetic structure may act as a source of residual confounding under current designs. Moreover, within our dataset we observed that the ChromoPainter coancestry matrix captured population structure to a greater extent than an unlinked GRM (Table 3.3), suggesting these matrices commonly used in GWAS likely undercorrect population stratification. Evidence is accumulating that population stratification remains despite correction in many published GWAS, in particular GWAS for height (Bhatia et al. 2016; Sohail et al. 2019; Berg et al. 2019), suggesting additional correction for structure is an avenue worth exploring. The effects of residual population stratification in GWAS are likely to have a large impact on methods combining information across many sub genome wide significant variants, such as polygenic scores. In addition our cross-island ChromoPainter results (Figure 3.2, 3.3, 3.4 and 3.6) show that ChromoPainter captures both global and local structure where multiple countries are involved, describing differences between and within the islands, suggesting it may be applicable to multi-population GWAS. We will further explore the application of haplotype sharing matrices to correct subtle population structure in GWAS in single and multiple populations in Chapter 5.

Our study also enriches the findings of the PoBI study through inclusion of samples from the Republic of Ireland, which were excluded from the 2015 study (Leslie et al. 2015). Primarily, through PCA of the ChromoPainter co-ancestry matrix, we demonstrate that the two islands are largely genetically distinguishable along the first axis of variation when haplotype sharing is considered (Figure 3.2). The notable exceptions to this rule are Scotland and Northern Ireland, which show strong signals of interisland exchange of genetic material (Figure 3.4), reflecting known historical movements of people between these regions. The inclusion of Irish data with British samples from the PoBI study provides an anchor for Celtic ancestry in the British Isles, filling out the genetic landscape of the islands. We see north-south covariation of the islands across the fourth axis of variation, demonstrating diversity in Celtic groups. Overall our study shows that haplotypes mirror geography across Britain and Ireland, revealing nuanced links between the two which may have important implications for future medical genetic studies.

# Chapter 4 - Dutch Population Structure, Movement and Demographic Change

*Now published in Byrne et al. Dutch population structure across space, time and GWAS design. Nature Communications 2020;11 4556.*

## 4.1 - Introduction

### 4.1.1 - Background

The Netherlands is a densely populated country on the northwestern edge of the European continent, bounded by Germany, Belgium and the North Sea. The country is divided into twelve provinces and has a complex demographic history, with occupation by several Germanic peoples since the collapse of the Roman Empire, including the Frisians, the Low Saxons and the Franks. Over 17 million individuals now inhabit this relatively small region (41,500km<sup>2</sup>), making it one of the most densely populated countries in Europe. Despite its small geographical size, previous genetic studies of the people of the Netherlands have demonstrated coarse population structure that correlates with its geography, as well as apparent heterogeneity in effective population sizes across provinces (Abdellaoui, Hottenga, de Knijff, et al. 2013; Genome of the Netherlands Consortium 2014). These observations suggest that the demographic past of the Dutch population has left residual signatures in its present regional genetic structure; however this has not been fully explained in the context of neighbouring populations and thus far the use of unlinked genetic markers have limited the resolution at which this structure can be described. This resolution limit also confines the extent to which the confounding effects of population structure can be controlled in genomic studies of health and disease such as genome-wide association studies (GWAS). As these studies continue to seek ever-rarer genetic variation with ever-increasing cohort sizes, intricate understanding and fine control of population structure is becoming increasingly relevant, but increasingly challenging (Lawson et al. 2019).

Recent studies have showcased the power of leveraging shared haplotypes to uncover and characterise previously unrecognised fine-grained genetic structure within populations, yielding novel insights into the demographic composition and history of Britain and Ireland (Leslie et al. 2015; R. P. Byrne et al. 2018; Gilbert et al. 2017) (Chapter 3), Finland (Kerminen et al. 2017), Japan (Takeuchi et al. 2017), Italy (Raveane et al. 2019), France (Pierre et al. 2020) and Spain (Bycroft et al. 2019). Haplotype sharing has also revealed genetic affinities between populations (Chacón-Duque et al. 2018), enabling inference of historical admixture events using modern populations as proxies for ancestral

admixing sources (Hellenthal et al. 2014). Furthermore, geographic information can be integrated to model genetic similarity as a function of spatial distance (Novembre and Peter 2016) to infer demographic mobility within or between populations; one approach uses the Wishart distribution to estimate and map a surface of effective migration rates based on deviations from a pure isolation by distance model (Petkova, Novembre, and Stephens 2016), allowing migrational cold spots to be inferred which may derive from geographical boundaries such as rivers and mountains. Almost half of the area of the Netherlands is reclaimed from the sea and its contemporary land surface is densely subdivided by human-made waterways and naturally-occurring rivers, including the Rhine (Dutch: *Rijn*), Meuse (*Maas*), Waal and IJssel. These rivers have been speculatively linked to genetic differentiation between northern and southern Dutch subpopulations in previous work (Abdellaoui, Hottenga, de Knijff, et al. 2013); however the explicit relationship between Dutch genetic diversity and movement of people within the Netherlands has not been directly modelled.

The Dutch have previously received special interest as a model population (Abdellaoui, Hottenga, de Knijff, et al. 2013; Genome of the Netherlands Consortium 2014) and form a major component of substantial ongoing efforts to better understand human health, disease, demography and evolution. For example, at the time of writing, over 10% of all studies listed in the NHGRI-EBI genome-wide association study (GWAS) catalogue (Buniello et al. 2019) include the Netherlands in their “Country of recruitment” metadata. As well as offering insights into demography and human history, refined population genetic studies are important to identify and adequately control confounding effects in genomic studies of health and disease, especially if spatially structured environmental factors contribute substantially to variance in phenotype, which in particular impacts rare variants (Mathieson and McVean 2012). In this chapter, we harness shared haplotypes to examine the fine-grained genetic structure of the Netherlands. We show that Dutch population structure is more granular than previously recognised, with notable genetic clusters forming both between and within provinces. We demonstrate that the population structure from north to south is strong and stable over many generations, while east to west structure likely emerged more recently. These major axes of structure appear to be tied both to opposing gradients of gene flow from the neighbouring countries Germany and Belgium, as well as by the internal geographic boundary of the Rhine. Notably this genetic structure demonstrably confounds GWAS which we will explore in a later chapter (Chapter 5). Finally we investigate major changes in population size across time in the Netherlands and their relationship with the observed population structure, noting regional signals of population decline following the arrival of the Black Death to the Netherlands

(~14th century) and a major increase in population growth countrywide in the 17th century, corresponding to a period of prosperity in the Netherlands known as the Dutch Golden Age.

#### 4.1.2 - Research aims

This chapter presents work carried out to characterise population structure, demographic change and the impact of internal and external migration on the modern Dutch population using data from Dutch ALS case control cohort. The research has four major aims:

- i.) To further dissect and characterise the population structure in the Netherlands in space and time using haplotype sharing metrics;
- ii.) To contextualise structure in terms of migration events and geneflow from neighbouring countries;
- iii.) To explore regional and global changes in population size across time using haplotype sharing and characterise their interplay with population structure;
- iv.) To investigate the impact of geographic boundaries on population structure.

Our work provides several insights into the extent and potential historical relevance of genetic structure within the Netherlands. Additionally by revealing a more finescale picture of genetic structure in another small north-west European region using haplotype sharing profiles we emphasise the need to critically assess the suitability of current methods for correcting confounding in genetic association studies. The work in this chapter is as such a crucial motivation for the work in chapter 5 in testing the use of haplotype sharing methods as a method of correcting population stratification in genome-wide association studies.

NB: The results of this chapter form the majority of an article now published in Nature Communications: (<https://www.nature.com/articles/s41467-020-18418-4>).

## 4.2 - Methods

### 4.2.1 - Data and quality control

We mapped fine-grained genetic structure in the Netherlands using a population-based Dutch ALS case-control dataset (n=1,626; subset of stratum sNL3 from a genome-wide association study (GWAS) for amyotrophic lateral sclerosis (van Rheenen et al. 2016)) and a European reference dataset subsampled from a GWAS for multiple sclerosis (Sawcer et al. 2011) (MS; n = 4,514; EGA accession ID EGAD00000000120). 1,422 Dutch individuals had associated residential data (hometown at time of sampling) which were used for geographical analyses. For population structure analyses, we applied quality control (QC) using PLINK v1.9 (Chang et al. 2015); briefly we removed samples with high missingness (>10%), high heterozygosity (>3 median absolute deviations from median) and single-marker PCA outliers (>5 standard deviations from mean for PCs 1-20). We also filtered out A/T and G/C SNPs and SNPs with minor allele frequency <0.05, high missingness (>2%) or in Hardy Weinberg disequilibrium ( $p < 1 \times 10^{-6}$ ). Before running Chromopainter/fineSTRUCTURE we retained only one individual from any pair or group that exhibited greater than 7.5% genomic relatedness ( $\hat{\pi}$ ) and removed SNPs with any missing genotypes as the algorithm does not tolerate missingness or relatedness well. For European reference data we also removed individuals past a missingness threshold defined in Chapter 3 (--mind 0.0005; to maximise SNP retention at zero missingness) and individuals suggested by the QC of the source study (Sawcer et al. 2011), extracting individuals only of European descent. As this European dataset included MS patients, we filtered out SNPs in a 15 Mb region surrounding the strongly associated HLA locus (GRCh37 position chr6:22,915,594–37,945,593) to avoid bias generated from this association, following previous works using this dataset (Leslie et al. 2015; Gilbert et al. 2017; R. P. Byrne et al. 2018).

The final Dutch and European reference datasets contained 374,629 SNPs and 363,396 SNPs respectively at zero missingness. The merge of these datasets contained 147,097 SNPs at zero missingness. Data were phased per chromosome with the 1000 Genomes Project phase 3 reference panel (Auton et al. 2015) using SHAPEIT v2 (Delaneau, Marchini, and Zagury 2011) (for ChromoPainter/fineSTRUCTURE) and Beagle v4.1 (S. R. Browning and Browning 2007) (for IBD estimation). For these and all subsequent runs of SHAPEIT and Chromopainter, we used the 1000 Genomes Project Phase 3 genetic map (Auton et al. 2015); while IBD analyses with Beagle were carried out using the Hapmap phase 2 genetic map (International HapMap Consortium et al. 2007), as used in the refinedIBD (B. L. Browning and Browning 2013a) and IBDNe (S. R. Browning and



Browning 2015) source papers. Both programmes were run with default settings; allele concordance was checked prior to phasing (SHAPEIT: --check; Beagle: conform-gt utility).

#### 4.2.2 - fineSTRUCTURE analysis

We used ChromoPainter/fineSTRUCTURE (Lawson et al. 2012) to detect fine-grained population structure using default settings. In brief, each individual was painted using all other individuals (-a 0 0), first estimating  $N_e$  and (switch rate and mutation rate) with 10 expectation-maximization (EM) iterations (using all samples and chromosomes), then the model was finally run using these parameter estimates. The fineSTRUCTURE Markov chain Monte Carlo (MCMC) model was then run on the resulting coancestry matrix with two chains for 3,000,000 burnin and 1,000,000 sampling iterations, sampling every 10,000 iterations. We extracted the state with the maximum posterior probability and performed an additional 10,000 hillclimbing iterations before inferring the final trees using both the climbtree and maximum concordance methods. For all subsequent analyses the maximum concordance tree was used.

#### 4.2.3 - Cluster robustness and differentiation

To assess the robustness of clustering in the Dutch data we calculated TVD (Leslie et al. 2015) and  $F_{ST}$ . TVD is a distance metric for assessing the distinctness of pairs of clusters, calculated from the ChromoPainter chunklength matrix. TVD is calculated as the sum of the absolute differences between copying vectors for all pairs of clusters, where the copying vector for a given cluster **A** is a vector of the average lengths of DNA donated to individuals in **A** by all clusters. Intuitively, the TVD of two clusters reflects distance between those clusters in terms of haplotype sharing amongst all clusters, and is a meaningful method for assessing the effectiveness of fineSTRUCTURE clustering. To assess whether the observed clustering performed better than chance we permuted individuals between cluster pairs (maintaining cluster size) and calculated the number of permutations that exceeded our original TVD score for that pairing of clusters. We used 1,000 permutations where possible, and otherwise used the maximum number of unique permutations. P-values were calculated from the number of permutations greater than or equal to the observed TVD divided by the total permutations; all p-values were less than 0.001, indicating robust clustering. We generated a TVD tree for clusters from the k=16 fineSTRUCTURE split by merging pairs of clusters with the lowest TVD successively using methods described previously (Kerminen et al. 2017), with the goal of providing an alternative representation of cluster relationships that is independent of sample size

(Appendix Figure 4.1). The tree was built in  $k-1$  steps, with TVD recalculated at each step from the remaining populations. Branch lengths were scaled proportional to the TVD value of the corresponding pair of populations using adapted code from the original paper (Kerminen et al. 2017). Finally, to assess cluster differentiation independently of the ChromoPainter model,  $F_{ST}$  was calculated between Dutch clusters using PLINK 1.9. For this analysis we used the SNP overlap between Dutch and European datasets, pruning for LD (--indep-pairwise 1000 50 0.25) and simultaneously calculated  $F_{ST}$  between European countries present in the European dataset (Sawcer et al. 2011) for comparison.

#### 4.2.4 - Ancestry profiles

We assessed the ancestral profile of Dutch samples in terms of a European reference made up of 4,514 European individuals (Sawcer et al. 2011) from Belgium, Denmark, Finland, France, Germany, Italy, Norway, Poland, Spain and Sweden. European samples were first assigned to homogeneous genetic clusters using fineSTRUCTURE as in previous work (R. P. Byrne et al. 2018) (Chapter 3) to reduce noise in painting profiles. We then modelled each Dutch individual's genome as a linear mixture of the European donor groups using ChromoPainter, and applied ancestry profile estimation as described previously (Leslie et al. 2015) and implemented in GLOBETROTTER (Hellenthal et al. 2014) (num.mixing.iterations: 0). This method estimates the proportion of DNA which is most closely shared with each individual from each donor group calculated from a normalised ChromoPainter chunklength output matrix, and then implements a multiple linear regression of the form:

$$Y_p = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_g X_g, \quad (6)$$

to correct for noise caused by similarities between donor populations. Here,  $Y_p$  is a vector of the proportion of DNA that individual  $p$  copies from each donor group, and  $X_g$  is the vector describing the average proportion of DNA that individuals in donor group  $g$  copy from other donor groups  $G$ , including their own. The coefficients of this equation  $\beta_1 \dots \beta_g$  are thus interpreted as the "cleaned" proportions of the genome that target individual  $p$  copies from each donor group, hence the ancestral contribution of each donor group to that individual. The equation is solved using a non-negative-least squares (NNLS) function such that  $\beta_g \geq 0$  and the sum of proportions across groups equals 1. We discarded European groups that contributed less than 5% total to any individual, and refit to eliminate noise. We then aggregated sharing proportions across donor groups (genetically homogenous clusters) from the same country to estimate total sharing between an

individual and a given country to investigate the regional distribution of sharing profiles. Autocorrelation of ancestry profiles was assessed by Moran's I and Mantel's test (10,000 permutations) in R version 3.2.3. Geographical directions of ancestry gradients were determined by rotating the plane of latitude-longitude between 0 and 360 in 1 degree steps and finding the axis Y that maximised the coefficient of determination for the linear regression  $Y \sim A_c$ , where  $A_c$  is the aggregated ancestry proportion for country c .

Additionally we compared the ancestry profiles estimated by the NNLS method to those estimated using the recently developed Bayesian algorithm SOURCEFIND (Chacón-Duque et al. 2018). We ran SOURCEFIND on the ChromoPainter output described above using 50,000 burnin and 200,000 MCMC iterations, sampling every 5,000 iterations. For each Dutch individual we took the weighted average (weighted by posterior probability) of ancestry estimates with the highest posterior probability taken from 50 independent runs of the algorithm. We aggregated sharing portions across donor groups from the same country to estimate total sharing between an individual and a given country and investigate the regional distribution of sharing profiles. Ancestry gradients generated by each method were regressed against one and other to estimate correlation. We report both the results of both NNLS (Figure 4.3) and SOURCEFIND (Appendix Figure 4.2) for comparison.

#### 4.2.5 - Identity by descent analysis

IBD segments were called in phased data using RefinedIBD (B. L. Browning and Browning 2013a) (default settings). Segments shared between pairs of individuals were summed to generate a pairwise matrix of the total length of IBD shared between these individuals, analogous to the ChromoPainter coancestry matrix. We additionally separated IBD segments by length into 1 centiMorgan (cM), 1.5cM and 2cM wide bins, taking a sliding window approach with increments of 0.1cM, and computed total IBD sharing matrixes for all length windows to explore temporal changes in IBD sharing. To identify population structure captured by IBD sharing patterns we performed PCA on these matrices using the `prcomp` function in R version 3.2.3 (CoreTeam 2015) and clustered the IBD matrices using a Gaussian mixture model implemented in the R package `mclust` (Scrucca et al. 2016). We note that while previous work (Lawson and Falush 2012) has shown that IBD matrices underperform the linked ChromoPainter matrix in identifying population structure, they are arguably more interpretable for visualising temporal change as they can be subdivided into cM bins corresponding to different time periods, a feature

leveraged by emerging work on local population structure (Al-Asadi et al. 2019). Patterns in IBD sharing that identify population subgroups in older (shorter) cM bins which are preserved in more recent (longer) bins are interpreted as persistent population structure that has been influenced by mating patterns both in old and recent generations. Structure which emerges in a specific cM bin and is lost is likely to reflect transient changes in panmixia that have not necessarily persisted. We approximated the age of segments in a given cM bin using equation s19 from Al-asadi et al. (Al-Asadi et al. 2019), under the assumption that the population is sufficiently large:

$$\lim_{N \rightarrow \infty} E[T | \mu \leq l \leq \nu] = 75 \left( \frac{1}{L_1} + \frac{1}{L_2} \right), \text{ (Al-Asadi et al. 2019)} \quad (7)$$

where T is the random coalescence time in generations, l is the length of a segment (in base pairs),  $\mu$  and  $\nu$  are the upper and lower segment length bounds of the interval (in base pairs) and L2 and L1 are the upper and lower bounds of the interval rescaled to centiMorgan (i.e. multiplied by 100r, where r is the recombination rate). For the age estimates given in Figure 4.4, we multiplied the expected coalescence time in generations by the approximate human generation time (28 years).

We generated an interactive web environment to visualise and explore population structure and clustering for all segment bins which is hosted at [bioinf.gen.tcd.ie/ctg/nlibd](http://bioinf.gen.tcd.ie/ctg/nlibd).

#### 4.2.6 - Inferring admixture dates

To infer and date admixture events from European sources we ran GLOBETROTTER (Hellenthal et al. 2014) with the Netherlands dataset as a whole and in individual cluster groups defined from the Dutch fineSTRUCTURE maximum concordance tree (Figure 4.1). To define European donor groups we used the European fineSTRUCTURE maximum concordance tree, as with previous work (R. P. Byrne et al. 2018)(Chapter 3) to ensure genetically homogenous donor populations. We used ChromoPainter v2 to paint Dutch and European individuals using European clusters as donor groups (estimating  $N_e$  and  $\mu$  using the weighted average of 10 EM iterations on chromosomes 1, 8, 15 and 20, using all samples). This generated a copying matrix (chunklengths file) and 10 painting samples for each Dutch individual. GLOBETROTTER was run for 5 mixing iterations twice: once using the null.ind:1 setting to test for evidence of admixture accounting for unusual linkage disequilibrium (LD) patterns and once using null.ind:0 to finally infer dates and sources. We further ran 100 bootstraps for the admixture date and calculated the probability of no admixture as the proportion of nonsensical inferred dates (<1 or >400 generations).

Confidence intervals were calculated from the bootstraps from the standard model (null.ind:0) using the empirical bootstrap method, and a generation time of 28 years.

#### 4.2.7 - ADMIXTURE analysis

We performed ADMIXTURE analysis (Alexander, Novembre, and Lange 2009) on the combined Dutch and European samples to explore single marker-based population structure in a set of 41,675 unlinked SNPs (LD-pruned using PLINK 1.9:  $r^2 > 0.1$ ; sliding window 50 SNPs advancing 10 SNPs at a time). ADMIXTURE was run for  $k=1-10$  populations, using 5 EM iterations at each  $k$  value. The  $k$  value with the lowest cross-validation error was selected for further analysis using 15 fold cross-validation; where two  $k$  values had equal CV-error the lower  $k$  value was taken for parsimony (Appendix Figure 4.3). We analysed the distribution of proportions for each ADMIXTURE cluster across the Dutch dataset, and its relationship with geography.

#### 4.2.8 - Estimating mean pairwise IBD sharing

We compared IBD sharing within and between both clusters and provinces (Appendix Figure 4.4) using the mean number of segments within a given length range (e.g. 1-2cM) shared between individuals. To calculate this mean for a single group of size  $N$  with itself the denominator was  $(N^2 - N)/2$ ; when comparing two groups of sizes  $N$  and  $M$  the denominator was  $NM$ .

#### 4.2.9 - Estimating recent changes in population size

We used IBDNe (S. R. Browning and Browning 2015) to estimate historical changes in  $N_e$ . IBDNe leverages information from the length distribution of IBD segments to accurately estimate effective population size over recent generations, with a resolution limit of about 50 generations for SNP data. We followed the authors' protocol and detected IBD segments using IBDseq version r1206 (B. L. Browning and Browning 2013b) with default settings and ran IBDNe on the resulting output with default settings, removing IBD segments shorter than 4cM (minibd=4, the recommended threshold for genotype data).

We compared estimated  $N_e$  with recorded census size

(<https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=1520261958200>) for approximately equivalent dates (starting at 1946 CE for generation 0 and assuming 1 generation is 28 years) and found that for generations 0 - 3 our  $N_e$  estimates were approximately  $\frac{1}{3}$  of the census population (Appendix Figure 4.5), which follows expectation if lifespan is  $\sim 3\times$  the generation time (S. R. Browning and Browning 2015;

Felsenstein 1971). The slope of the ratios for the three generations is near zero suggesting that our model tracks well with the census population; this is consistent with reported expectation (S. R. Browning and Browning 2015).

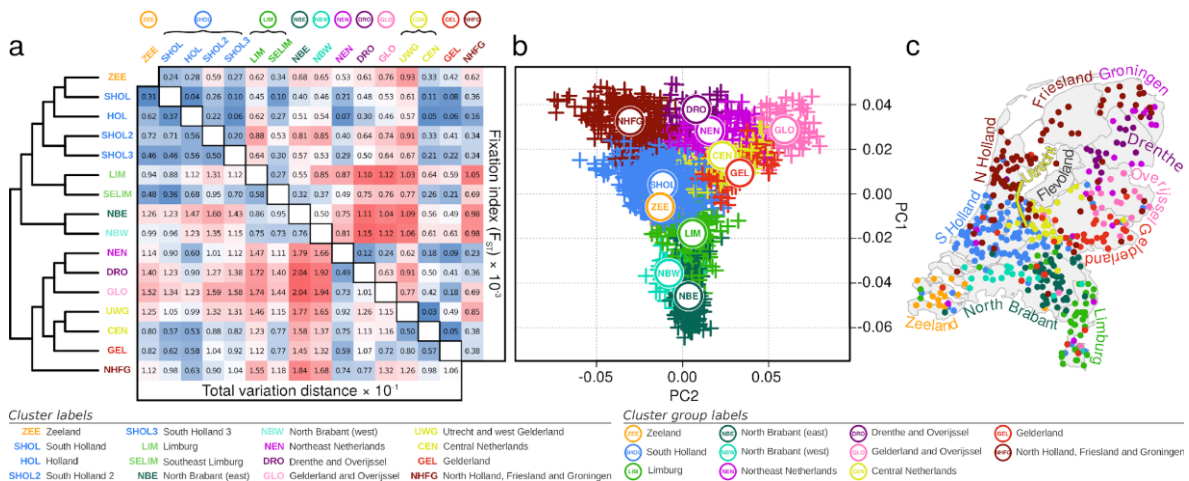
#### 4.2.10 - Estimating effective migration surfaces

To model geographic barriers to geneflow in the Netherlands we ran EEMS (Petkova, Novembre, and Stephens 2016). This software provides a visualisation of hot and coldspots for geneflow across a habitat using a geocoded genetic dataset. To run EEMS, we generated an average pairwise genetic dissimilarity matrix from our genotype data using the `bed2diffs` utility provided with the software. We initially ran the EEMS model with 10 randomly initialised MCMC chains for a short run of 100,000 burn-in and 200,000 sampling iterations, thinning every 999 iterations, to find a suitable starting point. For these runs we placed the data in 800 demes and used default settings with the following adjustments to the proposal variances: `qEffctProposalS2 = 0.000088888888`; `qSeedsProposalS2 = 0.7`; `mEffctProposalS2 = 0.7`. The resulting chain with the highest log-likelihood was then used as the starting point for a further ten chains for 1,000,000 burn-in iterations and 2,000,000 sampling iterations, thinning every 9,999 iterations. The model was run for these chains with the same adjustments to the proposal variances as above. We plotted the results of our analysis using the `rEEMSplot` package in R and modified the resulting vector graphics using Inkscape v0.91 to remove display artefacts caused by non-overlapping polygons. MCMC convergence was assessed by inspecting the log-posterior traces (Appendix Figure 4.6).

## 4.3 - Results

### 4.3.1 - The genetic structure of the Dutch population

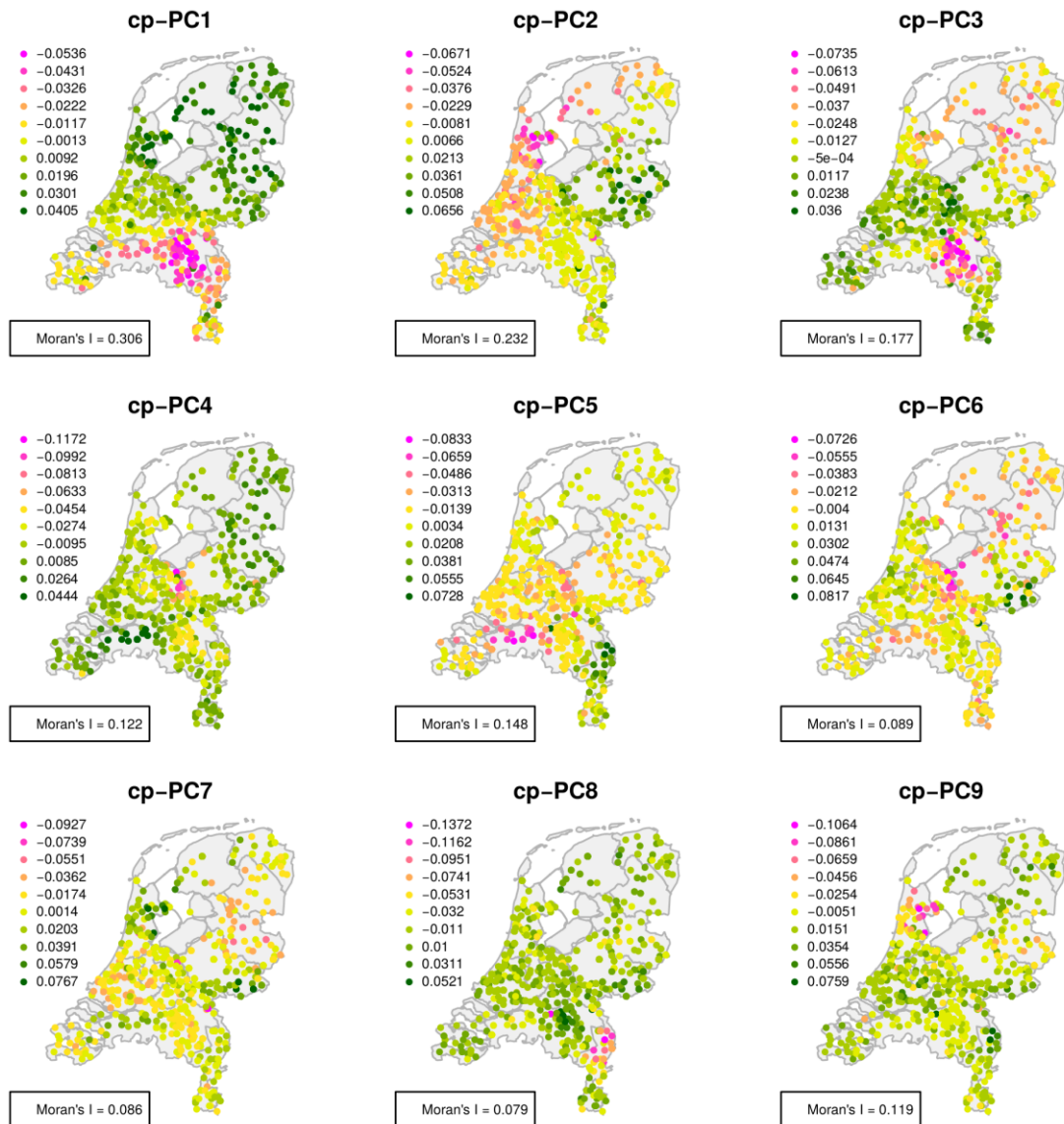
We summarised the haplotypic coancestry between 1,626 Dutch individuals using ChromoPainter (Lawson et al. 2012) and clustered the resulting matrix using fineSTRUCTURE (Lawson et al. 2012), identifying 40 genetic clusters at the highest level of the hierarchical tree which segregated with geographical provenance. We explored the clustering from the finest ( $k=40$ ) to the coarsest level ( $k=2$ ), settling on  $k=16$  as it captured the major regional splits sufficiently with little redundancy. Clusters at this level were robustly defined, as evidenced by permutation testing of total variation distance (TVD;  $p < 0.001$ ) which showed non-random differences in haplotype sharing between clusters. In addition clusters showed subtle differentiation even when measured using fixation index ( $F_{ST}$ ; Figure 4.1 a), which relies on unlinked markers rather than haplotype sharing; remarkably, some  $F_{ST}$  values between particularly differentiated Dutch clusters were comparable in magnitude to estimates between some European countries (calculated using data from the European reference dataset (Sawcer et al. 2011); Appendix Table 4.1). Some clusters had expansive geographical ranges (for example NHFG, representing individuals from North Holland, Friesland and Groningen), while others neatly distinguished populations on a sub-provincial level (for example, NBE and NBW, representing east and west regions of North Brabant). For visualisation we projected the ChromoPainter coancestry matrix in lower dimensional space using principal component analysis (PCA; Figure 4.1 b) and assigned cluster labels based on majority sampling location (available for 1,422 individuals), arranging neighbouring and genetically similar clusters into cluster groups, as with previous work (R. P. Byrne et al. 2018) (Chapter 3). The first principal component (PC) of coancestry followed a strong north-south trend (latitude vs mean PC1 per town  $r^2 = 0.52$ ;  $p = 6.8 \times 10^{-72}$ ) with PC2 generally explained by a west-east gradient (longitude vs mean PC2 per town  $r^2 = 0.29$ ;  $p = 3.4 \times 10^{-33}$ ). Further PCs demonstrated more complex non-linear relationships with geography, with evidence of significant spatial autocorrelation when tested using Moran's I (Figure 4.2). These PCs showed extreme values for extended geographic regions (e.g. the central band in PC3), or single provinces (e.g. North Holland on PC9), likely reflecting genetic affinities within these geographic areas relative to the rest of the Netherlands.



**Figure 4.1: The genetic structure of the people of the Netherlands.**

(a) fineSTRUCTURE dendrogram of ChromoPainter coancestry matrix showing clustering of 1,626 Dutch individuals based on haplotypic similarity. Associated total variation distance (TVD) and fixation index statistics between clusters are shown in the matrix. Permutation testing of TVD yields  $p < 0.001$  for all cluster pairs, indicating that clustering is non-random. Cluster labels derive from Dutch provinces and are arranged into cluster groups for genetically and geographically similar clusters (circled labels). (b) The first two principal components (PCs) of ChromoPainter coancestry matrix for all individuals analysed. Points represent individuals and are coloured and labelled by cluster group. (c) Geographical distribution of 1,422 sampled individuals, coloured by cluster groups defined in (a). Labels represent provinces of the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

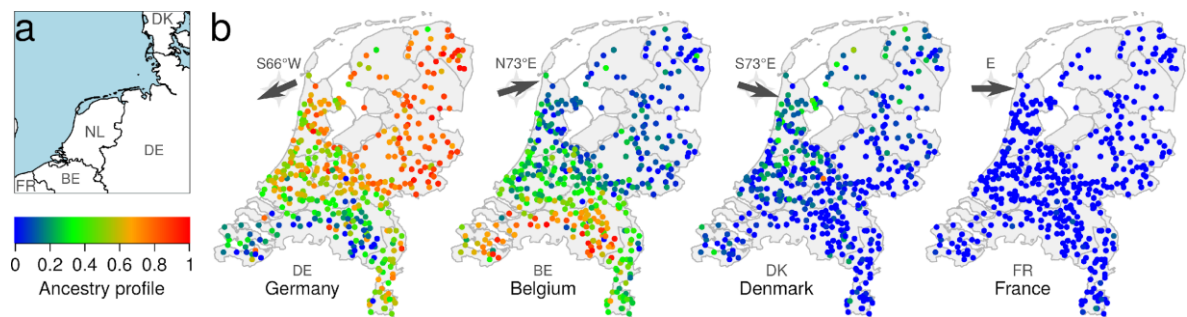




**Figure 4.2: ChromoPainter-PCs relationship to geography.**

Points on maps are coloured by the average ChromoPainter PC value per town of sampling. Each displayed PC shows a significant relationship with geography as tested by Moran's I ( $p < 0.0001$  for all PCs). PCs have been split into 10 bins for visualisation purposes as in previous works (Abdellaoui, Hottenga, de Knijff, et al. 2013). Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

As previously observed in Ireland (R. P. Byrne et al. 2018) (Chapter 3), the distribution of individuals across the PC1 vs PC2 projection generally resembled their geographic distribution (Figure 4.1 c), with some exceptions. For example, North Brabant is geographically further north than Limburg, but its major genetic cluster is further separated by PC1 from northern clusters than the Limburg cluster. We explored the possibility that this could instead be explained by relative ancestral affinities to neighbouring lands by modelling the genome of each Dutch individual as a linear mixture of European sources (obtained from the European dataset (Sawcer et al. 2011)) using ChromoPainter, retaining source groups that best matched Dutch individuals for at least 5% of the genome (Leslie et al. 2015) (Figure 4.3). The resulting profiles of German, Belgian and Danish ancestries were significantly autocorrelated ( $p_{DE}$  and  $p_{BE} < 0.0001$ ;  $p_{DK} < 0.001$ ; Moran's I and Mantel's test) and spatially arranged along geographical directions S66°W, N73°E and S73°E respectively, approximately corresponding to declining ancestry gradients directed away from the German and Belgian borders and the North Sea boundary (Figure 4.3;  $r_{DE}^2 = 0.31$ ;  $r_{BE}^2 = 0.35$ ;  $r_{DK}^2 = 0.12$ ;  $p_{DE} = 9.4 \times 10^{-119}$ ;  $p_{BE} = 2.7 \times 10^{-133}$ ;  $p_{DK} = 1.1 \times 10^{-39}$ ). In contrast the spatial distribution of French ancestry was comparatively uniform, with only a modest correlation due east ( $r_{FR}^2 = 0.014$ ;  $p_{FR} = 9.5 \times 10^{-6}$ ). The major trend across the Netherlands was thus of complementary Belgian and German ancestral affinities, decaying with distance from the respective borders. North Brabant clusters, however, showed a greater Belgian profile than Limburg clusters, despite similar, substantial Belgian frontiers in both Dutch provinces, potentially explaining the relative differentiation of North Brabant clusters from northern clusters compared to Limburg. Conversely, the German ancestry profile of Limburg greatly exceeded that of North Brabant, reflecting its 200-kilometre border with Germany and centuries of consequent demographic contact and likely genetic admixture. This analysis produced largely identical results when run with the NNLS (Figure 4.3) and SOURCEFIND methods (Appendix Figure 4.2), however while single marker ADMIXTURE runs demonstrate some degree of spatial structure in admixture components, they could not deconvolute the opposing ancestry gradients as clearly (Appendix Figure 4.7).



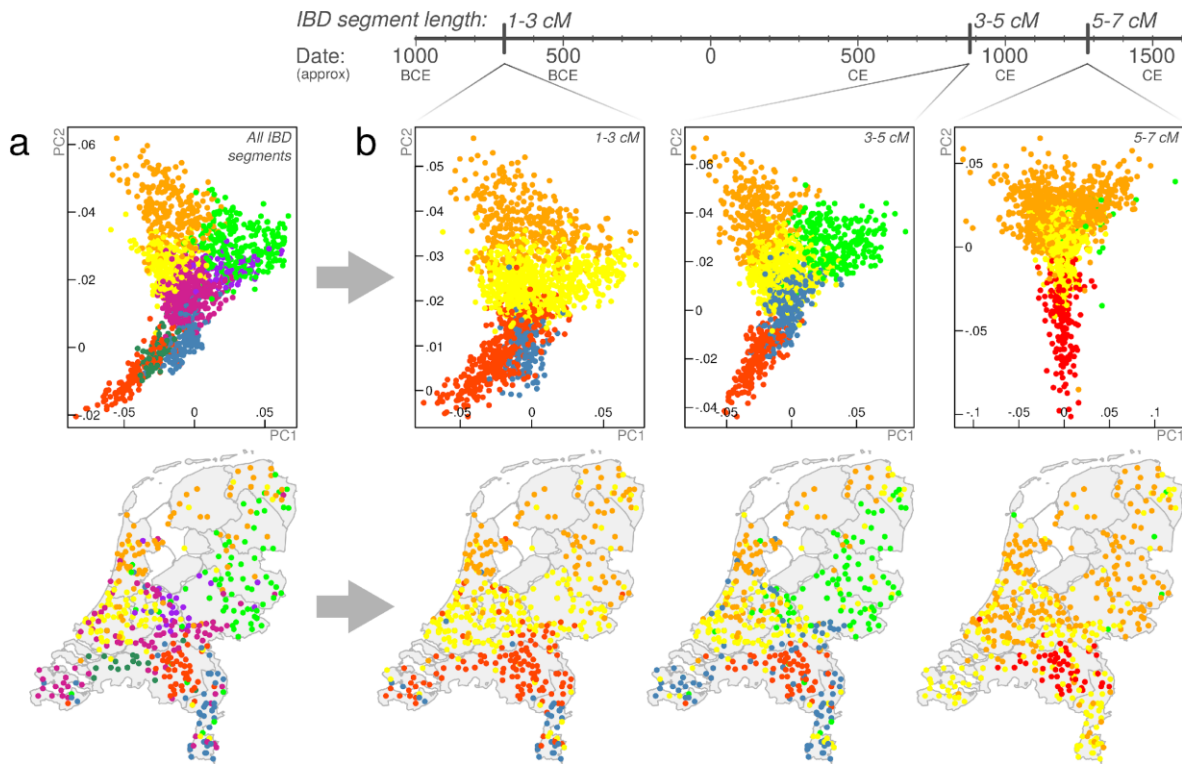
**Figure 4.3: The ancestry profile of the Netherlands.**

(a) The Netherlands and its geographical relationship to neighbouring lands. (b) German, Belgian, Danish and French haplotypic ancestry profiles for 1,422 Dutch individuals. Arrows indicate the predominant directions along which the ancestry gradients are arranged across the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>) and Natural Earth (<https://naturalearthdata.com>).

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

#### 4.3.2 - Genome flux and stasis in the Netherlands

To explore temporal trends in Dutch population structure we called genomic segments of pairwise identity-by-descent (IBD) using RefinedIBD (B. L. Browning and Browning 2013a). An IBD haplotype sharing matrix is conceptually similar to a ChromoPainter coancestry matrix (Lawson and Falush 2012), but trades some sensitivity to be more explicitly interpretable. As IBD segment length is inversely related to age (Palamara 2014; Al-Asadi et al. 2019), different length intervals can inform on structure at different time depths. Total pairwise IBD between Dutch individuals mirrored the structure observed with ChromoPainter (Figure 4.4 a), with 8 distinct clusters identified in the IBD sharing matrix that broadly segregated with geography and recapitulated some of the important splits obtained from fineSTRUCTURE, most strikingly the west-east split in North Brabant. Decomposing total IBD by centiMorgan (cM) length into short (1-3 cM), medium (3-5 cM) and long (5-7 cM) bins, we observed stability over time of north-south structure and the emergence of west-east structure embedded in 3-5 cM segments (Figure 4.4 b), corresponding to an expected time depth around 1,120 years ago (Al-Asadi et al. 2019). As this date and the structure observed is dependent on the (arbitrary) thresholds set for IBD segment length bins, we have also provided an interactive environment in which Dutch population structure can be explored across a range of IBD segment bins ([bioinf.gen.tcd.ie/ctg/nlibd](https://bioinf.gen.tcd.ie/ctg/nlibd)).



**Figure 4.4: The changing genomic structure of the Dutch population over time.** (a) Principal component (PC) analysis of pairwise total identity-by-descent (IBD) for 1,626 Dutch individuals (top) and their geographical provenance (bottom). Points represent individuals and are coloured by cluster assignment (mclust on pairwise IBD matrix). (b) PCs (top) and geographical provenance (bottom) for pairwise sharing of 1-3, 3-5 and 5-7 centiMorgan (cM) IBD segments, corresponding to point estimates of expected time depths at approximately 2,700, 1,120 and 720 years ago, respectively. Time depths for IBD segment bins have wide distributions (Al-Asadi et al. 2019); expected values presented here should be interpreted as a guide only and the changing west-east structure over time does not necessarily reflect (for instance) a precisely-timed admixture event. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>). Interactive visualisation of structure for different segment bins is available at: [bioinf.gen.tcd.ie/ctg/nlibd](https://bioinf.gen.tcd.ie/ctg/nlibd) (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

Although these observations could potentially be biased by differences in power to detect population structure in longer and shorter IBD segment bins, the temporally volatile west-east structure contrasts with the stability and persistence of old north-south structure and possibly represents a genomic signature of historical demographic flux in the region and its surrounding lands. With this in mind, we investigated possible admixture from outside demographic groups using GLOBETROTTER (Hellenthal et al. 2014) with 4,514 European individuals (Sawcer et al. 2011) representing modern proxies for admixing sources. Across the Dutch sample, a significant admixture event dating to 1088 CE (95%

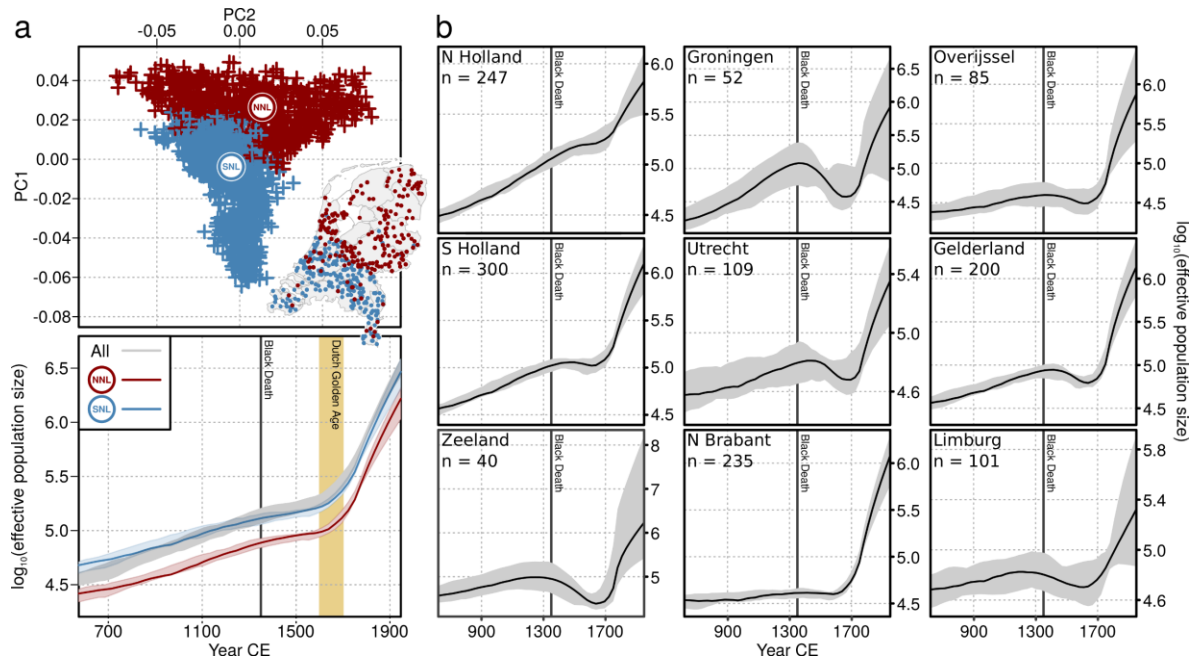
c.i. 1004-1111 CE) was inferred with the major contributing source best modelled by modern Germans and the minor source best modelled by southern European groups (France, Spain) (Table 4.1). This is supported by single-marker ADMIXTURE component estimates showing that the Netherlands has the closest profile to Germanic groups (Appendix Figure 4.7) and is consistent with the ancestry profile gradients detailed in Figure 4.3. The timing of the inferred 11<sup>th</sup> century event was stable across Dutch fineSTRUCTURE clusters (to varying degrees of confidence), suggesting that the signal represents an important period in the establishment of the modern Dutch genome (Table 4.1); however, given the state of demographic flux in Europe at the time, its exact historical correlate is open to interpretation. Notably, a significant admixture event with a major Danish source was inferred between 759 and 1290 CE in the NHFG cluster group (representing Dutch northern seaboard provinces); this period spans a historical period of recorded Danish Viking contact and rule in northern Dutch territories.

**Table 4.1: Date and source estimates for admixture into the Netherlands.**

Cluster group	Conclusion	Minor	Major	Prop	Date CE	95% c.i. CE	p
SHOL	one-date multiway	SPA-FRA(2)	GER(5)	0.25	1169	1086-1244	0
ZEE	one-date-multiway	FRA(8)	GER(5)	0.4	1172	771-1773	0
NBE	one-date-multiway	FRA(8)	GER(5)	0.4	1085	939-1262	0
NBW	one-date-multiway	GER(5)	BEL(5)	0.34	1013	668-1383	0
NEN	one-date	SPA-FRA(2)	GER(5)	0.19	1172	925-1364	0
DRO	one-date-multiway	FRA(8)	GER(5)	0.16	1390	1116-1932	0
GLO	one-date	SPA-FRA(2)	GER(5)	0.14	1128	893-1306	0
CEN	one-date	SPA-FRA(2)	GER(5)	0.18	1049	854-1244	0
GEL	one-date	SPA-FRA(2)	GER(5)	0.17	1189	1046-1391	0
NHFG	one-date	GER(9)	DEN(5)	0.36	1060	759-1290	0
LIM	one-date	ITA(8)	GER(5)	0.34	1162	1044-1351	0
<b>ALL</b>	one-date	SPA-FRA(2)	GER(5)	0.25	1088	1004-1111	0

Describes the GLOBETROTTER results for European admixture into the Netherlands for each cluster group. The Minor and Major column represent inferred proxy admixing sources. Prop represents estimated admixture proportion from the minor admixing source. Admixing sources are derived from ChromoPainter/fineSTRUCTURE clustering of 4,514 European reference individuals (Methods); labels represent principal country of origin (SPAin, FRAnce, GERmany, BELgium, DENmark) with cluster numbers arbitrarily assigned within countries. (Table reprinted from Byrne et al. (R. P. Byrne et al. 2020))

In addition to influence from outside populations, the population structure detailed in Figure 4.1 and Figure 4.4 has likely been shaped by independent regional demographic histories within the Netherlands. In support of this, we noted that short (1-2 cM) IBD segments shared between northern clusters and provinces outnumbered those shared between southern clusters and provinces (Appendix Figure 4.4), and, as observed previously (Genome of the Netherlands Consortium 2014), northern provinces shared more short segments with southern provinces than southern provinces shared amongst themselves. Together, these results suggest that the north had a smaller ancestral effective population size ( $N_e$ ) than the south and is probably derived from an ancient or historical founder event forming the northern population from a subset of southerners. We formally characterised ancestral trajectories in  $N_e$  for the north and the south of the Netherlands using the nonparametric method IBDNe (S. R. Browning and Browning 2015), using two subsamples representing the principal fineSTRUCTURE north/south split (Figure 4.5 a), retaining a random sample of 641 individuals from each group. We also characterised historical  $N_e$  for countrywide Dutch samples (Figure 4.5 a) and within individual Dutch provinces for which genotypes for more than 40 individuals were available (Figure 4.5b). Countrywide,  $N_e$  has grown superexponentially over the past 50 generations in the Netherlands (Figure 4.5 a) and has been consistently lower in the north than the south. Despite this, the pattern of growth in northern and southern groups was nearly identical, with a steady exponential growth up to around 1650 CE, when a major uptick in growth rate was observed. This corresponds to a period of substantial economic development in the Netherlands over the 17<sup>th</sup> century known to historians as the Dutch Golden Age. Preceding this period, historical  $N_e$  estimates for the entire country and for northern/southern groups showed only a modest response to the Black Death (*Yersinia pestis* plague pandemic) of the 14<sup>th</sup> century which claimed up to 60% of Europe's population (Herlihy 1997). Conversely,  $N_e$  estimation within individual Dutch provinces revealed a much more detectable impact of the Black death, showing evidence of population decline in the majority of provinces following this historical event (Figure 4.5 b).



**Figure 4.5: Dutch effective population size over time.**

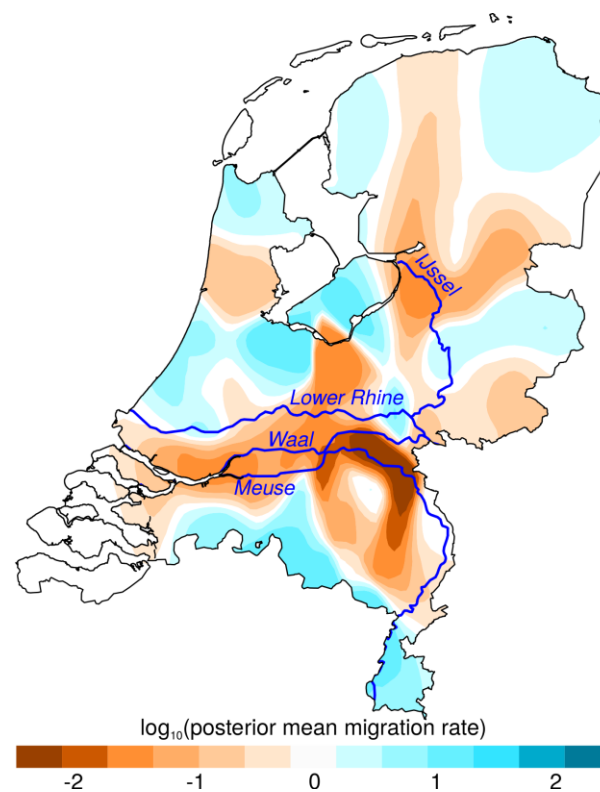
(a) Historical change in effective population size ( $N_e$ ) over the past 50 generations for all Dutch individuals and subsets of northerners and southerners. The top plot shows the principal components of ChromoPainter coancestry coloured by the first ( $k=2$ ) fineSTRUCTURE split, which separates the Dutch population into northern (NNL) and southern (SNL) genetic clusters; inset shows geographical distribution of these individuals. The bottom plot shows growth in effective population size countrywide or per fineSTRUCTURE cluster over the past 50 generations. (b) Historical  $N_e$  trajectories for individual Dutch provinces with more than 40 individuals sampled.  $N_e$  plots show estimates  $\pm 95\%$  confidence interval. and assume 28 years per generation and mean year of birth at 1946 CE. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

#### 4.3.3 - Genomic signatures of Dutch mobility

We noted that long ( $>7$  cM) IBD segments, which capture recent shared ancestry, were almost always shared within genetic clusters (and provinces), and rarely between (Appendix Figure 4.4). This indicates a propensity for genetically similar individuals (relatives) to remain mutually geographically proximal in recent years, suggesting a degree of sedentism that has likely influenced Dutch population structure over time. It has also previously been argued that genetic structure in the Netherlands may be partially rooted in geographic obstacles imposed by the country's major waterways (Abdellaoui, Hottenga, de Knijff, et al. 2013) so we explicitly modelled genetic similarity as a function of

geographic distance using EEMS (Petkova, Novembre, and Stephens 2016) to infer migrational hot and cold spots (Figure 4.6). The resulting effective migration surface showed several apparent barriers to gene flow, the strongest and most contiguous of which runs in an east-west direction across the Netherlands overlapping the courses of the Rhine, Meuse and Waal rivers. This inferred migrational boundary also approximately corresponds to the geographical division determining the principal fineSTRUCTURE split between northern and southern Dutch populations (Figure 4.5 a) as well as the geographical boundaries between clusters inferred from ancient IBD segments (Figure 4.4 b), suggesting that these rivers have been a historically persistent determinant of Dutch population structure.



**Figure 4.6: The effective migration surface of the Netherlands.**

Contour map shows the mean of 10 independent EEMS posterior migration rate estimates between 800 demes modelled over the land surface of the Netherlands. A value of 1 (blue) indicates a tenfold greater migration rate over the average; -1 (orange) indicates tenfold lower migration than average. The courses of major rivers are included to highlight their correlation with migrational cold spots. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>); river course data from Natural Earth (<https://www.naturalearthdata.com>).

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



## 4.4 - Discussion

In this chapter we have studied a densely sampled cohort of modern individuals from the Netherlands, harnessing information from shared haplotypes and recent developments in spatial modelling to gain insights into the geospatial distribution and likely origin of Dutch population genetic structure. The structure identified through shared haplotypes is surprisingly strong; some Dutch genetic clusters identified this way are more mutually distinct (by  $F_{ST}$ ) than pairs of European countries. We characterised changing population structure over time, using our novel method leveraging length-binned IBD sharing combined with PCA and Gaussian mixture models, revealing transient genetic structure layered over strong and stable north-south differentiation in the Netherlands. This north-south genetic split is contextualised by differing demographic histories between genetic groups in the Netherlands, with consistently lower  $N_e$  in the north than the south, suggesting an old split with little subsequent mixing between the groups. Combining genetic and geospatial data in a migration surface model suggests that east-west courses of the Rhine, Meuse and Waal form a boundary for this split, implicating impaired migration across these waterways as a potential cause for this population divide. Opposing ancestry gradients from Germany and Belgium overlaying this genetic structure suggest differential exchange with external groups likely contributes to the structure observed, which may be amplified by this internal boundary.

The population structure observed in the Netherlands is especially remarkable when considered in terms of the country's size and extensive infrastructure which might be expected to homogenise the population by reducing distance barriers to random mating; notably Denmark, which is roughly equal in geographical area, is in contrast genetically homogeneous, forming only a single cluster when interrogated using fineSTRUCTURE (Athanasiadis et al. 2016), despite its island-rich geography which would be expected to form isolated population sub groups. Moreover, while both the United Kingdom and Ireland exhibit at least one large indivisible cluster constituting a large fraction of their population when analysed with fineSTRUCTURE (Leslie et al. 2015; Gilbert et al. 2017; R. P. Byrne et al. 2018), no extraordinarily large clusters dominate the Dutch sample. In addition, mean  $F_{ST}$  between Dutch clusters is relatively high (mean  $F_{ST}=0.0005$ ) and outmeasures that observed between Irish clusters ( $F_{ST}=0.00036$ ; Chapter 3), suggesting that the extent of population differentiation between clusters is higher in the Netherlands, despite Dutch land area being less than half that of the island of Ireland. Given this comparatively strong structure in spite of its small geographic size, it is highly likely that factors beyond simple isolation by distance have led to the population structure observed

in the Netherlands including regional differences in gene flow from neighbouring countries and internal geographic boundaries.

While coarse geographical trends in Dutch genetic structure have previously been described using single-marker PCA (Abdellaoui, Hottenga, de Knijff, et al. 2013), our use of shared haplotypes reveals structure at a much higher resolution, differentiating subpopulations between, and sometimes within, provinces (Figure 4.1 and Figure 4.4). As a striking example, individuals from the east and west of North Brabant (NBE and NBW in Figure 4.1) are mutually genetically distinguishable and are more distinct from clusters to their north than Limburg clusters are, despite being geographically closer to these northern clusters. This deviation from haplotype sharing mirroring geography appears to be driven by strong genetic affinity to Belgium (Figure 4.3), reflecting a long history of demographic and political overlap across a 100 km frontier spanning the recently formed modern Dutch-Belgian border (formed circa 1840). In contrast, the majority of ancestral influence in Limburg, which also shares a substantial border with Belgium, is from Germany to the east, with lower contributions from Belgium (Figure 4.3; Appendix Figure 4.8). Notably, the Belgian border with the south of Dutch Limburg is almost entirely described by the course of the Meuse, which may have acted as a historical impediment to migration, thus distinguishing individuals in this region genetically. This is partially reflected in IBD clustering, in particular the distinction of southern Limburgish individuals from the rest of the Netherlands in short (1-3cM) segments, which otherwise only describe coarse north-central-south structure, with cluster boundaries roughly corresponding to the course of the Meuse (Figure 4.4). Future work explicitly modelling Dutch-Belgian and Dutch-German frontiers using additional Belgian and German genetic data with associated geography will resolve the historical and present-day role of the Meuse in distinguishing distinct population clusters in the south of the Netherlands.

Similarly to North Brabant, groups of individuals in North and South Holland show significant genetic separation despite mutual geographic proximity. While we have chosen to group the four South Holland clusters (SHOL cluster group) for visual brevity in Figure 4.1, they are robustly distinct by TVD permutation analysis (Figure 4.1;  $p < 0.001$ ), indicating that significant population differentiation exists even within South Holland. Migration and admixture in the highly urbanised *Randstad* has been proposed as a driver of genetic diversity and loss of geographic structure in this region (Abdellaoui, Hottenga, de Knijff, et al. 2013) which is supported by the overlaid geographical distribution of regional ancestry profiles (Figure 4.3) in this area. Previous studies have highlighted the correlation between decreasing autozygosity and increased urbanisation or population

density (Nalls et al. 2009), suggesting that urban centers are more outbred and diverse. In spite of this, the geographical ranges of the four South Holland clusters are somewhat independent (Appendix Figure 4.9), indicating that some degree of correlation between geographic and genetic structure has survived this urbanisation. Future work leveraging our length-binned IBD and Gaussian mixture model-based clustering approach on densely sampled data from this highly urbanised region may more explicitly delineate the interplay between urbanisation and population structure over time. To this end, highly urbanised areas such as the *Randstad* will be particularly informative. The inflated degree of diversity in populous areas like the *Randstad* is likely to reduce rates of recessive or rare variant mediated diseases in these areas through reduction of autozygosity (Nalls et al. 2009), with potential implications for study of disease in urban centres.

The principal fineSTRUCTURE split in the Netherlands describes north-south genetic differentiation (Figure 4.1) that is strong and persistent over time (Figure 4.4). We hypothesised that this reflects partially independent demographic histories so we estimated ancestral  $N_e$  for northern (NNL) and southern (SNL) Dutch fineSTRUCTURE populations, revealing superexponential growth in both populations with a sudden increase in rate during the 17<sup>th</sup> century (Figure 4.5 a). Historical  $N_e$  follows the same approximate trajectory for both populations but is consistently lower for the northern cluster, corroborating previous observations of increased homozygosity in northern Dutch populations (Abdellaoui, Hottenga, de Knijff, et al. 2013) and consistent with a model of northerners representing a founder isolate from southerners (although a more complex demographic model may better explain these observations) (Abdellaoui, Hottenga, de Knijff, et al. 2013; Genome of the Netherlands Consortium 2014). At first glance the apparent absence of  $N_e$  decline in 14<sup>th</sup>-century in the full Dutch dataset hints at the possibility that the Black Death had a weaker impact in the region than elsewhere in Europe (e.g. France which shows a decline in  $N_e$  around this time (Pierre et al. 2020)). Although this hypothesis agrees with the views of some historians, it is hotly contested by others (Roosen and Curtis 2019). However, regional  $N_e$  estimates display a prominent dip following the Black Death in the majority of provinces (Figure 4.5 b), suggesting that the Black Death has a detectable impact on the effective population size when measured in individual regions of the Netherlands. It is thus possible that merging non-randomly mating subpopulations (e.g. our genetic clusters) into a countrywide group (Figure 4.5 a) may have artificially inflated diversity, thus smoothing over the signal of population crash following the Black Death. This is because estimation of effective population size with IBDNe assumes random mating (S. R. Browning et al. 2018), meaning deviations from this assumption caused by population structure in our countrywide dataset may upwardly

bias  $N_e$  estimates. Conversely, defining overly homogeneous groups for estimation of  $N_e$ , such as genetic clusters identified by fineSTRUCTURE, might be expected to downwardly bias  $N_e$  estimates, hence our decision to define “regional” groups based on province of sampling here. Population structure is thus likely important when estimating  $N_e$  and trends countrywide and in NNL and SNL clusters (Figure 4.5 a) should be interpreted carefully. In spite of these concerns, the rate of exponential growth in countrywide  $N_e$  (Figure 4.5 a) is marginally shallower in the 10 generations following the Black Death (0.017; 95% c.i. 0.016-0.018) compared to the 10 generations prior (0.024; 95% c.i. 0.0235-0.0251), indicating some evidence of impact on the overall Dutch population prior to its recovery in the 17<sup>th</sup> century.

Previous works have hinted that north-south genetic differentiation in the Netherlands may have been facilitated by cultural division between the predominantly Catholic south and the Protestant north (Abdellaoui, Hottenga, de Knijff, et al. 2013). Given that the north-south structure observed in 1-3 cM IBD bins (expected time depth ~700 BCE) greatly precedes different forms of Christianity (Figure 4.4), our data support an alternative model in which the Protestant Reformation of the 16<sup>th</sup> and 17<sup>th</sup> centuries exploited pre-existing demographic subdivisions, leading to correlation between population structure and this cultural divide which has potentially been further strengthened by assortative mating among religious groups (Abdellaoui, Hottenga, Xiao, et al. 2013). Geographical modelling supports the role of migrational boundaries in establishing and maintaining this population substructure, especially rivers (Figure 4.6). A substantial belt of low inferred migration runs across the Netherlands, corresponding closely to the roughly parallel east-west courses of the Lower Rhine, Waal and Meuse rivers and correlating with the geographical boundary of the principal north-south fineSTRUCTURE split. Absolute assignment of causality to these geographical correlates is, however, not possible and, given the dense network of waterways in the Netherlands, could be misleading. For example, a strong migrational cold spot in the east of the Netherlands runs parallel to the IJssel (Figure 4.6), but could potentially be better explained by the course of the Apeldoorn Canal, a politically fraught waterway constructed in the early 19<sup>th</sup> Century. Similarly, a cold spot in the northwest directly overlays the North Sea Canal (completed in 1876). As both of these are human-made waterways, it is not certain whether their courses are consequences or determinants of low movement of people across their paths.

As well as internal geography, outside populations have also likely played an important role in the establishment and maintenance of population structure in the Netherlands (Figure 4.3; Table 4.1), with opposing gradients of Belgian-like and German-like ancestry overlaying the principal axes of structure in the country (Figure 4.3), and evidence of an

significant admixture event dating to 1088CE (1004-1111 CE) (Table 4.1). However due to the extent and variety of demographic upheaval and mobility of European populations over history, and regular changes in the political boundaries, interpretation of both our ancestry gradients and GLOBETROTTER admixture dates is complicated. For example Belgium only separated politically from the Netherlands in the past 180 years (~6 generations) suggesting that the high levels of ancestry matching a modern Belgian source seen in southern provinces may simply reflect mixing within the then unified political region as opposed to subsequent geneflow. Similarly ascribing a single historical source to the countrywide admixture in the 11th century is difficult. However there is one source of ancestry that appears to have a very clear origin historically, namely the small but significant contribution of Danish haplotypes in the north and west of the Netherlands, a possible vestige of Viking raids in coastal areas in the 9<sup>th</sup> and 10<sup>th</sup> centuries. This is corroborated by an inferred GLOBETROTTER single-date admixture event in the NHFG (North Holland, Friesland and Groningen) cluster (Figure 4.1) between 759 and 1290 CE with Danish haplotypes as a major admixing source (Table 4.1), which differs from the German and Belgian sources seen in other clusters. The extent of legacy left by more than a century of Danish Viking raids and settlement in the Netherlands has been the subject of some debate. However, from our data, it appears that the modern Dutch genome has indeed been partially shaped by historical Viking admixture. This Danish Viking contact is contemporaneous with a critical period in the establishment of the modern Dutch genome from other outside sources (1004-1111 CE; Table 4.1), although the precise historical correlates of the admixture events detected in the remaining Dutch regions are less obvious.

Future densely sampled ancient DNA datasets from informative time depths in the Netherlands and northwest Europe will enable direct estimation of ancestral population structure, admixture, demographic affinities and effective population sizes, improving precision over the current study which depends on proxy patterns of haplotype sharing between modern individuals. Similarly, regional ancestry and admixture inference are limited by the use of modern proxy populations in place of true ancestral sources; nevertheless, there are ample advantages to the use of modern data, including large sample size and relevance to research on modern human health and disease. In particular, as in our previous work in Ireland (R. P. Byrne et al. 2018) (Chapter 3), samples in the current Dutch dataset were not specifically selected to have pure ancestry in each geographical area (e.g. all grandparents from the same region (Leslie et al. 2015)) meaning the degree of structure observed is not idealised or exaggerated by sampling,

but instead representative of the structure expected in any GWAS that includes Dutch data. We will further explore the impact of the fine-scale genetic structure described in this study on GWAS statistics in chapter 5, where we will reflect on correcting GWAS performed on a larger Dutch dataset and a multi-population dataset using haplotype sharing profiles between samples.

# Chapter 5 - Detecting and Correcting Confounding in GWAS Using Haplotype Sharing Methods.

## 5.1 - Introduction

### 5.1.1 - Background

A history of non-random mating across human populations has led to detectable systematic differences in allele frequencies across populations in a phenomenon commonly referred to as population structure. These allele frequency differences can lead to statistical inflation and spurious associations in genome wide association studies (GWAS) where the phenotype being studied is unevenly distributed or stratified across these populations. Notable early examples of false genetic associations driven by population stratification include the apparent genetic associations between an immunoglobulin Gm haplotype (Gm3;5,13,14) with type two diabetes in Native Americans (Knowler et al. 1988) and an apparent association of a *CYP3A4-V* promoter variant with prostate cancer in African Americans (Kittles et al. 2002), both of which were due to confounding from admixture in their studied groups. A wide range of techniques have been developed to identify and address this problem, including genomic control (Devlin and Roeder 1999; Devlin, Roeder, and Wasserman 2001), structured association (Pritchard et al. 2000; Devlin, Roeder, and Wasserman 2001), principal component analysis (Price et al. 2006) and linear mixed models (Yang et al. 2011), which have greatly reduced the incidence of false positives in GWAS and controlled the level of genome wide statistical inflation caused by population structure. This is evidenced by linkage disequilibrium score regression results (LD score regression), which show that statistical inflation in many (but not all (Bhatia et al. 2016)) complex traits mostly results from polygenicity rather than confounding (Bulik-Sullivan, Loh, et al. 2015).

However, emerging evidence suggests that residual effects of stratification can persist in GWAS in spite of these corrections; both single markers and polygenic scores for a number of traits show strong association with birth location in the UK Biobank even following correction for population stratification with 40 PCs (Haworth et al. 2019) suggesting that PCs cannot fully account for population stratification. It has been hypothesised that the residual correlations between polygenic scores and geography result from recent internal migrations in the UK driven by socioeconomic status (Abdellaoui et al. 2019). Emerging work supports this framework of residual structure resulting from recent migration, showing that principal components from common variants cannot fully account for or adjust for structure from recent demographic events (Zaidi and

Mathieson 2020). Residual confounding appears to have particularly strong impact on methods that rely on large numbers of effect size estimates from variants below genome wide significance in a studied trait (Lawson et al. 2019), as recently seen for estimates of polygenic signals of adaptation in height which were overestimated due to the residual effects of population stratification the source GWAS (Berg et al. 2019; Sohail et al. 2019). Similar biases have been proposed for other methods such as polygenic risk scores (Kerminen et al. 2019), and genome-wide SNP heritability estimates (Dandine-Roulland et al. 2016) in the presence of uncorrected population structure. These observations all motivate the search for novel methods to correct for population stratification that can account for both recent and long-standing structure in the data and improve effect estimates to limit the biases seen in polygenic scores and other methods leveraging genome-wide effect estimates for complex traits.

As discussed in previous chapters (Chapters 3 and 4), matrices recording patterns of haplotype sharing between samples such as the ChromoPainter coancestry matrix (Lawson et al. 2012) and IBD sharing matrices (Lawson and Falush 2012) increase the resolution of population structure detected in genotype datasets compared to approaches based on unlinked markers (e.g. the covariance matrix used in SNP PCA). So far the ChromoPainter method has largely been applied to detecting and describing structure in single country populations (Leslie et al. 2015; R. P. Byrne et al. 2018; Takeuchi et al. 2017; Kerminen et al. 2017; Raveane et al. 2019; Gilbert et al. 2017, 2019; Bycroft et al. 2019; Pierre et al. 2020), however it may be well suited to the application of correcting residual confounding in GWAS. As stated in the methods paper for ChromoPainter, principal components from the normalised ChromoPainter coancestry matrix are a “natural extension” of the standard unlinked SNP PCA method proposed by Price et al. (Price et al. 2006) for cases where we have information on linkage relationships between densely genotyped markers (Lawson et al. 2012). This means ChromoPainter PCs are analogous to the EIGENSTRAT PCs (Price et al. 2006) commonly used to control for confounding in GWAS. The authors additionally show that in models where recombination rates approach infinity, PCs of the ChromoPainter coancestry matrix reduce approximately to those from the EIGENSTRAT method, suggesting standard SNP PCA is a “special case” of their method (Lawson et al. 2012). This indicates that ChromoPainter PCs should represent ancestry between samples more accurately in human populations, where recombination rates are much lower than this (International HapMap Consortium 2005). It follows that PCs from the coancestry matrix may be suitable for correction of population structure in GWAS, given that they too are orthogonal continuous proxies for sample ancestry, affording them many of the benefits which motivated the initial choice of standard SNP



PCA (Price et al. 2006). In addition as the ChromoPainter coancestry matrix simultaneously describes sharing patterns for both old (short) and recent (long) ancestry segments, ChromoPainter PCs likely capture patterns of both recent and older population structure. Hence ChromoPainter PCs may better address the residual confounding in GWAS datasets from recent demographic events (Abdellaoui et al. 2019; Zaidi and Mathieson 2020) that likely shapes biases in polygenic scores, while adequately controlling for structure caused by older events.

Application of ChromoPainter to large datasets is currently hindered by computational cost, which scales quadratically with sample size, making it impractical for use in large GWAS. Fortunately a fast approximate chromosome painting method has been developed by Richard Durbin and Daniel Lawson (PBWT-paint; <https://github.com/richarddurbin/pbwt>; -paint switch) which exploits the quick haplotype matching of the Positional Burrows-Wheeler transform (Durbin 2014) to increase the speed and efficiency of the ChromoPainter approach, and is capable of handling hundreds of thousands of haplotypes. This method may provide a scalable alternative to ChromoPainter in a GWAS setting.

In addition, downstream analysis of ChromoPainter output such as the fineSTRUCTURE clustering algorithm (Lawson et al. 2012), which identifies homogeneous sub-population groupings based on patterns of haplotype sharing within a dataset, is also much too computationally costly for large datasets. This limits finescale studies of population structure using these methods to modestly sized datasets ( $n < 10,000$ ), restricting analysis of local population structure in large multi-population datasets. Community detection algorithms which rapidly search for densely connected groups in large networks (e.g. social networks) may be a suitable alternative for finding homogeneous clusters at larger scales. One such method known as the Louvain community detection method (Blondel et al. 2008) has been successfully applied to identify genetic clusters in a huge IBD sharing dataset from the US ( $n = 770,000$ ) (Han et al. 2017) and is readily adaptable to this application. Hence analogous methods to those used in Chapters 3 and 4 are close to coming of age for use in large scale datasets, with potential applications in both describing and correcting for population structure in modern GWAS datasets.

In this chapter we explore the application of ChromoPainter PCs and PBWT-paint PCs to detect and correct residual population structure in a small Dutch ( $n = 4,753$ ) and large multi-population ( $n = 36,052$ ) ALS GWAS dataset, comparing their performance with standard SNP PCA covariates. ChromoPainter PCs appear to reduce inflation due to confounding better than SNP PCs (measured by LD-score regression) when applied to a

GWAS conducted on a small ALS case control dataset (n=4,753) from the Netherlands, suggesting they can better correct for confounding caused by local structure in the Netherlands (see Chapter 4). Scaling to the full GWAS dataset, the approximate chromosome painting method PBWT-paint ran magnitudes faster than ChromoPainter while producing highly correlated results when tested on smaller datasets from Chapters 3 and 4 suggesting it is a suitable alternative for larger datasets. When applied to the full 2016 GWAS dataset (van Rheenen et al. 2016) (n=36,052) t-SNE of PBWT-paint PCs shows much clearer separation of sampling regions than t-SNE of SNP PCs, indicating the method captures additional structure missed by SNP PCA. We found that the Louvain Community detection method could identify sensible subpopulation groupings in this large haplotype sharing dataset, making it a viable alternative to fineSTRUCTURE for descriptive population groupings at this scale. Additionally the Louvain method identified splits missed by fineSTRUCTURE in the POBI dataset (Leslie et al. 2015) from Chapter 3, motivating application to smaller datasets. Finally, PBWT-paint corrected GWAS summary statistics calculated from the 2016 ALS GWAS dataset (van Rheenen et al. 2016) showed significantly lower inflation than SNP-corrected summary statistics while retaining the power to detect known ALS hits, implying the method is more stringent but doesn't limit power excessively. Polygenic risk scores calculated from these PBWT-paint PC corrected GWAS summary statistics showed different distributions to those generated from SNP PC corrected GWAS, indicating the correction method affects not only association signal but distribution of effect sizes. These scores predicted the phenotype less accurately than SNP PC corrected scores, but also had lower signals of residual confounding from population structure. The loss of prediction accuracy is thus likely partially explained by the removal of residual confounding meaning scores generated using this approach might be more meaningful due to lower bias.

### 5.1.2 - Research aims

This chapter presents work carried out on individual-level genotype data from the 2016 ALS GWAS (van Rheenen et al. 2016) with the aim of exploring novel methods for detecting and correcting latent population structure in large GWAS. In particular we will explore the impact of using principal components (PCs) derived from haplotype sharing matrices in place of standard SNP PCs, while also considering the scalable clustering methods in place of fineSTRUCTURE. We divide this global aim into three component aims:

- i.) To explore the application of ChromoPainter PCs to reduce inflation in a preliminary GWAS conducted in a single population with known local structure (The Netherlands).

ii.) To explore the scalability and accuracy of the PBWT-paint haplotype sharing method and the Louvain community detection method for detecting subtle population structure in a large GWAS dataset (2016 ALS GWAS; n=36,052).

iii.) To explore the application of PBWT-paint PCs to reduce inflation in the full 2016 ALS dataset, and assess the impact of this correction on polygenic scores.

Through this research we hope to potentially provide a remedy for confounding from population structure not addressed by current methods using unlinked markers such as SNP PCA, with potential benefits for polygenic methods such as PRS and unbiased heritability estimation.

NB: Several results from this chapter feature in an article now published in Nature Communications: (<https://www.nature.com/articles/s41467-020-18418-4>).

## 5.2 - Methods

### 5.2.1 - Datasets and initial quality control

Analyses in this chapter were carried out on two main datasets derived from the 2016 ALS GWAS (van Rheenen et al. 2016) (see Appendix Table 5.1 for sample breakdown). Firstly to explore the utility of haplotype sharing analysis in correcting confounding from finescale population structure in a single country GWAS context we analysed an ALS case/control dataset with samples from the Netherlands only ( $n = 4,753$ ; strata sNL1, sNL3 and sNL4 from the 2016 GWAS (van Rheenen et al. 2016)), excluding sNL2 due to severe case control imbalance in this stratum (case:control  $\sim 1:34$  in this stratum). Following this we then analysed the application of haplotype sharing analysis to correcting confounding in larger multi-population GWAS using the full 2016 ALS GWAS dataset (van Rheenen et al. 2016) ( $n = 36,052$ ) from which this Dutch subset was derived. In addition data from previous chapters (Chapters 3 and 4) were analysed following QC described in those chapters.

For our baseline haplotype sharing analyses we extracted 1,060,224 zero-missingness Hapmap phase 3 SNPs (International HapMap 3 Consortium et al. 2010) that passed QC in the source paper for the dataset (van Rheenen et al. 2016) for both the Dutch and multi-population dataset. This SNP set was chosen as Hapmap3 SNPs are typically well imputed and lack A/T and C/G SNPs that could cause phasing errors. These SNPs also produce unbiased estimates when used in LD score regression. Reducing our dataset to these SNPs also lowered computational costs associated with ChromoPainter. Related individuals ( $\hat{\pi} > 0.1$ ) and SNPs with greater than zero missingness (`--geno 0`) were removed using plink v1.9 (Chang et al. 2015) as cryptic relatedness and missingness confound both ChromoPainter and GWAS analyses. In addition following an initial round of painting with PBWT-paint (see below) extreme haplotype PCA outliers in the multi-population dataset ( $>20$  SD from mean on haplotype PC1-10) were removed, followed by repainting as an additional QC step. Final Dutch and multi-population datasets contained 4,753 samples and 35,985 (12,510 cases and 23,475 controls) samples respectively.

Finally for GWAS analyses carried out on the data described below we scaled back up to all variants passing QC in the source 2016 GWAS which had a mean imputation score greater than 0.9 across all strata (6,767,915 SNPs).

### 5.2.2 - Phasing

Samples from each dataset were phased by running SHAPEIT v2 (Delaneau, Marchini, and Zagury 2011) using the 1000 Genomes Project phase 3 samples as a reference panel and the 1000 Genomes phase 3 genetic map (Auton et al. 2015). Allele concordance with the reference panel was checked prior to phasing (--check) in order to remove variants not aligned properly to the reference panel. Samples per dataset were phased together one chromosome at a time.

### 5.2.3 - Haplotype painting analysis

We estimated haplotypic sharing profiles between individuals in our i.) Dutch and ii.) multi-population datasets separately as follows.

#### i.) Dutch dataset:

We ran ChromoPainter v2 on the phased Dutch dataset to paint all individuals in terms of one another (--a 0 0). For this analysis we first ran 10 EM iterations on chromosomes 1, 8, 15 and 20, using 10% of samples (chosen at random) to estimate model parameters  $N_e$  and  $\mu$ . We then took the weighted average of these parameter estimates across chromosomes for use in a final run with all samples.  $N_e$  and  $\mu$  were compared to estimates from the Dutch dataset in chapter 4 to evaluate if downsampling severely affected parameter estimation and was found to be consistent.

#### ii.) Multi-population dataset:

As ChromoPainter is computationally intractable for more than ~10,000 samples, we instead ran PBWT-paint (<https://github.com/richarddurbin/pbwt>; -paint switch) on our phased multi-population dataset (n=36,052) using the default settings. PBWT-paint is a fast approximate implementation of ChromoPainter suitable for large datasets.

For each run we combined (summed) all 22 per-chromosome haplotype sharing matrices to create a genome-wide coancestry matrix for use in GWAS correction. We additionally constructed 22 matrices leaving one chromosome out (LOCO; one for each chromosome) to describe population structure on all but the target chromosome, allowing us explore the possibility of over-correcting in GWAS which may occur if our haplotype sharing profile describes variation in a target SNP or disease haplotype which is not simply due to background population structure.

#### 5.2.4 - Principal component analysis and t-SNE

For a number of analyses in this chapter we compare principal components calculated from unlinked genotypes (SNP PCs) to principal components from a haplotype sharing matrix calculated using ChromoPainter or PBWT-paint. SNP PCs were calculated in each dataset using PLINKv1.9 (--pca), first removing long range LD regions (Price, Weale, et al. 2008) ([https://genome.sph.umich.edu/wiki/Regions\\_of\\_high\\_linkage\\_disequilibrium\\_\(LD\)](https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD))) and pruning for LD (--indep-pairwise 500 50 0.8). Haplotype PCs were calculated using the fineSTRUCTURE R tools (<https://www.paintmychromosomes.com>) on the normalised coancestry matrices for each dataset.

Additionally SNP and haplotype PCs were used to initialise t-SNE (t-distributed stochastic neighbour embedding) analysis in our multi-population dataset in order to compare the resolution of structure observed. Samples from this dataset were projected onto a 2-dimensional embedding using Rtsne (<https://github.com/jkrijthe/Rtsne>) to describe both global and local structure. For this analysis we ran Rtsne for 5,000 iterations using the top 100 PCs from each method (haplotypic and unlinked markers) as a starting point using perplexity of 30.

#### 5.2.5 - Benchmarking PBWT-paint against ChromoPainter

To evaluate the i.) accuracy and ii.) computational cost of PBWT-paint, we benchmarked it against ChromoPainter as follows:

##### i.) Accuracy:

We compared PBWT-paint and chromopainter outputs for the Irish and Dutch datasets from Chapter 3 and Chapter 4 for accuracy. PBWT-paint was run with standard settings on the phased Dutch and Irish datasets and compared to ChromoPainter output described previously in Chapters 3 and 4. We regressed PCs 1 and 2 of the coancestry matrices from each method (ChromoPainter vs PBWT-paint) against each other to evaluate if the two methods were picking up similar major trends in finescale ancestral structure described in previous chapters. Next we calculated Pearson's  $\rho$  between the coancestry matrices from the two methods to evaluate the overall pairwise correlation across the full matrixes. As the ChromoPainter model is the gold standard here, high correlation across these metrics was deemed as high accuracy.

##### ii.) Efficiency:

The comparative runtime of ChromoPainter and PBWT-paint were estimated using a

single chromosome (chromosome 20) for datasets of different sizes. Analyses were run on a single core on chromosome 20 of three datasets a.) Dutch single stratum dataset (n=1,626; Chapter 4), b.) Dutch multi stratum dataset (n=4,753) and c.) the multi-population dataset (n=36,052).

Estimates of runtime for all PBWT-paint runs were taken using the bash time command using all samples for each dataset, while runtime estimates for ChromoPainter were calculated by painting a single individual in terms of all others in the sample and multiplying by the sample size to reduce computational cost. While there may be slight variation in runtime speed across samples for ChromoPainter and hence taking the runtime for all samples would be more realistic than extrapolating from one sample each time, this estimate is purely to give a scale of runtime for comparison to PBWT-paint and hence does not need to be exact.

As ChromoPainter runtime scales roughly quadratically with sample size (see <https://people.maths.bris.ac.uk/~madjl/finestructure/manualse8.html>), and thus runs are close to intractable for datasets over ~10,000, estimates of runtime for ChromoPainter for the largest dataset (n=36,052) were extrapolated from the runs on smaller datasets using the formula:

$$E[\text{runtime}(N_2)] = \text{runtime}(N_1)/N_1^2 \times N_2^2, \quad (8)$$

where  $N_1$  is the sample size of the dataset we are extrapolating the runtime from and  $N_2$  is the size of the dataset we are extrapolating to.

### 5.2.6 - GWAS analyses

#### Haplotype sharing PCA vs SNP PCA covariates:

We ran GWAS on both the single population and multi-population datasets using the logistic function in Plinkv1.9 (--logistic), fitting independent marker PCs (SNP PCA) and haplotype sharing PCs (ChromoPainter PCs or PBWT-paint PCs) as covariates separately. For these analyses we fitted 10, 20, 30 and 40 PCs to test how the number of covariates included affects the inflation of each method (Appendix Figure 5.1; see LD score regression for further details). We included all SNPs passing QC in the source paper (van Rheenen et al. 2016) with imputation info scores greater than 0.9 for each dataset in this analysis. Statistics of inflation were calculated for these GWAS using LD score regression as described below (Methods section 5.2.8).

In addition to fitting PCs from the full genome wide haplotype sharing matrix to correct for population structure, we adopted a leave one chromosome out (LOCO) analysis approach analogous to those used in many LMM softwares (e.g. GCTA-LOCO (Yang et al. 2014, 2011)). In this analysis logistic regression GWAS was run on each chromosome separately using PCs calculated on haplotype sharing matrices constructed from all chromosomes excluding the chromosome being tested. In doing this we hoped to avoid capturing disease-specific variation on haplotypes at the locus being tested for association in our covariates and over-correcting. This adjustment should help to prevent penalising real effects, while adequately describing population structure using the remaining chromosomes. This approach assumes that variation due to population structure does not massively vary across chromosomes, and hence should adequately be described by the remaining chromosomes included. We also tested inflation in this approach using LD score regression and compared results to the standard method using all chromosomes.

#### ALS GWAS excluding Dutch subset for PRS:

We excluded all samples in the Dutch dataset from the full 2016 GWAS dataset to create a training dataset ( $n = 31,299$ ; 10,606 cases and 20,693 controls) for testing for enrichment in ALS signal in Dutch GWAS and PRS (See 5.2.10) under various correction schemes (eg. PBWT-paint PCs as covariates vs SNP PCs as covariates). We fit two logistic regression GWAS on this dataset fitting 20 SNP PCs or 20 PBWT-paint PCs as covariates respectively.

#### 5.2.7 - Estimating variance explained in phenotype

In order to understand how well suited ChromoPainter PCs are to correcting population stratification we first compared how well these haplotype sharing PCs and standard SNP-based PCs predicted ALS case/control status in our two datasets. The rationale here is that the ancestral structure described by PCs should only predict phenotype in cases where it is subtly stratified across population structure in our sample (assuming no ancestry by disease interaction), hence the variance in phenotype explained by each method should scale with the level of stratification captured by each method.

To quantify the full variance explained by each method, we ran two separate logistic regression models per dataset, fitting 100 PCs from either the haplotype sharing or the SNP-based method. We estimated the cumulative variance in phenotype explained by successive PCs from each method using logistic regression in R (`glm()` function), evaluating the goodness of fit of these models using Nagelkerke  $R^2$  calculated with the `fmsb` package (Nakazawa 2018). Curves of this cumulative phenotypic variance explained



were plotted to inform an appropriate number of PCs appropriate for correction of GWAS based on where the slope begins to plateau (~20 for both the multi-population and Dutch only GWAS).

#### 5.2.8 - Estimating confounding with LD score regression

Summary statistics from GWAS run using SNP PCs and haplotype sharing PCs as covariates were formatted for use with the LD score regression software using the “munge\_sumstats.py” script provided with the software, retaining only SNPs present in the HapMap phase 3 dataset, which are typically well imputed. We ran univariate LD score regression in each GWAS using precomputed LD scores for European individuals: ([https://data.broadinstitute.org/alkesgroup/LDSCORE/eur\\_w\\_ld\\_chr.tar.bz2](https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2)). The LD score intercept was used to compare confounding between GWAS runs using different corrections.

To test whether the differences in LD-score intercept were significant for each pair of methods in the same dataset we calculated Z scores as follows:

$$Z_{diff-intercept} = \frac{(intercept_1 - 1) - (intercept_2 - 1)}{\sqrt{se(intercept_1)^2 + se(intercept_2)^2}} \sim N(0,1) \quad (9)$$

Here we subtract 1 from each intercept as this is the expectation of the intercept when there is no confounding present (See Chapter 1, Equation 1).

#### 5.2.9 - Clustering haplotype sharing datasets using the Louvain method for community detection

We explored the iterative application of the Louvain method for community detection (Blondel et al. 2008) as an alternative clustering method to fineSTRUCTURE for use in large haplotype co-ancestry matrices. We first tested this method in the relatively small Dutch dataset from Chapter 4, and in the POBI dataset from Chapter 3 to assess whether it captured similar structure to fineSTRUCTURE, and then applied it to the larger multi-population dataset.

Haplotype sharing networks were constructed between individuals (nodes) from the respective coancestry matrices using the “graph\_from\_adjacency\_matrix” function in the R igraph package (Csardi, Nepusz, and Others 2006). Here the edges between individuals were weighted based on the number of chunks individuals shared in the

coancestry matrix. Next clusters were identified using the “cluster\_louvain” function in igraph which greedily assigns samples to communities maximising a measure known as modularity, which compares the density of links within a community to those between communities. The modularity value can take any value from -0.5 (no modularity) to 1 (full modularity), and changes in the modularity are used to inform moves of individuals to different communities. Hence this method can identify structured groups of individuals where individuals share more haplotypes within the group than with individuals from other groups. Once the algorithm converged on a solution for the full dataset (iteration one), we took the cluster assignments and constructed new independent networks based on sharing within these clusters, repeating the process (iteration two to n). In this way the algorithm can be applied iteratively to its own output to reveal increasingly granular structure in the datasets until no structure is present in remaining groups. For our analysis we ran at most 3 iterations of this method, ensuring that modularity remained significantly greater than 0 for each iteration to avoid over splitting.

#### 5.2.10 - Polygenic risk scores and residual confounding

We calculated ALS polygenic risk scores (PRS) in our Dutch only dataset using GWAS summary statistics calculated with all other samples from the 2016 GWAS (n=31,299, Dutch excluded) corrected using either 20 SNP PCs or 20 PBWT-paint PCs as covariates. PRS were computed using PRSice v2.1.2 (Euesden, Lewis, and O’Reilly 2015) with default settings, setting the prevalence of ALS to 1 in 400 (Johnston et al. 2006), and using 20 SNP PCs as covariates in the model to correct for structure in the target population. We evaluated the impact of using haplotype sharing PCs relative to SNP PCs in the training population by comparing the resulting geographic distributions of PRS in the Netherlands after adjusting for phenotype (i.e residual population stratification) and variance explained by the models (model performance) under various correction schemes.

For analysis of residual population stratification in PRS we measured geographic autocorrelation of scores using Moran’s I. We first regressed the phenotype out of PRS scores for our samples to avoid measuring correlation with geography caused by geographic distribution of the phenotype in our samples as this would not represent confounding from population stratification. We then ran Moran’s I on the mean residuals of this regression for each unique geographic coordinate in our data using the “gearymoran” function in the ADE4 R package setting “nrepet” to 10,000 and using a two-sided test for significance. Moran’s I can range between 1 and -1, with values of 1 representing complete spatial clustering of a variable (i.e. similar values are adjacent spatially), values

of 0 representing spatial randomness and values of -1 representing complete spatial dispersal (i.e. distinct values are adjacent spatially). A Bonferroni adjusted significance threshold of 0.0041 (0.05/12) was used to assess significance in these analyses.

#### 5.2.11 - Stringent heritability estimation using haplotype sharing covariates

We constructed a GRM using HapMap3 SNPs for the full 2016 ALS dataset using GCTA (Yang et al. 2011; S. H. Lee et al. 2011) to explore the impact of using haplotype sharing PCs as covariates in GREML analysis. We ran GREML (S. H. Lee et al. 2011) on the resulting GRM, fitting either 20 SNP PCs or 20 PBWT-paint PCs as covariates. We also split the data into minor allele frequency bins (MAF; 0.01-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and 0.4-0.5) and by chromosome and estimated heritability per allele frequency bin and chromosome per method to compare the effects of the two population structure correction methods.

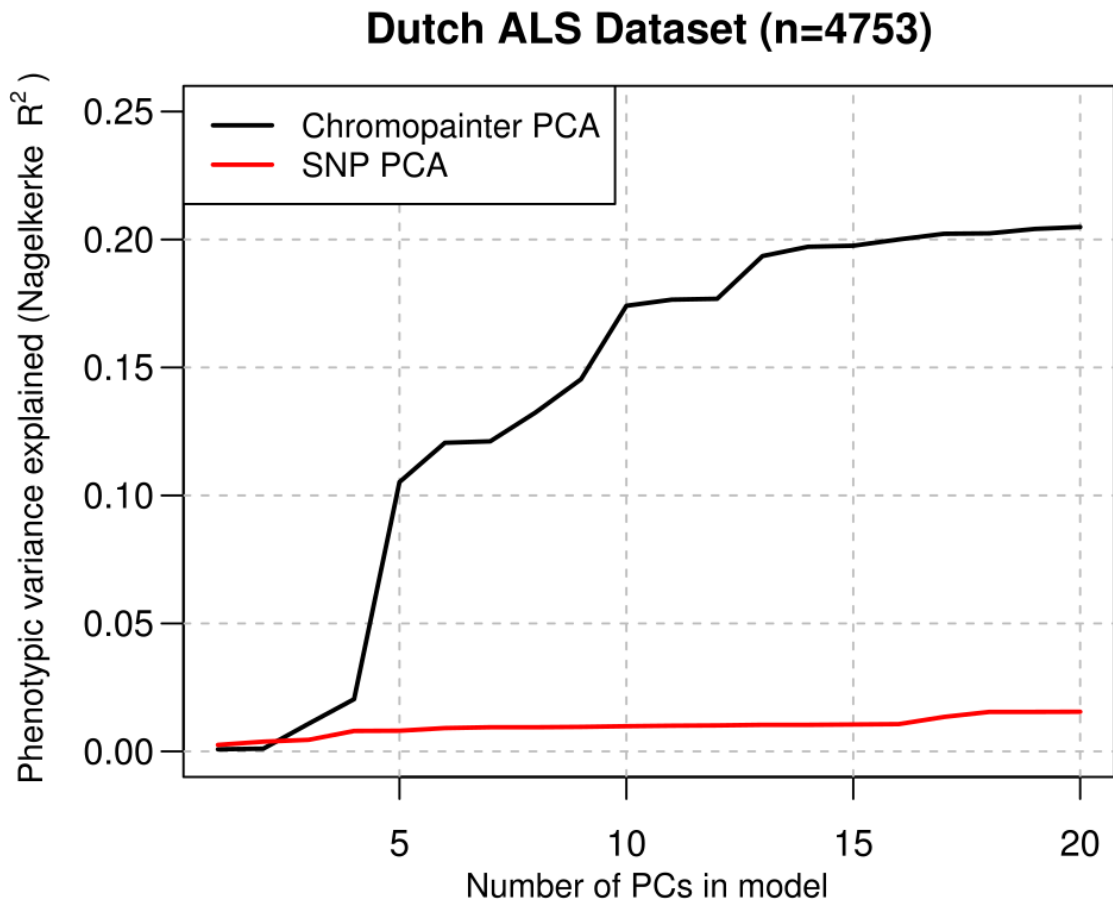
To compare heritability estimates between methods for full datasets or partitions we calculated a z-score for the difference in estimates using the following equation :

$$Z_{diff-heritability} = \frac{h_{method1}^2 - h_{method2}^2}{\sqrt{se(h_{method1}^2)^2 + se(h_{method2}^2)^2}} \sim N(0,1) \quad (10)$$

## 5.3 - Results

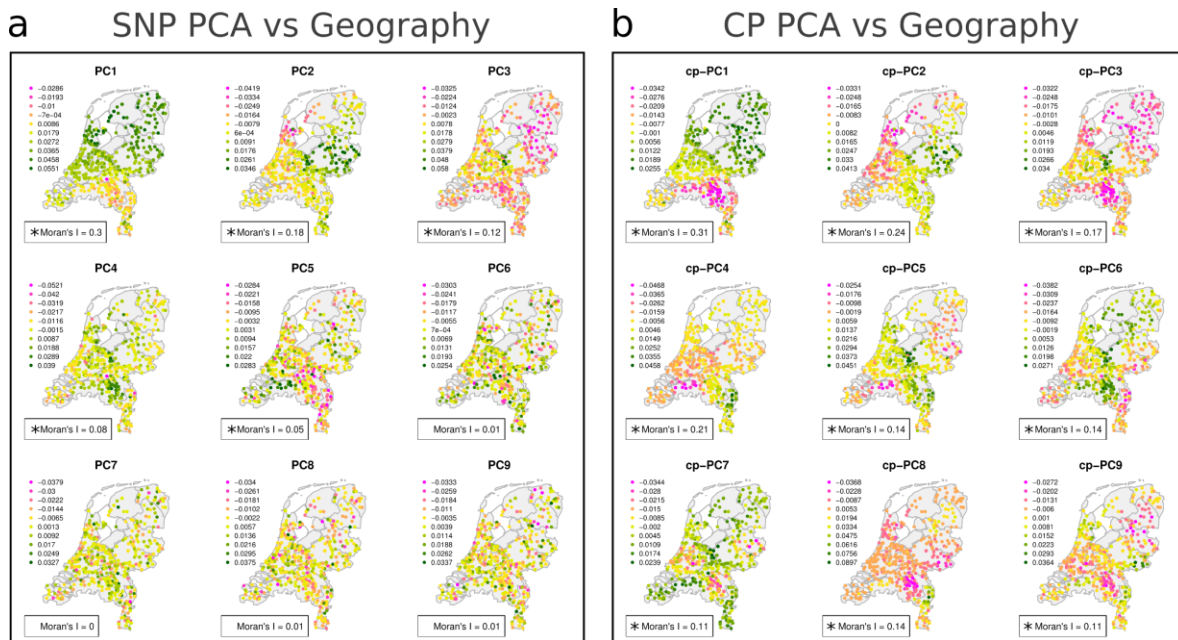
### 5.3.1 - Correcting GWAS confounding in a single population dataset (Dutch)

In order to test the impact of the finescale structure identified in the Dutch population in Chapter 4 on GWAS association statistics we explored the effectiveness of using ChromoPainter PCs (cp-PCs) as covariates in a larger Dutch case-control dataset (n=4753: 1971 cases; 2782 controls) composed of strata sNL1, sNL3 and sNL4 from the 2016 ALS GWAS (van Rheenen et al. 2016). Firstly, we assessed how well haplotype-based PCs and standard SNP-based PCs described the effects of stratification of ALS cases and controls over population structure using logistic regression of case-control status versus successive PCs from each method. Comparing the cumulative variance in phenotype explained by logistic regression of successive PCs from either method versus ALS case-control status (Figure 5.1), we observed that ChromoPainter PCs explain a large portion of the variance in case-control status (Nagelkerke's R-squared for 20 ChromoPainter PCs = 0.205), while SNP PCs explain very little variance (Nagelkerke's R-squared for 20 SNP PCs = 0.015; ~14 fold lower). The rate of increase in variance explained appears to plateau for both methods at ~20 PCs (with minor increases thereafter), suggesting that this is an appropriate number of covariates to include in a GWAS in this context. The significant increase in phenotypic variance explained by ChromoPainter PCs likely indicates that ChromoPainter PCs detect significant stratification of samples across local ancestry groups that is not captured by standard SNP PCs. Alternatively, it could mean that ChromoPainter PCs capture differential haplotypic sharing between ALS cases and controls due to certain haplotypes associating with the phenotype, however this is less likely given that early haplotypic PCs tend to describe geographically structured patterns of ancestry, making population structure a more parsimonious explanation (Figure 5.2; also see Chapter 3 and Chapter 4). In fact, while at least the first 20 ChromoPainter PCs show significant evidence of geographic clustering by Moran's I, only the first 5 SNP PCs geographically cluster (Appendix Table 5.2), indicating SNP PCs miss substantial patterns in local population structure.



**Figure 5.1: Phenotype stratification captured by ChromoPainter PCs and SNP PCA in Dutch dataset.**

Displayed are plots of the cumulative variance in phenotype (amyotrophic lateral sclerosis) explained by principal components (PCs) of the ChromoPainter coancestry matrix and standard SNP PCs in a Dutch only GWAS dataset (n=4,753). ChromoPainter PCs explain a greater amount of variance in phenotype and are hence expected to better correct for confounding due to stratification in GWAS. Phenotypic variance explained by ChromoPainter PCs increases rapidly, and then appears to plateau between 15-20 PCs, suggesting this is a suitable number of PCs to include as covariates to correct population stratification.



**Figure 5.2: Comparing the relationship of cp-PCA and SNP PCA to Dutch geography.**

Displayed are maps demonstrating the geographic distribution of a.) SNP PCs and b.) ChromoPainter PCs for our Dutch only dataset ( $n=4,753$ ). Points on the maps are coloured by the average PC value per town of sampling for a.) SNP PCA and b.) ChromoPainter PCA (cp-PCA) calculated using 4,753 Dutch samples (NB: Geography is only available for 1,352 samples). PCs have been split into 10 equally sized bins for visualisation purposes. Moran's I values are included for each map representing the degree of geographic clustering. ChromoPainter PCs show a stronger relationship with geography for a greater number of PCs than SNP PCA as denoted by significant positive values of Moran's I for these and further PCs (Appendix Table 5.2). Asterisks (\*) signify PCs are significantly clustered in geographic space by Moran's I, passing a Bonferroni corrected p-value threshold ( $p < 0.0025$ ). Exact values for these and further PCs are available in Appendix Table 5.2. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).

Next to test how well ChromoPainter corrects for inflation of GWAS summary statistics we performed a GWAS for ALS in the Dutch data adjusting for i.) 20 standard PCs; ii.) 20 ChromoPainter PCs or iii.) 20 ChromoPainter PCs in a leave one chromosome out (LOCO) style analysis, recording the mean chi-squared and lambda GC result for each regression to test for inflation (Table 5.1). We also calculated the LD-score intercept resulting from each correction (Table 5.1), which is a metric that can distinguish inflation resulting from confounding from inflation resulting from polygenicity (Bulik-Sullivan, Loh, et al. 2015) making it particularly useful in this context. Inclusion of the ChromoPainter PCs led to subtly lower mean chi-squared, lambda GC and LD-score intercept values, all of which approached 1, signifying deflation of the GWAS summary statistics compared to SNP PC corrected GWAS. Notably our ChromoPainter LOCO analysis reduced inflation to

practically the same level as our standard ChromoPainter analysis, suggesting that the observed effect on LD score intercepts is due to correction of population stratification rather than loss of signal due to haplotype sharing mirroring variation in disease-associated loci. This is because the LOCO analysis does not correct for haplotype sharing patterns at the locus being tested, instead using background sharing patterns to adjust for population stratification, hence it should not overcorrect at a given locus in theory. These results indicate that inclusion of ChromoPainter PCs as covariates lowers inflation in our GWAS without necessarily penalising disease-associated SNPs.

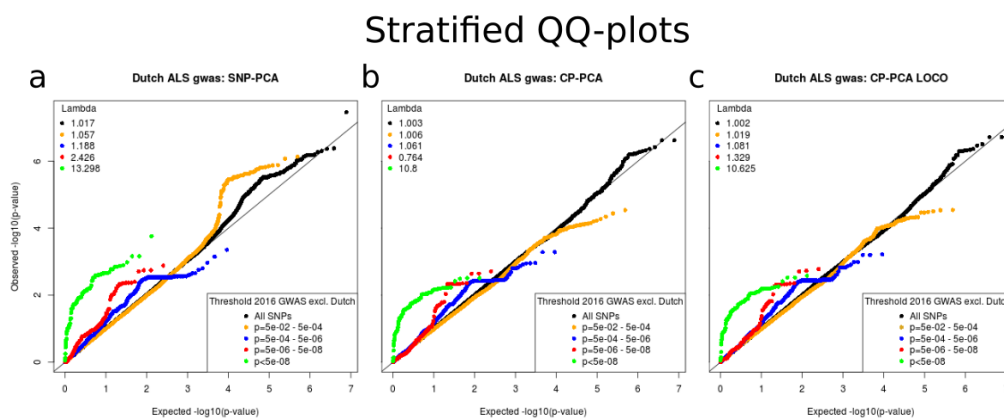
**Table 5.1: Measures of inflation in Dutch only GWAS of ALS with haplotypic (cp-PCA) and unlinked (SNP PCA) ancestry covariates on HapMap SNPs.**

Covariates included	LD score intercept	LD score intercept SE	Lambda GC	Mean chi-squared
20 SNP PCs	1.0292	0.0061	1.0195	1.0149
20 cp-PCs	1.0159	0.0061	1.0046	1.0064
20 cp-PCs (LOCO)	1.0164	0.0061	1.0046	1.0053

To further explore the differences in the inflation between SNP PCA corrected and cp-PCA corrected GWAS we generated QQ-plots for each correction method (Figure 5.3). Firstly considering all SNPs (Figure 5.3; black) we observed that our SNP PC corrected GWAS showed the greatest overall inflation ( $\lambda=1.017$ ) as shown by a deviation from the expected uniform null distribution in the QQ plot for  $-\log(p\text{-values})$ , which was not present in our cp-PCA ( $\lambda=1.003$ ) or cp-PCA LOCO ( $\lambda=1.002$ ) corrected GWAS (Figure 5.3; black). This inflation could represent disease association, but could also be due to population stratification. To distinguish these sources of inflation we stratified the SNPs into bins based on their p-values in a larger independent subset of the 2016 ALS GWAS ( $n = 31,299$ ; Dutch samples excluded) (van Rheenen et al. 2016), which we take as a measure of confidence that they have a “true” association with ALS. This allowed us to explore whether inflation resided in likely disease-associated SNPs or not. The SNP PCA corrected GWAS shows clear inflation across all bins, including the bin for variants with the weakest association (Figure 5.3 a; orange) in the larger ALS dataset, with increasing inflation in bins more strongly associated with ALS. While the increasing inflation across these confidence bins suggests that this GWAS is enriched for association signal in variants likely associated with ALS, inflation in the weakest association bins

means it is likely also subject to some level of confounding. In contrast cp-PCA and cp-PCA LOCO show almost no inflation in the weakest association bins, while retaining clear enrichment in the strongest association bin (Figure 5.3 b and c; Green), suggesting that they correct out noise missed by SNP PCA while preserving signal at important loci.

However the cp-PCA correction shows evidence of deflation in the second strongest association bin (Figure 5.3 b; red, lambda = 0.764), which may imply it overcorrects potentially disease associated variants. Signal in this bin appears to be regained when using the cp-PCA LOCO approach (lambda = 1.329), which due to design is less likely to overcorrect. These results coupled with our LD score regression intercepts suggest that cp-PCA based correction approaches correct GWAS more stringently while retaining power to detect signal at variants likely associated with ALS. We will further explore this in a larger multi-population GWAS setting in section 5.3.3 affording us greater power to detect real associations.



**Figure 5.3: Stratified QQ-plots under unlinked and linked correction methods.**

QQ-plots for the Dutch ALS GWAS ( $n=4,753$ ) corrected using a.) SNP PCs as covariates, b.) cp-PCs as covariates and c.) cp-PCs as covariates in a leave one chromosome out analysis (LOCO). Plots are stratified by p-value thresholds from an ALS GWAS run on samples from van Rheenen et al. (van Rheenen et al. 2016) excluding these Dutch individuals to demonstrate the relative effects of each correction on SNPs related to ALS at different confidences.

a.) Considering all SNPs (black), SNP PCA corrected GWAS deviates significantly from the null for demonstrating clear inflation. This inflation is spread across all significance bins from the larger GWAS dataset, suggesting that it must be at least partially due to non-disease related signal (e.g. confounding from population structure). b-c.) In contrast the GWASes run with cp-PC covariates show significantly reduced total inflation (black), and demonstrate lower inflation in bins with less evidence of association with ALS (orange and blue) suggesting this confounding is better corrected. Both cp-PCA and cp-PCA LOCO retain inflation in SNPs with strongest evidence of association with ALS (green) suggesting they retain ALS associated signal. Lambda for each SNP set is the observed median chi-squared divided by 0.4549 (the expected median of a chi-squared distribution with 1 d.f.).

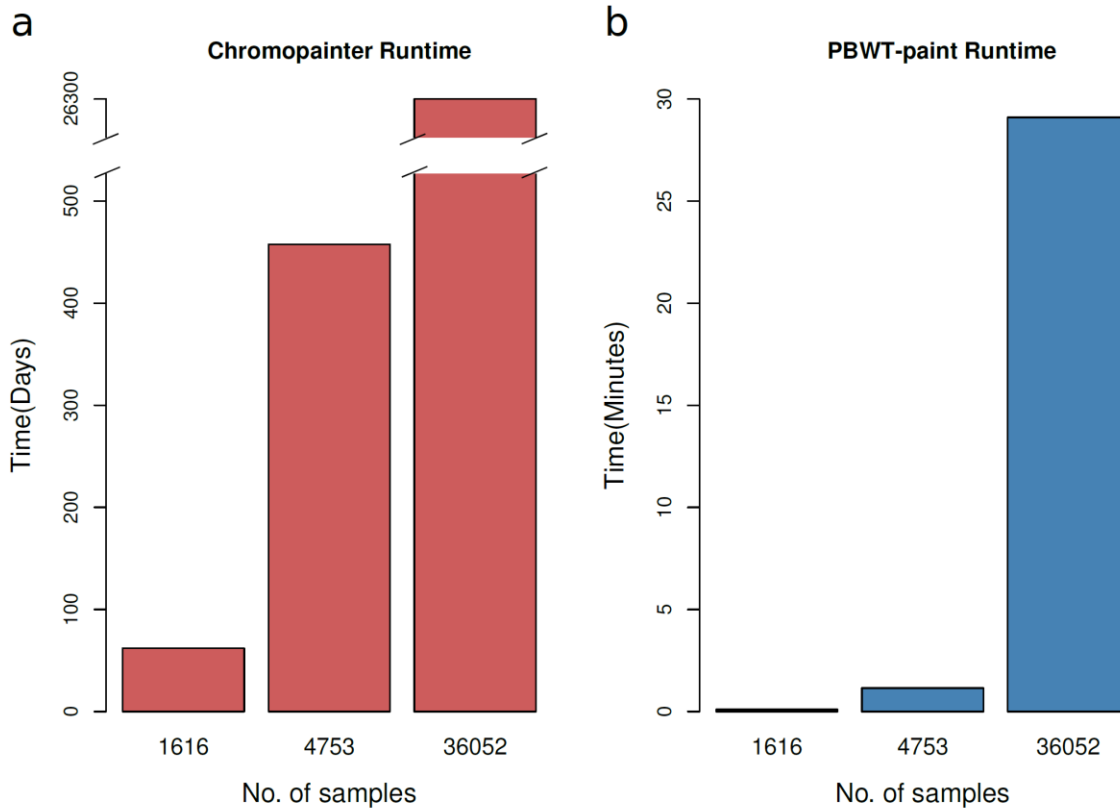


### 5.3.2 - Detecting population structure in large datasets with scalable methods

As our Dutch-only haplotype sharing-based GWAS correction showed subtle reductions in confounding compared to SNP PCA (Table 5.1) we aimed to scale up to the full 2016 ALS GWAS dataset (n=36,052) and explore the effects of haplotypic correction in a larger multi-population setting. However ChromoPainter is extremely computationally costly to run and runtime scales quadratically with sample size making this analysis intractable at this scale. Hence we investigated the potential application of PBWT-paint (<https://github.com/richarddurbin/pbwt>; -paint switch), a tool which approximates the ChromoPainter algorithm using the Positional Burrows-Wheeler transform (PBWT) (Durbin 2014) to speed up haplotype matching in large datasets.

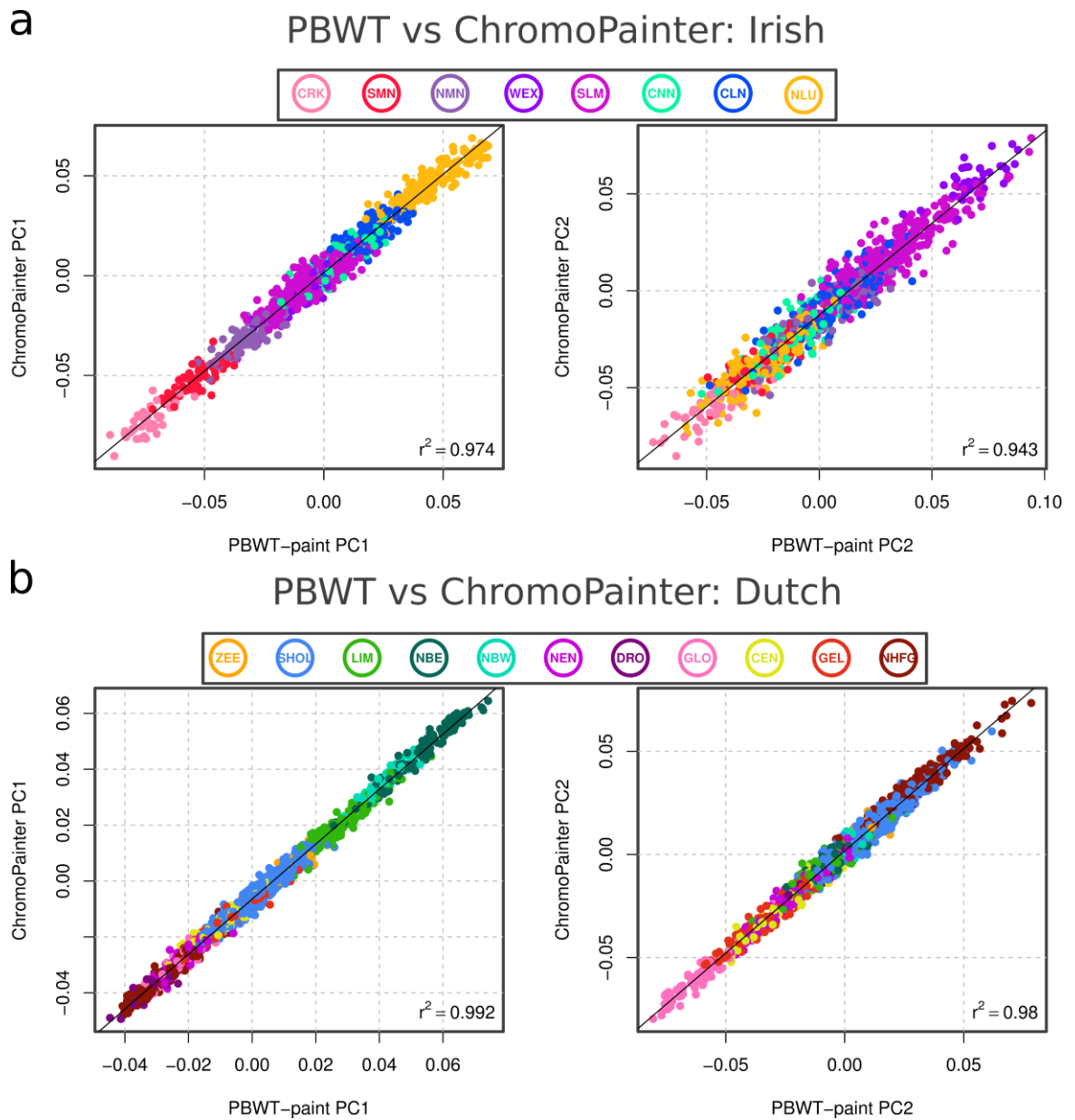
We first estimated the runtimes of the PBWT-method for a single chromosome in datasets of varying sample sizes and compared them to the projected runtimes for ChromoPainter in these datasets to assess the computational gains (Figure 5.4). While PBWT-paint takes only ~30 minutes to run for chromosome 20 with 36,052 samples on a single core, ChromoPainter is projected to take ~26,300 days for the same data, hence PBWT-paint is magnitudes (~1.26 million fold) faster for our desired application. This implies that the PBWT-paint method is computationally scalable for detecting population structure in large GWAS datasets. With this in mind we next compared the accuracy of PBWT-paint to ChromoPainter.

We ran PBWT-paint on the Irish dataset from Chapter 3 and the Dutch dataset from Chapter 4 to benchmark its accuracy against ChromoPainter. We assessed whether major trends in the haplotype sharing methods were consistent by regressing the first two PCs of each method against each other, which demonstrated strong concordance (Figure 5.5). Additionally we estimated Pearson's correlation coefficient for all pairwise entries between co-ancestry matrices produced with the two methods for the Irish and Dutch dataset (Pearson's  $\rho$  Irish = 0.817 (0.816-0.818); Pearson's  $\rho$  Dutch = 0.8204 (0.820-0.8208);  $p < 2.2 \times 10^{-16}$  for both datasets), which demonstrated strong correlation between the outputs of each algorithm regardless of dataset used.



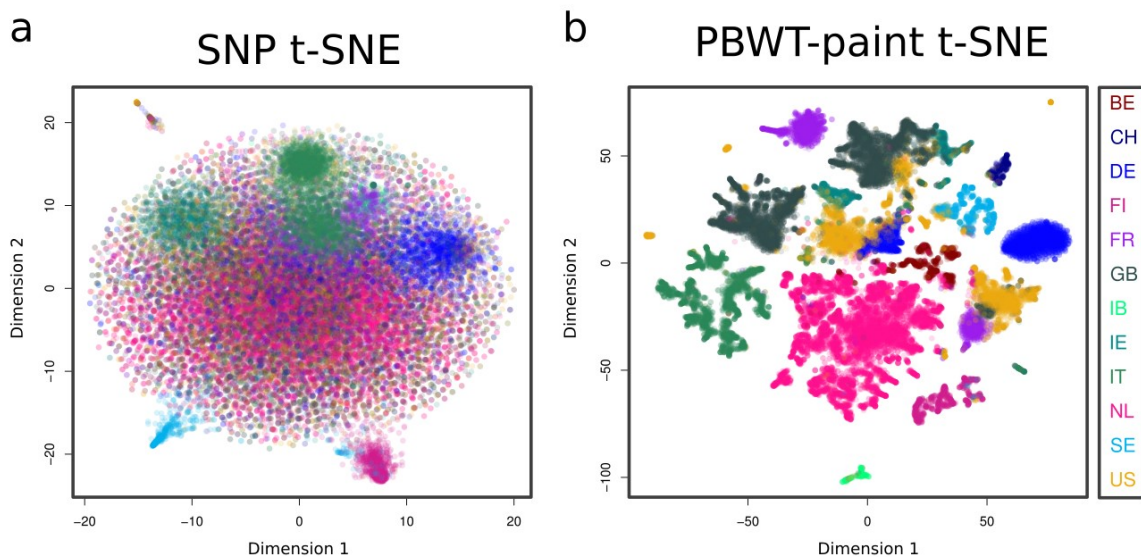
**Figure 5.4: Runtimes for PBWT-paint compared to projected runtimes for ChromoPainter.**

Example runtimes for a.) ChromoPainter (days) and b.) PBWT-paint (minutes) for chromosome 20 calculated on a single core for three datasets (Dutch  $n=1,626$ ; Dutch  $n=4,753$  and 2016 GWAS 36,052). The axis for ChromoPainter runtimes is discontinuous due to large jump in runtimes between datasets. ChromoPainter runtimes are extrapolated from painting a single individual in terms of the remaining individuals in the dataset due to computational expense while PBWT-paint runtimes are directly measured. PBWT-paint shows magnitudes faster runtime particularly as sample sizes increase.



**Figure 5.5: Benchmark of PBWT-paint vs ChromoPainter in Irish and Dutch data.** Scatterplots comparing the first two principal components (PCs) of the coancestry matrices produced by ChromoPainter and PBWT-paint for the a.) Irish dataset (Chapter 3) and b.) Dutch dataset (Chapter 4). The two painting methods show strong correlation ( $r^2 > 0.94$  for all plots) indicating the methods produce similar results regardless of dataset. Points are coloured by cluster groups defined in Figures 3.8 (Irish) and 4.1 (Dutch). For the full set of pairwise comparisons in the Irish coancestry matrices Pearson's  $\rho = 0.817$  (0.816-0.818;  $p < 2 \times 10^{-16}$ ) and for the Dutch coancestry matrices Pearson's  $\rho = 0.82$  (0.82-0.821;  $p < 2 \times 10^{-16}$ ).

Given the high fidelity and significantly reduced runtime of the PBWT-paint method on small test datasets we applied this method to detect structure in the full multi-population dataset ( $n=35,985$ ; 67 haplotype sharing outliers removed). The resulting haplotype sharing matrix identified structure at a far more granular level than SNP PCA, as evidenced by t-SNE initialised with haplotype sharing PCs (Figure 5.6 b) which shows both strong global structure, separating samples from different countries into broad clusters in t-SNE space and even subdividing clusters into more fine-grained local subgroups within countries. For example t-SNE of this haplotype sharing matrix forms sub clusters within the Italian samples consistent with previous findings of finescale population structure in Italy (Raveane et al. 2019). In contrast, while t-SNE of SNP PCA shows some clustering of samples within countries (Figure 5.6 a) it largely forms a diffuse cloud of points with overlapping samples from many countries rather than tight clusters. This indicates that the structure detected using SNP PCs is less resolved. We later explore the relative gains of using these haplotype sharing PCs to correct confounding in GWAS (Section 5.3.3).



**Figure 5.6: Describing structure in large multi population datasets using t-SNE initialised with PBWT-paint PCs and SNP PCs.**

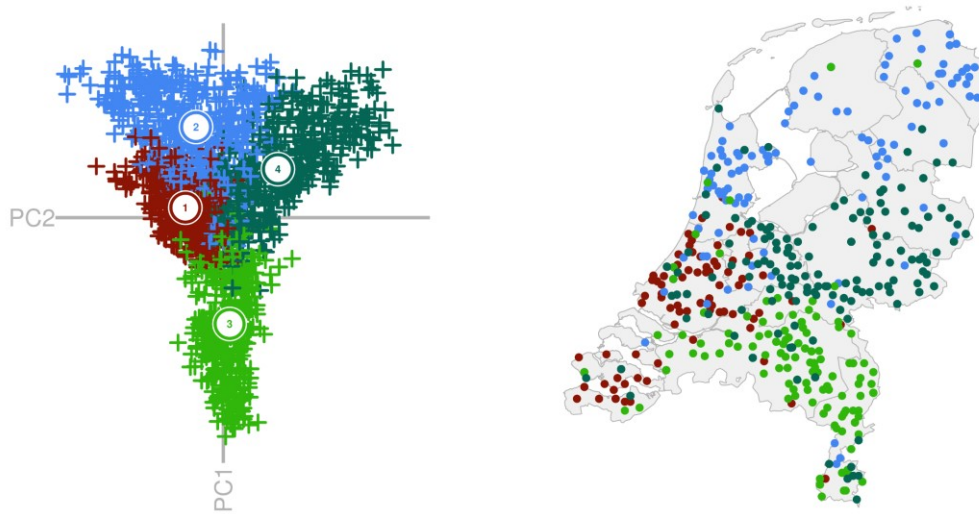
t-distributed stochastic neighbour embedding (t-SNE) plots for the multi-population dataset ( $n=35,985$ ) initialised on a.) 100 SNP PCs and b.) 100 PBWT-paint PCs. Points are coloured by the country of sampling and semi-transparent to emphasise overlap of samples. Labels (right) follow ISO 3166-1 country codes, except IB, which was labelled Iberia (containing Spanish and Portuguese data) in the original GWAS dataset. The PBWT-paint method (a) shows discrimination of both between-country and within-country genetic structure, identifying several tight country and subcountry clusters. In contrast the SNP PCA initialised t-SNE projection forms a dispersed cloud of points with less obvious genetic structure.

As the clustering observed in our t-SNE plots is largely for visualisation purposes, and does not define meaningful boundaries for population subgroups we attempted to algorithmically define clusters to further describe the extent of structure seen in the dataset in a manner analogous to fineSTRUCTURE (used in Chapters 3 and 4). Given that fineSTRUCTURE is computationally intractable for large datasets such as this one, we adopted the Louvain community detection method (Blondel et al. 2008) as implemented in the R igraph package (Csardi, Nepusz, and Others 2006) which is used to detect densely connected groups in large networks. This algorithm has previously been successfully applied to identify recent population structure in a large IBD sharing dataset (Han et al. 2017). To ensure this algorithm returned sensible genetic clusters in this context we first applied it to the 1,626 Dutch samples from chapter 4 (Figure 5.7 a and b), before applying it to the PBWT-paint matrix of the multi population dataset (Figure 5.7 c and d). For these analyses we constructed haplotype sharing networks treating individuals as nodes and weighting edges between these nodes based on the number of haplotypic “chunks” they shared. Following the initial assignment of individuals to clusters, we applied the algorithm for a second iteration treating these clusters as independent networks to identify further subdivisions in the data. Our testing in the Dutch dataset from Chapter 4 produced genetic clusters with tight geographic distributions both in the first and second iteration, suggesting this method is suitable for identifying population subgroups (Figure 5.7 a and b). The groupings seen in the second iteration strongly resemble the clusters defined by fineSTRUCTURE in this data and include notable splits from our analysis in chapter 4, including the separation of eastern and western groups in North Brabant. These clusters have a mean  $F_{ST}$  on par with the fineSTRUCTURE clusters from Chapter 4 (mean  $F_{ST} = 5.9 \times 10^{-4}$ ).

## Dutch preliminary data (n=1,626)

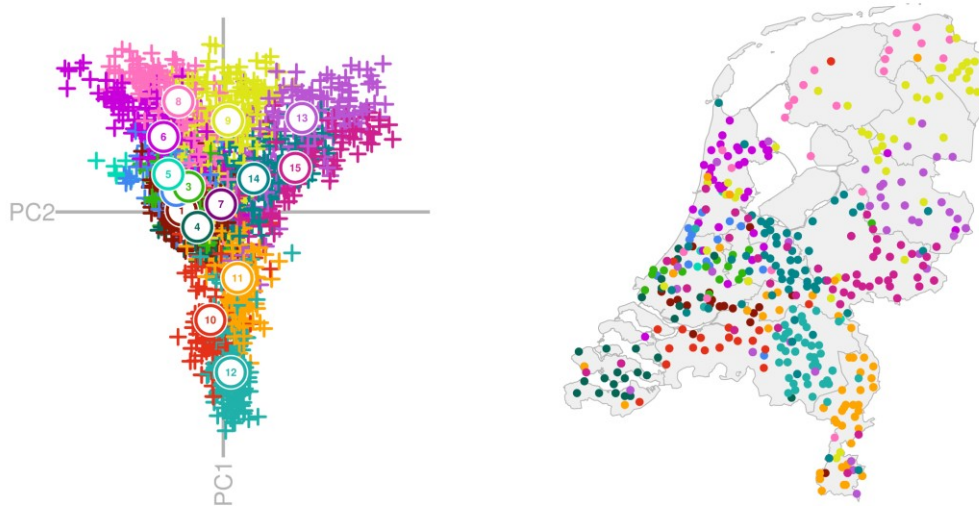
a

Louvain clustering: Iteration 1



b

Louvain clustering: Iteration 2



**Figure 5.7: Louvain community detection method for identifying population subgroups in preliminary data.**

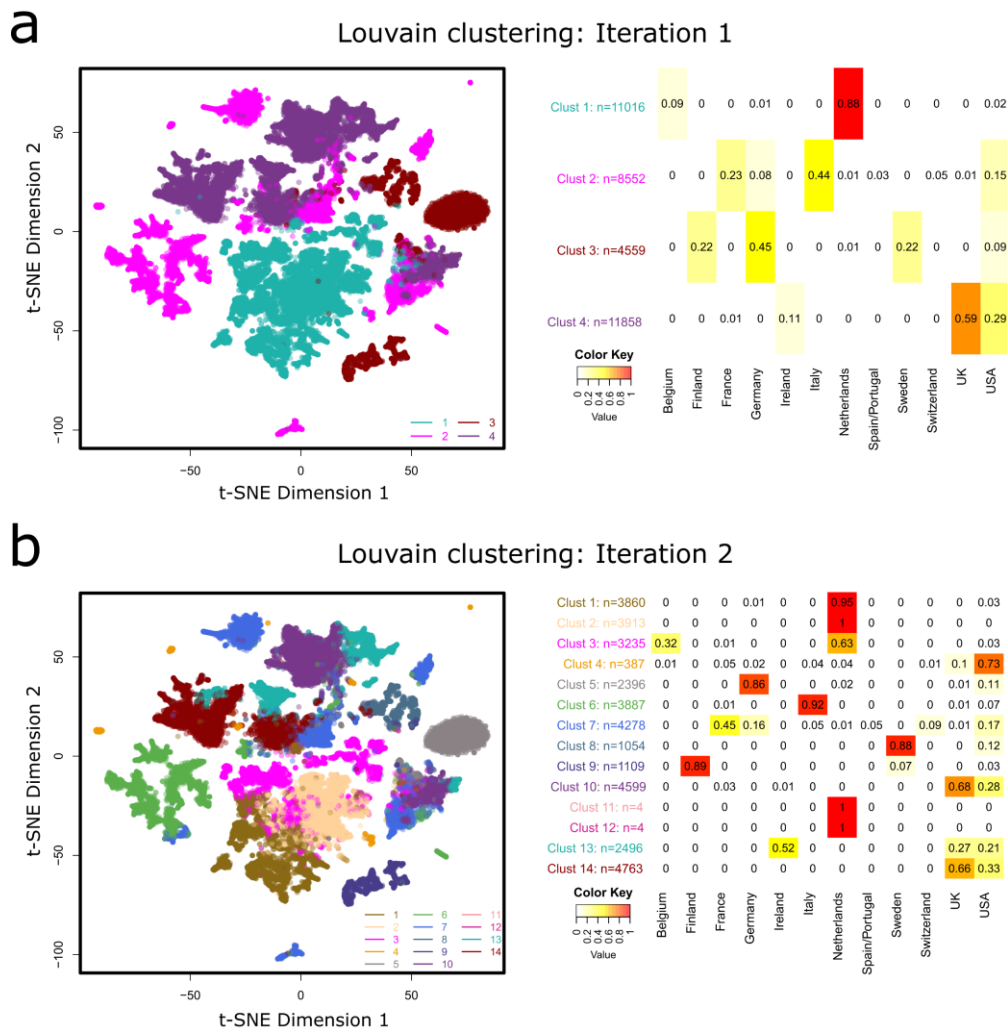
Displays genetic clusters identified using the Louvain community detection method on haplotype sharing data from a preliminary test set of 1,626 Dutch samples (data from Chapter 4). Samples are plotted on the first two principal components of the ChromoPainter coancestry matrix and according to geographic origin and coloured by clusters defined using the Louvain community detection method using a.) a single iteration and b.) two iterations. Clusters identified show clear definition in PC space and segregate well with local geography. Clusters from the second iteration capture most of the major sub populations identified by fineSTRUCTURE in this dataset in Chapter 4.

When applied to the larger multi population dataset the algorithm identified subgroups well defined in t-SNE space (Figure 5.8). After two iterations these subgroups typically consist of samples mostly from a single country of origin (i.e. Clust 1 - the Netherlands; Figure 5.8 b), or mixtures of closely related groups (i.e. Clust 3 - the Netherlands and Belgium; Figure 5.8 b) suggesting this approach can successfully identify ancestral subpopulations in large datasets. Further iterations (Appendix Figure 5.2) identify several subgroups within single countries such as Italy, Finland, the Netherlands and the UK, all of which have previously been shown to harbour local fine scale genetic structure (Abdellaoui, Hottenga, de Knijff, et al. 2013; Leslie et al. 2015; Kerminen et al. 2017; Raveane et al. 2019), suggesting that both broad and local population structure can be simultaneously detected in these large datasets using haplotype sharing and network clustering approaches. This means that the current study design of exploring finescale population structure using haplotype sharing as seen in Chapters 3 and 4 has come of age for application to large multi-population datasets allowing us to identify subtle structure both within and across countries.

Unlike fineSTRUCTURE this approach does not have a clear protocol for deciding when to stop splitting, given that one can simply subset a given cluster and rerun the algorithm on the resulting network until modularity plateaus at its minimum value of -0.5. Hence this approach leaves some level of judgement to the user. While one could define a minimum modularity threshold at which to stop performing iterations in a particular chain of clusters, choosing a cutoff based on a minimum value of genetic differentiation between groups such as  $F_{ST}$  or TVD may prove a more sensible approach for this particular application as these values have more interpretable meanings in a population genetics context. An advantage of the loose termination criteria is that one can explore potentially meaningful subtle substructure even where methods such as fineSTRUCTURE would algorithmically assign a single cluster. As an example of this we ran a single iteration of Louvain clustering on the haplotype sharing matrix for a large fineSTRUCTURE cluster of individuals from the “South East England” cluster from Chapter 3 (SEE, also seen in the source PoBI paper (Leslie et al. 2015)) and observed meaningful subclusters (Figure 5.9) in this supposedly homogeneous cluster. These genetic clusters segregated reasonably well with geography (Figure 5.9 b) suggesting they may reflect real non-randomly mating subpopulations in the data. Hence it is possible that fineSTRUCTURE’s likelihood function under-split this group of individuals due to the minimal gains to model fit from separating them. Notably,  $F_{ST}$  estimates between these Louvain clusters are extremely low (mean  $F_{ST} = 6.02 \times 10^{-5}$ ) meaning that despite their geographic separation, these clusters are quite close to panmixia, and their differences are on a very fine scale. In fact, these clusters

have a significantly lower mean  $F_{ST}$  than the Irish clusters seen in Chapter 3 (mean  $F_{ST}=3.5\times 10^{-4}$ ), despite occupying an equivalent geographic area. This demonstrates that Louvain clustering can divide highly homogeneous groups into extremely finescale, but potentially meaningful sub clusters.

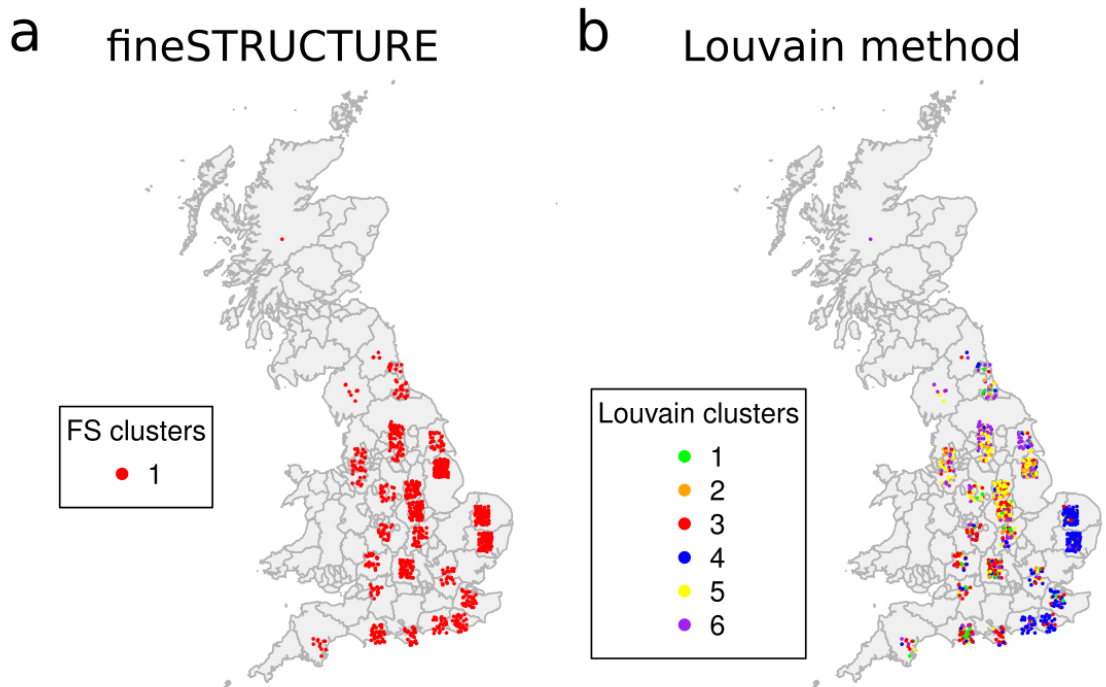
### Multi-population data (n=35,985)



**Figure 5.8: Detecting population structure in large haplotype sharing datasets with the Louvain community detection method.**

Displayed are clustering results obtained by running the Louvain community detection algorithm on the PBWT-paint co-ancestry matrix for the multi-population dataset (n=35,985) for a.) one iteration or b.) two iterations. Samples are projected into t-SNE space based on haplotype sharing patterns and coloured by cluster assignment (left). To the right are heatmaps of the proportion of samples from each cluster (rows) originating from a given country (columns). Cluster names are coloured according to the same scheme in the t-SNE plots for ease of reference. By the second clustering iteration, clusters are mostly composed of samples from a single country (i.e. Clust 1 and Clust 2 are mostly samples from the Netherlands) or ancestrally related countries (i.e. Clust 13 is mostly from Ireland, with large numbers of samples from the UK and US, likely due to ancestral ties and migration).





**Figure 5.9: Louvain method for community detection identifies extremely subtle splits missed by fineSTRUCTURE in the “indivisible” SEE cluster.**

a.) The geographic spread of a large seemingly indivisible fineSTRUCTURE cluster in the POBI dataset from chapter 3 ( $n=966$ ; SEE; Appendix Figure 3.4). These samples form a single cluster at the finest level of the fineSTRUCTURE tree despite their wide geographic range. A similar indivisible cluster was observed spanning this geographic range in the original analysis of the data ( $n=1,006$ ) (Leslie et al. 2015).

b.) The geographic spread of six sub-clusters identified within this large fineSTRUCTURE cluster using the Louvain method for community detection on the haplotype sharing matrix for these samples. Genetically defined clusters show a clear relationship with geography indicating they represent meaningful subpopulations. Notably these clusters are very genetically similar (mean  $F_{ST}=6 \times 10^{-5}$ ) suggesting the differentiation detected here is extremely finescale.

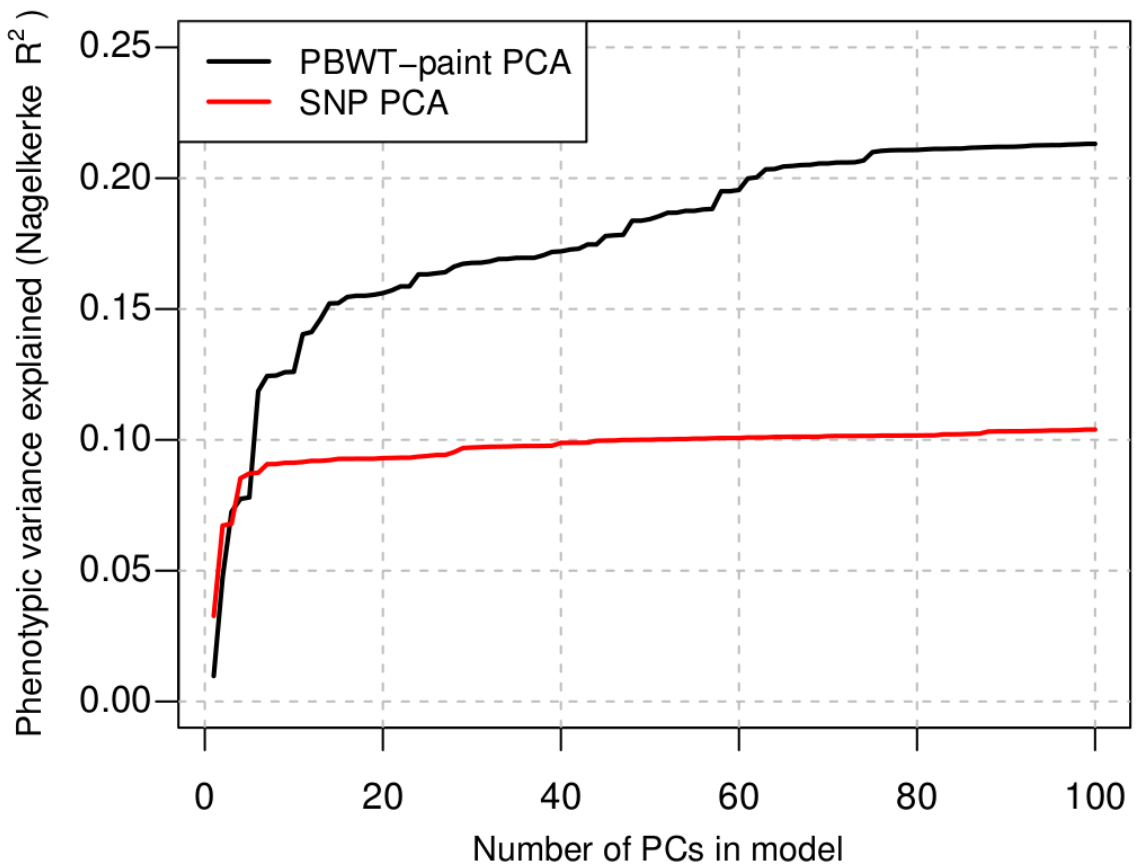
NB: As only sampling regions are available, points are jittered around a fixed point in the sampling region to separate overlapping samples from the same sampling region.

### 5.3.3 - Correcting multi-population GWAS structure with haplotype sharing methods

As the preliminary work exploring correction of GWAS using haplotype sharing PCs (section 5.3.1) was performed on a small underpowered GWAS ( $n=4,753$ ), it is difficult to fully evaluate how well the method performs. For example, while ChromoPainter appears to correct confounding more stringently than SNP PCA (Table 5.1), there is little inflation to begin with; additionally it is difficult to assess whether the method overcorrects or produces more false negatives than SNP PCA corrected GWAS as no loci reach genome-wide significance in the dataset under any correction scheme considered. Hence, we next expanded this analysis to the multi-population dataset using PBWT-paint PCs (calculated in section 5.3.2) as covariates.

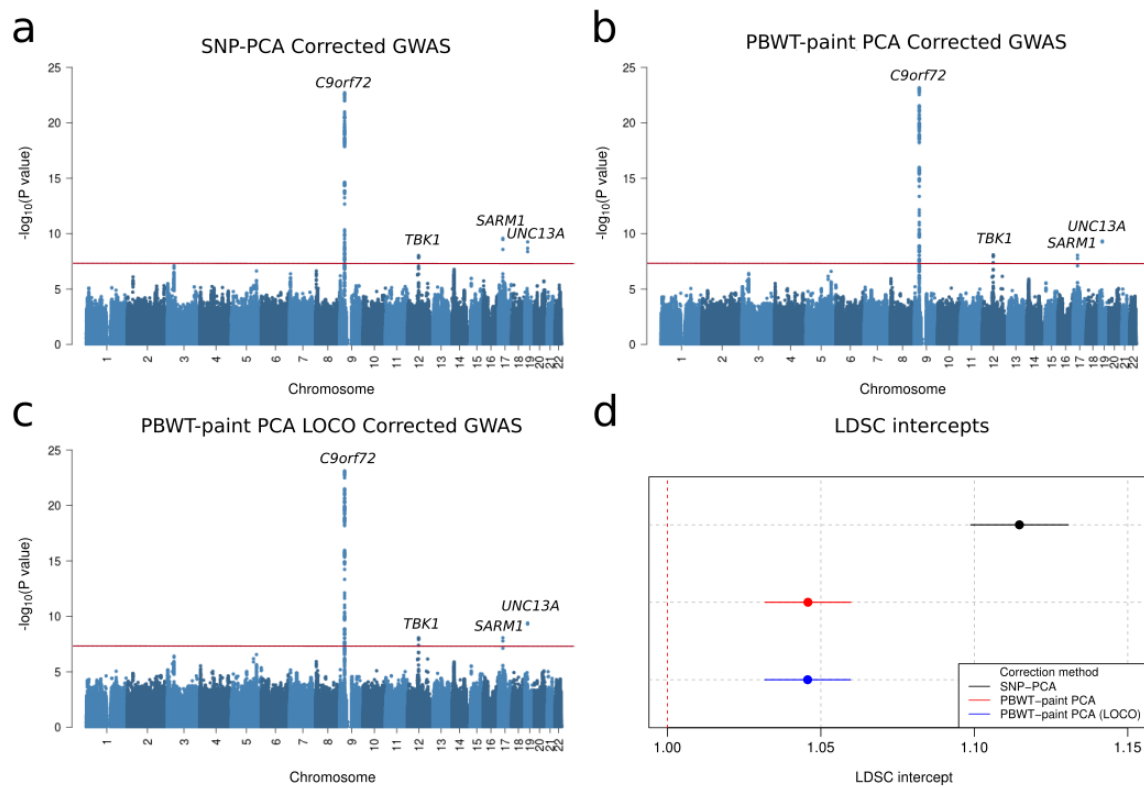
This dataset has two major advantages, the first being sample size and the second being that it is sampled from multiple populations, making its design closer to that commonly used in modern GWAS (with the exception of single population biobanks). Top PCs of PBWT-paint coancestry explained substantially more variance in phenotype than SNP PCs (Figure 5.10), recapitulating our result from the Dutch-only GWAS (Figure 5.1). The majority of variance explained by PCs from both methods was captured by PCs 1-20, motivating the choice of 20 PCs in downstream analyses. Strikingly, LD score regression intercepts show that GWAS statistics calculated including PBWT-paint PCs as covariates were significantly less confounded than statistics corrected by SNP PCA ( $p=1.5\times 10^{-10}$ ; Figure 5.11 d), while retaining the power to detect the same known ALS hits (Figure 5.11 a-c). Together these results indicate that genetic structure described by PBWT-paint PCs captures patterns of stratification of ALS cases and controls in the data missed by standard SNP PCA and hence better corrects confounding when included as a covariate. Additionally it appears that PBWT-paint PCs do not overcorrect loci associated with the disease given that the method detected the same hits as SNP PCA corrected GWAS. As previously (Section 5.3.1) we also ran a GWAS using a leave one chromosome out (LOCO) approach when fitting PBWT-paint PCs to minimise overcorrection caused by patterns of haplotype sharing at the locus being tested. Notably the LD score intercept of this LOCO analysis was practically indistinguishable from the LD score intercept run using PCs calculated from the full PBWT-paint matrix ( $p=0.99$ ), indicating that the decrease in LD score intercept from PBWT-paint is due to reduction of confounding, rather than loss of ALS-associated GWAS signal.

### Multi-population ALS Dataset (n=35,985)



**Figure 5.10: Phenotype stratification captured by PBWT-paint PCs and SNP PCA in multi-population dataset.**

Cumulative variance in phenotype (amyotrophic lateral sclerosis) explained by principal components (PCs) of the PBWT-paint coancestry matrix and standard SNP PCs in the multi-population GWAS dataset (n=35,985). As in the single population setting (Figure 5.1) PBWT-paint PCs explain a greater amount of variance in phenotype and are hence expected to better correct for confounding due to stratification in GWAS.



**Figure 5.11: Comparing GWAS power and inflation when corrected with haplotype sharing or SNP PCA.**

a-c.) Manhattan plots for GWAS carried out on the multi-population ALS dataset ( $n=35,985$ ) corrected using a.) SNP PCs as covariates b.) PBWT-paint PCs as covariates and c.) PBWT-paint PCs as covariates following a leave one chromosome out approach (LOCO). Each method appears similarly powered and detects the same genome-wide significant loci.

d.) LD-score regression intercepts coloured by GWAS correction method demonstrate that PBWT-paint PCA corrected GWAS summary statistics (LOCO and regular) are subject to significantly ( $p=1.5 \times 10^{-10}$ ) less confounding than SNP PCA corrected GWAS.

### 5.3.4 - Addressing bias from residual stratification in polygenic methods using haplotype sharing PCs

#### Polygenic risk scores:

We next explored the effect of PBWT-paint PC correction of ALS GWAS training data on the distribution of polygenic risk scores in our Dutch dataset and their relationship to geography. GWAS conducted on a training ALS dataset excluding the Dutch data were run using a.) 20 SNP PCs or b.) 20 PBWT-paint PCs as covariates. The best fit PRS model using summary statistics from a SNP PC corrected GWAS explained more variance in phenotype than the best PBWT-paint corrected model (Table 5.2). However, this model included variants from all significance thresholds, which may imply overfitting given the relatively low polygenicity of ALS (see Chapter 1; Figure 1.2), meaning not all variants are expected to have predictive power for this phenotype. In fact, most of the predictive power in the SNP PCA corrected model comes from variants which had a p-value greater than 0.05 in the training dataset, supporting the hypothesis of overfitting (Figure 5.12 a). In contrast the best fit PBWT-paint PC corrected model only includes SNPs with p-values below a threshold of  $2.15 \times 10^{-3}$  in the base dataset which have much greater evidence of association with ALS. Higher p-value thresholds contribute almost no predictive power to the PBWT-paint corrected model (Figure 5.12 a), indicating it is very unlikely to be overfitted (though overcorrection is possible).

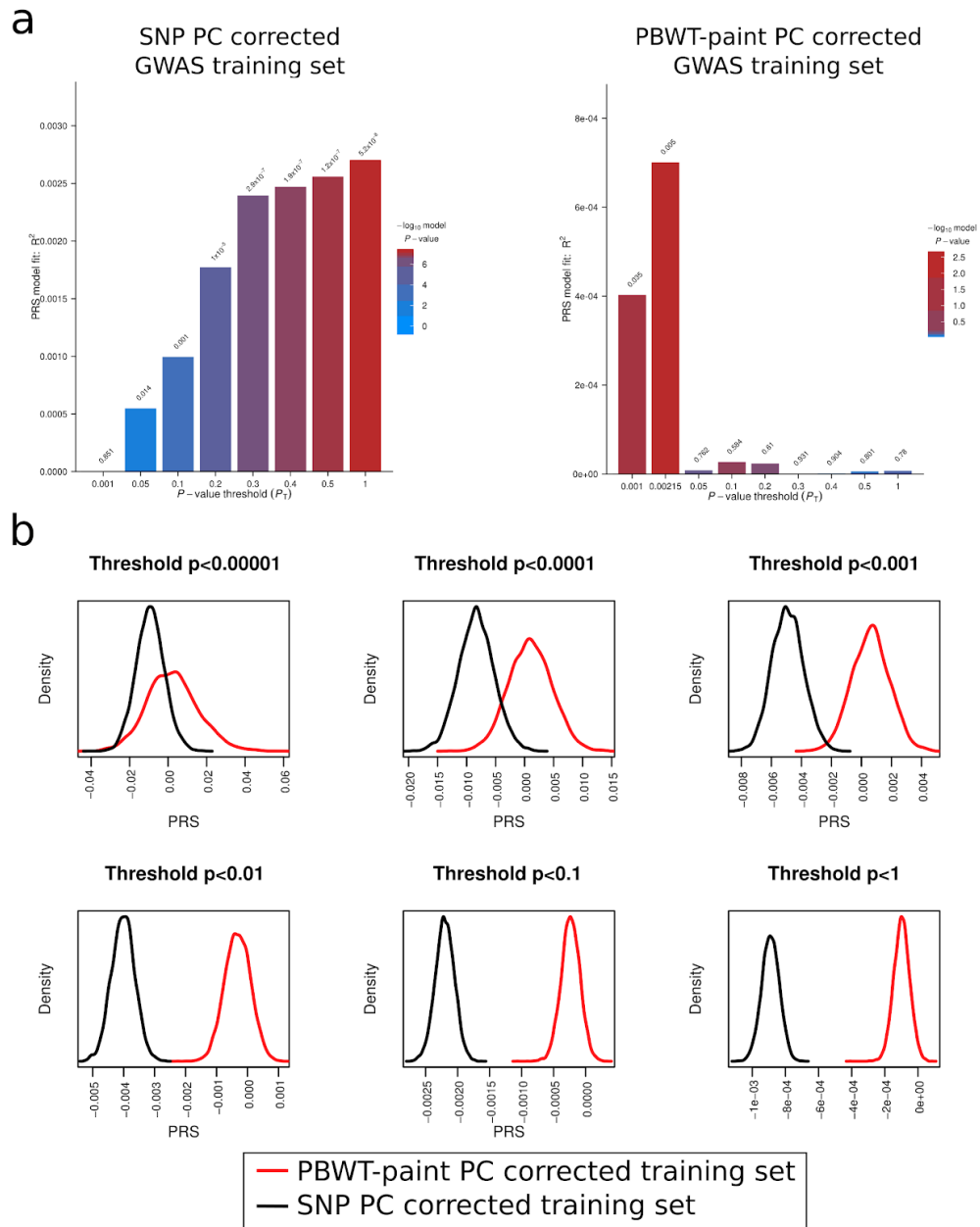
**Table 5.2: Best fit PRSice model details for SNP and PBWT-paint PC correction methods.**

Covariates in training GWAS	PRS Model fit ( $R^2$ )	Model fit (p)	SNP Inclusion threshold ( $p < x$ in training set)	# SNPs included
20 SNP PCs	$2.7 \times 10^{-03}$	$5.16 \times 10^{-08}$	1	304,806
20 PBWT-paint PCs	$7.0 \times 10^{-04}$	0.005	0.00215	3,317

Model fits (liability scale) for prediction of phenotype from polygenic risk scores generated from SNP PC-corrected and PBWT-paint PC-corrected GWAS for ALS. The SNP PC-corrected method better predicts ALS case-control status in the Dutch data, however this model uses all variants, including those with little association with ALS, meaning overfitting is probable.

Distributions of PRS for the two correction methods are notably different across all SNP inclusion thresholds (Figure 5.12 b). While PBWT-paint PC-corrected PRS scores are roughly normally distributed around 0 for all inclusion thresholds, SNP PC-corrected PRS scores trend away from zero as more variants with less evidence of association with ALS are added. This growing separation of the distributions as variants with less evidence of

association with ALS are added suggests that the SNP PC-corrected PRS scores (and GWAS effect sizes) may be biased by residual confounding at these non-significant variants. To test for residual confounding in the PRS scores produced by each method we looked at the relationship between PRS scores and geography using Moran's I, which measures the spatial autocorrelation of a variable. We first regressed the phenotype out of the PRS scores to get residual PRS scores which are non-correlated with phenotype (i.e. represent the component of the PRS that does explain the phenotype). This was done to correct for correlation between the phenotype and geography, which would confound our estimation of residual population structure. We then calculated Moran's I for these residual PRS scores for Dutch samples with geographic information (n=1,352) to capture the relationship of the non-disease related component of the PRS scores with geography (representing population structure). Residual PRS scores calculated with all variants had a significant relationship with geography for the SNP PC-corrected analysis (Table 5.3; Moran's I=0.06 ;  $p=9.9\times 10^{-05}$  ), but not for the PBWT-paint corrected method (Table 5.3; Moran's I=-8.3 $\times 10^{-03}$  ;  $p=0.22$  ). This suggests that the SNP PC-corrected training set produces scores that are predictive of population structure that does not relate to the phenotype at this threshold. Hence the differences in model fit and PRS distribution (Figure 5.12) are likely at least partially due to residual confounding from population structure in the effect size estimates from the SNP PC-corrected GWAS. Additional SNP inclusion thresholds show significant spatial autocorrelation in the SNP PC-corrected analysis (Table 5.3), all with values significantly greater than zero suggesting spatial clustering of PRS values. In contrast the PBWT-paint PC-corrected PRS scores showed no significant sign of bias due to residual confounding. Hence PBWT-paint corrected summary statistics appear to generate less biased PRS scores.



**Figure 5.12: PRS model fit and distribution is affected by correction method.**

a.) PRS model fit (liability scale) for predicting ALS case-control status in the Dutch-only dataset ( $n=4,753$ ) using the remaining data from the 2016 GWAS ( $n=31,299$ ) corrected with SNP PCs or PBWT-paint PCs. The SNP PC-corrected analysis shows a trend towards better prediction (y-axis; PRS model fit:  $R^2$ ) as variants in less stringent p-value thresholds are included. The PBWT-paint PC-corrected analysis, however, shows best fits for lower p-value thresholds and limited signal as additional SNPs are added.

b.) Distributions of PRS scores from both correction approaches at a range of p-value thresholds show differences in distributions. PRS from PBWT-paint PC corrected PRS are distributed closer to a mean value of zero while SNP PC-corrected PRS are shifted towards more negative scores. Differences in the distributions are most striking for the model with all SNPs included (threshold  $p < 1$ ) which includes many variants which should have a null effect on ALS status.

**Table 5.3: Measures of spatial autocorrelation for PRS under SNP PC and PBWT-paint PC correction.**

SNP inclusion threshold ( $p < x$ Training set)	SNP PC-corrected training set PRS		PBWT-paint PC-corrected training set PRS	
	Moran's I	p	Moran's I	p
1	$6.0 \times 10^{-02}$	<b><math>9.9 \times 10^{-05}</math> *</b>	$-8.3 \times 10^{-03}$	$2.2 \times 10^{-01}$
0.1	$4.1 \times 10^{-02}$	<b><math>9.9 \times 10^{-05}</math> *</b>	$-5.3 \times 10^{-03}$	$5.5 \times 10^{-01}$
0.01	$1.9 \times 10^{-02}$	<b><math>2.0 \times 10^{-04}</math> *</b>	$-3.0 \times 10^{-03}$	$9.1 \times 10^{-01}$
0.001	$3.5 \times 10^{-03}$	$1.9 \times 10^{-01}$	$6.9 \times 10^{-04}$	$5.0 \times 10^{-01}$
0.0001	$4.1 \times 10^{-03}$	$1.6 \times 10^{-01}$	$-1.3 \times 10^{-02}$	$3.3 \times 10^{-02}$
0.00001	$-5.6 \times 10^{-03}$	$4.9 \times 10^{-01}$	$-2.1 \times 10^{-03}$	$9.2 \times 10^{-01}$

Moran's I values measure spatial autocorrelation for residuals of polygenic risk scores (PRS) with phenotype regressed out. Several inclusion thresholds show significant spatial clustering for the SNP PC corrected method suggesting the method is biased by residual population structure. In contrast none of the thresholds show significant spatial autocorrelation for the PBWT-paint corrected method. Values marked with an asterisk (\*) are significant at the Bonferroni corrected p-value of  $4.1 \times 10^{-03}$ .

Estimates of SNP heritability:

Controversially, the estimate of SNP heritability for ALS obtained using LD score regression on summary statistics from a larger ALS GWAS from 2018 (Nicolas et al. 2018) ( $h^2_{\text{SNP ALS 2018}} = 0.016$ ; 95% c.i: 0.009-0.023) is substantially lower than published GREML  $h^2_{\text{SNP}}$  estimate from the 2016 ALS GWAS (van Rheenen et al. 2016) ( $h^2_{\text{SNP ALS 2016}} = 0.085$ ; 95% c.i: 0.0752-0.0948) suggesting that bias from population structure may have inflated this earlier heritability estimate. To explore this possibility we ran GREML heritability analysis on the 2016 GWAS dataset controlling for population structure with 20 SNP-based PCs or 20 PBWT-paint PCs and compared the estimates of  $h^2_{\text{SNP}}$ . The genome-wide estimate of  $h^2_{\text{SNP}}$  controlling for PBWT-paint PCs was significantly lower than using standard SNP PC based correction (Table 5.4;  $p=9.6 \times 10^{-5}$ ). This could mean that population stratification in this dataset is leading to overestimation of SNP-based heritability, or alternatively that the method is overcorrecting by capturing similar patterns of relatedness to the GRM. Notably our SNP PC-corrected estimate ( $h^2_{\text{SNP}}=0.066$ ) is also lower than the original published estimate (van Rheenen et al. 2016), however that estimate used only 10 PCs to correct for structure, meaning this difference may also be reconciled by residual structure.



**Table 5.4: Total unpartitioned GREML SNP-based heritability estimates from 2016 ALS GWAS under SNP and PBWT-paint PC corrections.**

Correction method	$h^2_{\text{SNP}}$	SE
20 SNP PCs	6.6E-02	4.8E-03
20 PBWT-paint PCs	4.0E-02	4.6E-03

Estimates of total unpartitioned SNP heritability from the 2016 ALS GWAS dataset corrected using 20 SNP PCs or 20 PBWT-paint PCs as covariates in the model. The PBWT-paint PC corrected estimate is significantly lower than the SNP PC corrected estimate ( $p=9.6\times 10^{-5}$ )

To explore if this decrease in heritability differentially affected regions of the genome, we partitioned the heritability estimates by chromosome using both correction methods. Per chromosome heritability estimates from each method were strongly linearly correlated (Appendix Figure 5.3;  $r^2=0.877$ ;  $p=9.2\times 10^{-11}$ ), and showed no significant per-chromosome differences (Appendix Table 5.3). This suggests that while correcting with PBWT-paint PCs reduces the total heritability estimate, it does not substantially alter the distribution of heritability across the genome. Additionally, following van Rheenen et al (van Rheenen et al. 2016) we explored the relationship between chromosome length and heritability to investigate if the signal of polygenicity in ALS remains after PBWT-paint correction. We found that both SNP PC corrected heritability estimates ( $r^2=0.43$ ,  $p=5.8\times 10^{-4}$ ) and PBWT-paint heritability estimates ( $r^2=0.24$ ,  $p=0.013$ ) show significant linear correlation with chromosome length, providing evidence of polygenicity under both models (Appendix Figure 5.4). Notably while the relationship between SNP heritability and chromosome length is statistically stronger for the SNP PC-corrected estimate, the slopes of these regressions are close to identical (Appendix Figure 5.4), indicating that a similar polygenic signal is present following both corrections.

To further dissect this difference in SNP heritability under the two correction methods we compared the minor allele frequency (MAF) partitioned heritability estimates, from both methods, examining whether any allele frequency bins were more affected by this proposed latent structure (Table 5.5). Notably, while all MAF bins showed a slight decrease in  $h^2_{\text{SNP}}$  when using PBWT-paint PCs as covariates in the model, only the lowest frequency bin considered (MAF=0.01-0.1) was significantly lowered ( $p=0.0042$ ). This could either mean that PBWT-paint selectively overcorrects low-frequency variants, or that heritability is overestimated at low-frequency variants due to uncorrected latent structure missed by SNP PCA (Figures 5.6, 5.10 and 5.11 d). If the latter of these options were true this would have important implications for the study of ALS, given that the current model

driving mass genome wide sequencing of ALS patients is that the disease is mediated by a rare variant architecture (Nicolas et al. 2018).

**Table 5.5: MAF partitioned GREML SNP-based heritability estimates from the 2016 ALS GWAS under SNP and PBWT PC correction.**

MAF	SNP PC correction		PBWT-paint PC correction		p difference
	h <sup>2</sup>	SE	h <sup>2</sup>	SE	
0.01-0.1	2.18E-02	3.63E-03	7.66E-03	3.34E-03	4.19E-03*
0.1-0.2	1.72E-02	3.85E-03	8.04E-03	3.59E-03	8.27E-02
0.2-0.3	2.01E-02	3.79E-03	1.73E-02	3.65E-03	5.93E-01
0.3-0.4	2.32E-03	3.44E-03	2.11E-03	3.33E-03	9.64E-01
0.4-0.5	1.04E-02	3.12E-03	6.03E-03	2.95E-03	3.04E-01

Estimates of SNP heritability for ALS from the 2016 ALS GWAS dataset partitioned into MAF bins, using 20 SNP PCs or 20 PBWT-paint PCs as covariates in the model. Asterisk (\*) denotes a significant difference between the estimates at the Bonferroni corrected level (i.e.  $p < 0.01$ ).

## 5.4 - Discussion

In this chapter we explored the utility of applying haplotype based PCs to detect and correct for population structure in the 2016 GWAS dataset (n=36,052) (van Rheenen et al. 2016), revealing a number of interesting findings which warrant application of this approach to further GWAS datasets. Preliminary work in a small Dutch subset (n=4,753) of this data showed ChromoPainter PCs corrected confounding in GWAS more comprehensively than SNP PCA, while retaining signal at variants with evidence of association with ALS in the remainder of the dataset (n=31,299). This motivated extension of the technique to the full dataset, however computational constraints prohibited use of ChromoPainter on this number of samples. Testing of an alternative haplotype sharing method, PBWT-paint, demonstrated that it performed nearly identically to ChromoPainter in small datasets, while saving significant computational time for large datasets, making application of haplotype sharing methods to the full dataset viable. PCs from the resulting haplotype sharing matrix detected structure at significantly higher resolution in the data than standard SNP PCs. Clusters from this matrix defined by the Louvain community detection method and t-SNE decomposition reveal that this matrix contains sufficient information to separate samples into country level and even finer subpopulations. GWAS summary statistics from the full dataset corrected by PBWT-paint PCs show reduced inflation while retaining the power to identify the same ALS associated loci as SNP PC-corrected summary statistics, indicating the method is more stringent but likely doesn't overcorrect. Finally two methods combining variants across the genome, namely polygenic risk scores and heritability estimates, behaved differently when corrected using PBWT-paint compared to SNP PCA, consistent with predictions of potential bias in these methods from residual population structure. SNP heritability estimates were also lower when corrected with PBWT-paint PCs which could mean SNP PC-corrected estimates were inflated by residual population structure. Strikingly, PRS for ALS calculated with SNP PC corrected summary statistics showed geographic clustering in the Dutch dataset after correcting out the relationship between geography and the phenotype, suggesting residual population structure biases these scores. In contrast, PBWT-paint PC-corrected PRS removed this geographic stratification of PRS, potentially eliminating this bias. While these results are promising there are numerous caveats and future directions that accompany them which will be addressed below.

Results from our initial analysis in the Dutch-only dataset builds on work presented in Chapter 4 characterising the local population structure and demographic history of the Netherlands and indicate that this finescale structure has a measurable effect on

confounding in GWAS (Table 5.1). This could have important implications for single country biobank datasets, such as the UK Biobank (Bycroft et al. 2018) and the Biobank Japan (Nagai et al. 2017), especially given that both countries have been shown to harbour subtle but extensive local structure in smaller datasets using ChromoPainter (Leslie et al. 2015; Takeuchi et al. 2017). In fact, mounting evidence is emerging supporting the existence of residual confounding from uncorrected population structure in GWAS performed on the UK Biobank dataset; after correction using ancestry-informative principal components and even linear mixed models, clear associations have remained between birth location and genetic variants, as well as distributions of PRS (Haworth et al. 2019; Abdellaoui et al. 2019; Cook, Mahajan, and Morris 2020). While haplotype PCs may partially address this issue of residual confounding in large scale biobanks by providing more refined vectors of ancestry it is unclear how well the results observed in the Netherlands will translate to larger single-country datasets due to disparities in sample size. One important reason for this is that as sample size increases, so too does the power to detect the structure present in the sample, even for subtle structure (Lawson et al. 2019), which could mean that unlinked methods such as SNP PCA might perform comparably to haplotype sharing PCs at these scales. This increase in power is apparent when comparing the extensive finescale structure that has been detected in the UK Biobank and Biobank Japan using single marker methods (Diaz-Papkovich et al. 2019; Sakaue et al. 2020) to measures taken in early smaller datasets from these countries. However, it is feasible that even at large sample sizes single markers may remain incapable of detecting certain types of population structure. In fact recent simulations showing SNP PCA from common variants cannot fully correct structure resulting from recent demographic events (Zaidi and Mathieson 2020), which were better corrected by rare variant PCA and IBD PCA. As haplotypes from chromosome painting are expected to capture both old and recent structure this might afford the method an edge even where large sample sizes drastically improve unlinked estimates of population structure. Hence future work is required to properly assess the effects of sample size and demographic history on the relative performance of haplotype sharing and unlinked methods.

While the full dataset differs from the Dutch subset not only in size but in ancestry composition and overall diversity, it is reassuring given the above considerations that haplotype sharing methods still appear to correct more stringently than SNP PCA even in this much larger sample. This suggests that our observations are not simply driven by the low power of SNP PCA to detect population structure in small samples. In fact the difference in LD score intercepts is more pronounced in this larger dataset, achieving significance by our Z test ( $p=1.5\times 10^{-10}$ ) where the Dutch only sample was not significant

( $p=0.12$ ). This may be partially driven by the increased inflation that occurs in larger GWAS samples, though it is more likely driven by increased diversity of the dataset, given that greater differences in allele frequencies are expected across subpopulations from different countries within the dataset. While SNP PCA might be expected to correct most major differences between these countries, multiple sources of recent finescale structure between and within subpopulations may cumulatively account for the large differences in both variance explained in phenotype and estimates of confounding between haplotype sharing and SNP based PCA.

It is also alternatively possible that haplotypic PCA may be overcorrecting disease associated variation here. The two most likely reasons for overcorrection would be if haplotype sharing was directly capturing and penalising patterns of disease-associated variation at a given locus or if causal variants were associated with ancestry (Lawson et al. 2019). The first of these possibilities is mostly ruled out by the similar results produced by the PBWT-paint LOCO analysis (Figure 5.11) which avoids overcorrection at a locus based on variation at that locus by only measuring population structure on the remaining chromosomes. However, it is difficult to measure or prevent overcorrection where ancestry is associated with causal variants, making the second of these possibilities a clear caveat of this analysis. Estimating whether disease is causally associated with true population structure or ancestry (and not simply environmental variation overlaying this) remains a difficult challenge which must be addressed to prevent overcorrection when using any ancestry-informed measure of population structure.

Although our PBWT-paint corrected summary statistics for ALS identify the same hits as present in the SNP corrected method (Figure 5.11), both methods miss four loci identified by the powerful mixed linear model approach applied in the original analysis of the data (van Rheenen et al. 2016) (i.e. *MOBP*, *SCFD1*, *LOC101927815* and *C21orf2*). Associated variants for one of these loci, *C21orf2* were not included in our analysis due to exclusion during QC. In fact these variants initially failed the GWAS QC in the original analysis due to low frequency in a Swedish cohort (van Rheenen et al. 2016). Hence this loss cannot be attributed to overcorrection here but simply exclusion. A second locus *LOC101927815* did not replicate in the replication phase of the original analysis (van Rheenen et al. 2016) indicating it may be a false association, while the remaining two loci *MOBP* and *SCFD1* do not replicate in the larger 2018 ALS GWAS (Nicolas et al. 2018), meaning the veracity of association at these loci is currently under debate pending replication in subsequent larger GWAS. Additionally, variants in *MOBP* showed evidence of sex-specific association (Chapter 2) which means they may only affect ALS risk in a fraction of the dataset, making

them difficult to detect in the full GWAS. Hence it is currently unclear whether the disagreement of our approach and the mixed model approach at these loci represents a real loss of power, or more accurate correction. Notably, the PBWT-paint corrected LD score intercept ( $\text{intercept}_{\text{pbwt-paint}}=1.046$ ; 95% c.i: 1.031-1.06) was also significantly lower ( $p=4.76\times 10^{-8}$ ) than the intercept reported from a mixed linear model analysis ( $\text{Intercept}_{\text{mlma}}=1.1$ ; 95% c.i: 1.08 -1.12 ) (van Rheenen et al. 2016), suggesting that as a trade-off for its improved power compared to our method, the mixed linear model analysis of this dataset is subject to more inflation attributable to confounding genome-wide. Hence, while the PBWT-paint corrected GWAS shows comparable power to detect known (replicated) loci, it also reduces inflation resulting from confounding genome-wide. This may be important for generating unbiased effect sizes for use in polygenic methods such as polygenic risk scores.

The differences in PRS distributions for ALS in the Netherlands, and changes in prediction accuracy when using summary statistics corrected with SNP PCA versus haplotype PCA appear to be partially explained by residual population structure in the SNP PC-corrected scores as evidenced by geographic clustering of these scores (Table 5.3 and Figure 5.12). This is mainly evident for scores calculated with more permissive p-value thresholds, which include more variants with less significant association with the trait. This could suggest that variants less associated with a trait may contribute more bias to polygenic scores than strongly associated variants. However we would caution against extrapolating this result to the general application of PRS without further testing in a range of traits and datasets. Given that ALS is not a highly polygenic trait (see Figure 1.2) it is possible that PRS estimates here are more susceptible to noise from residual confounding in less associated variants than would be expected for highly polygenic traits such as height or schizophrenia, which have risk associated variation spread more evenly genome wide. This is because null variants are expected to be more abundant for ALS GWAS, and the variation in effect sizes (which weight PRS) at these null variants should likely only represent noise and confounding rather than having a true relationship to the disease. Hence the effects of residual stratification on PRS may be more visible in this trait due to its architecture. This motivates testing in other traits of varying polygenicity to assess the effect of trait polygenicity on differences in PRS distributions between methods.

The study design behind the 2016 ALS GWAS is another factor to consider when assessing the degree of residual stratification we observed in SNP-corrected PRS. This dataset combines case-control cohorts sourced from collaborators at many European ALS centres with slightly different sampling practices at each centre (van Rheenen et al. 2016),

which may introduce a source of bias in the GWAS that tracks directly with ancestry. This sampling bias could inflate the level of residual population structure in the training GWAS to levels uncommon for GWAS performed in biobanks or under more standardised sampling schemes, which in turn could translate into abnormally biased PRS. Additionally, as geographic information was only available for a small number of individuals in our target dataset (1,352 of 4,753), and only in the form of current postal address rather than birth location, it is possible that the correlation of PRS with geography we observed is under- or over-estimated in this dataset. The ideal sample for testing how generalisable our findings are would thus be one with deep phenotyping of multiple traits, standardised sampling and rich associated geographic information such as the UK Biobank (Bycroft et al. 2018). Such a dataset would enable us to assess the effects of haplotype-based correction on both prediction from and distribution of PRS in multiple traits with different architectures with standardised sampling schemes. In addition the detailed information on both birthplace and current address would enable us to assess how much the use of current address can inflate or deflate the observed relationship between PRS and geography compared to the more appropriate birthplace, due to migration in the current population.

Following the trend of our PRS scores, our GREML SNP heritability estimate for ALS calculated using PBWT-paint PCs as covariates in place of SNP PCs was also significantly deflated (Table 5.4, Appendix Table 5.3 and Table 5.5). While this could be due to overcorrection as a result of a real interaction between ancestry and disease, or PBWT-paint PCs capturing non-ancestry related disease variation, this estimate approaches univariate LD score regression estimates of heritability from the larger 2018 ALS GWAS (Nicolas et al. 2018) ( $h^2_{\text{SNP}_{2018}}=0.016$ ; 95% c.i: 0.009-0.023), suggesting it may not be an overcorrection. Notably the 2018 ALS GWAS also shows significantly lower LD score intercept ( $\text{Intercept}_{2018}=1.021$ ; 95% c.i: 1.008-1.035 ) than the 2016 dataset corrected with SNP-based PCA ( $p=6.95e-19$ ), indicating that confounding is better controlled in the 2018 GWAS. Combined, these observations support a model where the differences in heritability are likely partially driven by uncorrected population structure in the 2016 ALS dataset (Figure 5.6, 5.10 and 5.11). Hence the lower SNP heritability estimates in PBWT-paint PC corrected GREML analyses could simply be due to better correction of population structure in the dataset. The potential inflation of SNP heritability due to latent structure in the data has implications for the analysis in Chapter 2 which relied extensively on the use of SNP based covariates to correct population structure when estimating male specific and female specific heritability.

Partitioning heritability estimates per chromosome revealed a strong linear relationship between estimates from both methods (Appendix Figure 5.3) and no significant differences in per-chromosome heritability (Appendix Table 5.3). This indicates that while total heritability is reduced when using PBWT-paint PC correction, this reduction is spread relatively uniformly across the genome. Additionally per chromosome heritability estimates from both methods showed significant positive linear relationships with chromosome length with comparable slopes, supporting the model that ALS is polygenic even with this more stringent correction (Appendix Figure 5.4). The similar slopes in this regression indicate that the polygenic signal is preserved when correcting with PBWT-paint PCs, suggesting that it does not overcorrect.

However, when considering MAF partitioned estimates of  $h^2_{\text{SNP}}$ , a significant decline was seen for the lowest frequency variants (Table 5.5), seeding doubt on the observation that low frequency variants contribute the most to ALS heritability (van Rheenen et al. 2016). If true this could have important implications for study design in ALS, as this observation has partially motivated a large uptake of whole genome sequencing of ALS patients in consortia such as Project MinE (van Rheenen et al. 2018), with the aim of identifying causal variants in the lower frequency spectrum. However, given the low SNP heritability of ALS under any model considered here, it is still probable that SNP chip technology does not adequately capture genetic variation associated with ALS. Furthermore exome studies have since identified associated rare variants associated with ALS in genes including *NEK1* (Kenna et al. 2016) and *KIF5A* (Nicolas et al. 2018) with far smaller sample sizes than this GWAS, suggesting that this approach remains valid regardless of whether this initial observation supporting a rare variant architecture holds up to scrutiny.

In addition to our exploration of correcting residual population structure with haplotype based PCs, this chapter also demonstrates that finescale population genetic analyses such as those seen in Chapters 3 and 4 are scalable to large multi-population datasets. We demonstrated the power of applying the Louvain clustering method to both small and large scale haplotype sharing datasets as an alternative to conventional methods such as fineSTRUCTURE. This method produced similar splits to fineSTRUCTURE in our Dutch coancestry matrix from Chapter 4 with the comparable mean  $F_{\text{ST}}$ , suggesting that it is a viable alternative clustering method for identifying homogeneous subgroups in haplotype sharing datasets. The major advantage of this method is that it can be applied to massive datasets which are prohibitively computationally expensive to analyse with fineSTRUCTURE. Iterative application of this approach can also yield finer and finer splits in the data, allowing the researcher to choose the resolution of structure they wish to



investigate (presumably limited by the amount of structure contained in the coancestry matrix). To this end, the method was capable of identifying sub-clusters with overlapping but distinct geographic ranges when applied to previously “indivisible” cluster in the southeast of England (Figure 5.9).  $F_{ST}$  values between these subclusters were almost an order of magnitude smaller than those between fineSTRUCTURE clusters identified in Chapters 3 and 4, suggesting this method can describe extremely finescale population structure. This algorithm should enable efficient clustering in haplotype sharing datasets from large GWAS datasets, greatly expanding the sample sizes viable for finescale population genetic analyses such as those carried out in Chapters 3 and 4. Combined with geospatial data such as that available for large biobank-scale datasets, PBWT-paint and Louvain clustering may yield important insights into recent finescale population structure underrepresented in smaller single-country datasets. In addition it should enable us to explore finescale links between countries, as seen in the clustering of Belgium and the Netherlands, or samples from the US and European groups in this chapter. Simultaneous analysis of haplotype sharing in large multi-country datasets may yield insights missed when projecting structure from an external reference panel into a small single country dataset as seen in Chapters 3 and 4. Hence our analysis highlights potential advances for both genetic epidemiology and the study of modern population genetics which are likely to yield interesting results for both fields when applied to large modern GWAS datasets.

## Chapter 6 - Discussion

This thesis describes work carried out across several distinct but complementary projects as described in Chapter 1 with the global aim of improving our understanding of the complex genetics of ALS and how residual finescale population structure impacts commonly used GWAS analysis techniques. Briefly, this work studied: the potential genetic overlap between ALS and comorbid cognitive and psychiatric traits (Chapter 2); the role of sex in ALS genetics (Chapter 2); the extent of finescale genetic structure and historical inferences from ALS GWAS data sampled across Ireland (Chapter 3) and The Netherlands (Chapter 4); and finally the application of haplotype sharing methods to correct for the observed finescale population structure seen in these countries and others in large GWAS (Chapter 5). This chapter will discuss the future directions for study that these analyses open up.

### 6.1 - Future directions

The results from this thesis, alongside the ideas and methods explored within, point towards several important avenues for future research.

#### 6.1.1 - Replication and functional validation of putative ALS loci

The putative ALS loci identified in our sex specific scans and pleiotropic analysis (cFDR analysis) from Chapter 2 appear plausible based on their enrichment in disease related functional categories and tissue of expression. However, to be fully certain of their association with ALS these loci require replication in an independent cohort.

For the both sex specific loci and the pleiotropic loci, independent samples from the 2018 GWAS (Nicolas et al. 2018) may be an appealing sample for the replication stage.

However, due to significant sample overlap with the 2016 ALS GWAS used in our study, exclusion of samples would lead to relatively small replication cohorts. Indeed, based on table s1 from the 2018 ALS GWAS (Nicolas et al. 2018) this would leave us with a total of 8,229 ALS cases for the replication GWAS (3,433 female and 4,796 male cases), which may prohibitively reduce power for these analyses especially when split into male and female strata. Additionally, as the majority of controls from the 2018 ALS GWAS were gathered from 18 distinct studies (see table s1 from Nicolas et al. (Nicolas et al. 2018)), and show significant bias towards females (~69% female controls) gathering an unbiased control set for these replications would not be trivial. Instead, a more viable and powerful option might be to rerun both the sex-specific and multi-trait analysis in data from the

much larger upcoming Project MinE GWAS for ALS (n~150,000, Jan Veldink and Wouter van Rheenen, personal communication), using the data from Chapter 2 as a replication cohort. This approach would afford far greater power to identify robustly associated sex-specific and pleiotropic variants associated with ALS and potentially be more efficient considering the limitations of both the 2016 and 2018 ALS GWAS datasets. Hence, while our multi-trait and sex-specific analyses have revealed novel loci with potential roles in ALS, they should at present be treated as exploratory results. Finally, following replication and fine-mapping of variants, further functional work, for example in mouse models or induced pluripotent stem cell-derived motor neurones, is required to validate the role in ALS pathology.

#### 6.1.2 - Updating multi-trait analysis for ALS in the face of growing GWAS datasets

Analysis of genetic overlap between ALS and secondary psychiatric and cognitive traits in Chapter 2 of this thesis were performed on recent sets of GWAS summary statistics. However, while this work outlines interesting findings regarding the shared genetic components of these traits, datasets are constantly growing with larger GWAS releases occurring regularly for each of these traits, meaning this analysis is not the final picture. To illustrate this point, at the time of writing (13th of September 2020), the psychiatric genetics consortium has just released summary statistics for a massive GWAS of schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. 2020) studying 69,369 patients with schizophrenia and 236,642 controls. As these GWAS datasets grow, their increasing sample size leads to more robust estimates of SNP effect sizes with smaller errors, which will in turn increase the power of multi-trait analyses including these traits. Indeed, with an increase in sample size and power, we might expect additional psychiatric traits to show significant correlations with ALS, as seen for bipolar disorder in our study. Given that the projected sample size for ALS GWAS face greater limitations than more common traits due to the relatively low number of patients presenting with the disease at a given year, future studies informed by pleiotropic overlap of ALS and with larger GWAS from genetically correlated traits may become an important resource for boosting power to discover new ALS loci.

#### 6.1.3 - Towards greater diversity in the study of ALS genetics

Analyses in this thesis are unfortunately restricted to European datasets and hence are not representative of the genetic architecture of ALS in the global population, which is a limitation both for the study of ALS and the field of GWAS in general (Need and Goldstein 2009; Popejoy and Fullerton 2016). This lack of diversity not only creates a biased view of the genetic variation behind traits, which causes findings to translate poorly to individuals

of non-European ancestry (e.g. for PRS (A. R. Martin et al. 2017; Duncan et al. 2019)), but also lowers power for detecting causal variants which are at low frequency in European populations (Wojcik et al. 2019), limiting overall discovery of trait-relevant variation. In addition, fine-mapping has been shown to be more sensitive in mixed ancestry cohorts due to different LD blocks reducing the search space for the causal variant, improving detection of causal variants (Wojcik et al. 2019). Aside from these general advantages of diverse GWAS, there is mounting evidence that the prevalence of certain known ALS genes may differ across ancestries. For example, while the *C9orf72* hexanucleotide repeat expansion is the most common known genetic cause of ALS in Europeans, it is rare in China, where the most common mutation is instead *SOD1* (Xiaolu Liu et al. 2018). Hence diversity is also an important factor for proper epidemiological characterisation of ALS risk worldwide. Thus a clear future direction would be to move towards greater diversity in ALS GWAS and other genetic studies, which may be achieved through expanding collaboration with ALS centres worldwide.

Initial work in this direction meta-analysing ALS cases and controls from China (1,234 cases and 2,850 controls) and Europe (2016 ALS GWAS (van Rheenen et al. 2016)) has shown some promising returns (Benyamin et al. 2017). Most notably, this cross-ethnic meta-analysis identified an association at the *TNIP1-GPX3* locus (Benyamin et al. 2017), which had prior functional evidence of involvement in ALS, and has since been identified in larger GWAS (Nicolas et al. 2018). The lead SNP showed higher frequency in the Chinese cohort than the European cohort in this analysis, which may have improved power to detect this association. However, to date there are no further multi-ethnic ALS GWAS studies, and very little study of ALS genetics has been done outside of Europe and east Asia. Thus substantial further work dedicated to the inclusion of more diverse ancestry in ALS GWAS may yield even further returns and remedy European bias in resulting research.

#### 6.1.4 - Evaluating correction of confounding using haplotype sharing PCs in other traits and simulation studies

The findings in Chapter 5 suggest that haplotype sharing PCs reduce bias resulting from latent population structure in GWAS and other analysis such as PRS and heritability estimation. However these findings are based on a single empirical dataset (the 2016 ALS GWAS) which means that they should be treated with due caution until replicated in other traits and simulated datasets. Given the rich phenotyping data for multiple complex traits and associated geographic data available for samples in the large UK Biobank (Bycroft et

al. 2018) this may be a suitable dataset for further evaluating the method. Furthermore, as there is emerging evidence of potentially uncorrected recent structure biasing both single marker associations and PRS in this dataset (Haworth et al. 2019), which both show a residual association with sample location, this dataset presents both a metric to assess performance of our method and a problem in need of solving. Should this method prove useful in this setting, the singular coancestry matrix generated from such a study could become a widely used resource for all researchers studying the many traits within the UK Biobank.

However, there are limitations to testing this method in empirical datasets, namely that values such as trait heritability or true associated loci are unknown meaning we cannot conclusively separate useful correction from over-correction in some cases. Hence simulation studies where we can control which variants are associated with the trait, what the level of trait heritability is, how trait-associated variants are distributed across the genome and how recently the population structure emerged would be invaluable for evaluating whether the method comes with any of its own biases. Knowing which variants are associated would allow us to assess the false positive and false negative rate under the various trait architectures, and demographic models studied. Similarly knowledge of trait heritability would enable us to conclusively assess how well this method deals with bias from population structure in heritability estimation. Finally being able to test how well this method performs in data with differing levels and types population structure (i.e. recent vs old; admixed vs not admixed) would be crucial for understanding how applicable it is to non-European populations. Developing such a detailed simulation combining both realistic population structure and trait genetic architecture is a non-trivial exercise, however a similar set of simulations to those carried out by Zaidi et al. (Zaidi and Mathieson 2020) may prove a useful starting point.

#### 6.1.5 - Application of haplotype sharing methods to rare variant association studies

In Chapters 3-5 we showed that haplotype sharing PCs can detect local population structure more sensitively than SNP PCs calculated from unlinked variants, and perform better at correcting latent population structure affecting the common variants studied in GWAS. However the question remains as to how well these PCs correct confounding in rare variants untyped by standard SNP arrays, which have been shown to suffer from differential stratification to common variants where environmental risk is unevenly distributed in simulation studies (Mathieson and McVean 2012). This is an important question as it is possible that much of the remaining differences between true narrow-sense heritability and SNP-based heritability estimates lies in untyped rare variants, as

suggested by empirical studies of height using WGS data (Wainschtein et al. 2019). While this empirical study argues that its results are unbiased due to inclusion of large numbers of principal components, recent simulations have shown that other methods estimating heritability concentrated in rare variants using IBD as a proxy (IBD-GREML) are extremely sensitive to population stratification (Evans et al. 2018). Thus correcting for stratification of rare variants confounding both genome wide association studies using rare variation and indeed heritability estimates considering rarer variants may be crucial to further understanding the genetic roots of many traits. Promisingly, a study of IBD sharing in Finland has shown a clear relationship between haplotype sharing and variants of a rarer allele frequency (A. R. Martin et al. 2018), suggesting that haplotype sharing tracks well with rare variant sharing. It follows that PCs based on a haplotype sharing across all samples could suitably describe the population structure affecting rare variants, while also correcting common variants as seen above. This proposed method, however, comes with the caveat of overcorrection as these haplotype PCs could theoretically correct out patterns of disease rare variant sharing associated with the trait being studied. Future work assessing the suitability of this correction method in datasets with suitably rare variation is thus warranted.

#### 6.1.6 - Expanding our understanding of human history using large scale GWAS data

Analyses in Chapters 3 and 4 applied several population genetics techniques to explore the genetic variation in modern individuals from Ireland and the Netherlands, evaluating local structure, signatures of historical admixture, changes in structure over time and changes in demography. These datasets were both repurposed from a much larger GWAS dataset, yet many of their findings were consistent with other studies from these countries (e.g. Gilbert et al. (Gilbert et al. 2017) and Abdellaoui et al. (Abdellaoui, Hottenga, de Knijff, et al. 2013)), suggesting that ascertainment hasn't massively impacted the usefulness of this data to answer questions about how the past has shaped modern genomes. This suggests that the large modern GWAS datasets collected to study a range of traits may be similarly repurposed to uncover more features of human history. Indeed, many of the techniques used in these chapters are coming of age to be used effectively in extremely large datasets: a fast efficient implementation of the GLOBETROTTER method developed by Hellenthal et al. (Hellenthal et al. 2014) to detect and date human admixtures is undergoing development (Wangkumhang 2020); fast scalable IBD-calling methods using the PBWT algorithm have been released (e.g. RaPID (Naseri et al. 2019)), which could be used in our length-binned IBD approach from Chapter 5; and our pipeline substituting PBWT-paint and Louvain community detection for

the ChromoPainter/fineSTRUCTURE pipeline (see Chapter 5) greatly improves computational efficiency in detecting population subgroups from large datasets. Future work applying these and related methods to large GWAS datasets will likely yield vast insights into historical events (e.g. the impact of the Black Death in other countries) shaping modern individuals.

## 6.2 - Concluding remark

In conclusion this thesis has deepened our knowledge of the complex genetics of ALS, the population genetics of Ireland and the Netherlands, and finally the extent and impact of finescale population structure on both GWAS and downstream analyses. While our work on the genetic architecture of ALS in the first results chapter provides novel insights into the genetic overlap between ALS and secondary psychiatric and cognitive traits, and demonstrates differences in the genetic architecture of ALS across sexes, it is reliant on the robustness of modern GWAS analysis to confounding from latent factors. Indeed, work in subsequent chapters (Chapters 3 and 4) highlights the extent to which a potential confounder, namely finescale population structure, pervades this GWAS data. Given that this structure is poorly detected by currently used methods such as SNP PCA, these findings have strong implications for GWAS study design in general. To address this concern we apply the methods built to detect this finescale structure to correct GWAS analysis in the final chapter, demonstrating clear reductions in statistical inflation, with little loss of power. Thus, in addition to novel findings regarding genetics of ALS and the population history of Ireland and the Netherlands, this thesis contributes an important methodological advance which may prove crucial to the robust study of complex traits.



## References

- Abdellaoui, Abdel, Jouke-Jan Hottenga, Peter de Knijff, Michel G. Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, et al. 2013. "Population Structure, Migration, and Diversifying Selection in the Netherlands." *European Journal of Human Genetics: EJHG* 21 (11): 1277–85.
- Abdellaoui, Abdel, Jouke-Jan Hottenga, Xiangjun Xiao, Paul Scheet, Erik A. Ehli, Gareth E. Davies, James J. Hudziak, et al. 2013. "Association between Autozygosity and Major Depression: Stratification Due to Religious Assortment." *Behavior Genetics* 43 (6): 455–67.
- Abdellaoui, Abdel, David Hugh-Jones, Loic Yengo, Kathryn E. Kemper, Michel G. Nivard, Laura Veul, Yan Holtz, et al. 2019. "Genetic Correlates of Social Stratification in Great Britain." *Nature Human Behaviour* 3 (12): 1332–42.
- Al-Asadi, Hussein, Desislava Petkova, Matthew Stephens, and John Novembre. 2019. "Estimating Recent Migration and Population-Size Surfaces." *PLoS Genetics* 15 (1): e1007908.
- Al-Chalabi, Ammar, Andrea Calvo, Adriano Chio, Shuna Colville, Cathy M. Ellis, Orla Hardiman, Mark Heverin, et al. 2014. "Analysis of Amyotrophic Lateral Sclerosis as a Multistep Process: A Population-Based Modelling Study." *Lancet Neurology* 13 (11): 1108–13.
- Al-Chalabi, Ammar, F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye, and F. Rijdsdijk. 2010. "An Estimate of Amyotrophic Lateral Sclerosis Heritability Using Twin Data." *Journal of Neurology, Neurosurgery, and Psychiatry* 81 (12): 1324–26.
- Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64.
- Andreassen, Ole A., Wesley K. Thompson, Andrew J. Schork, Stephan Ripke, Morten Mattingsdal, John R. Kelsoe, Kenneth S. Kendler, et al. 2013. "Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate." *PLoS Genetics* 9 (4): e1003455.
- Athanasiadis, Georgios, Jade Y. Cheng, Bjarni J. Vilhjálmsson, Frank G. Jørgensen, Thomas D. Als, Stephanie Le Hellard, Thomas Espeseth, et al. 2016. "Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity." *Genetics* 204 (2): 711–22.
- Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.
- Badano, Jose L., and Nicholas Katsanis. 2002. "Beyond Mendel: An Evolving View of Human Genetic Disease Transmission." *Nature Reviews. Genetics* 3 (10): 779–89.
- Bandres-Ciga, Sara, Alastair J. Noyce, Gibran Hemani, Aude Nicolas, Andrea Calvo, Gabriele Mora, ITALSGEN Consortium, et al. 2019. "Shared Polygenic Risk and Causal Inferences in Amyotrophic Lateral Sclerosis." *Annals of Neurology* 85 (4): 470–81.
- Bayliss, Alex, and Peter Woodman. 2009. "A New Bayesian Chronology for Mesolithic Occupation at Mount Sandel, Northern Ireland." *Proceedings of the Prehistoric Society* 75: 101–23.
- Bayot, Aurélien, Sacha Reichman, Sophie Lebon, Zsolt Csaba, Laetitia Aubry, Ghislaine Sterkers, Isabelle Husson, Malgorzata Rak, and Pierre Rustin. 2013. "Cis-Silencing of PIP5K1B Evidenced in Friedreich's Ataxia Patient Cells Results in Cytoskeleton Anomalies." *Human Molecular Genetics* 22 (14): 2894–2904.
- Bayot, Aurélien, and Pierre Rustin. 2013. "Friedreich's Ataxia, Frataxin, PIP5K1B: Echo of a Distant Fracas." *Oxidative Medicine and Cellular Longevity* 2013 (September):

725635.

- Beeldman, Emma, Joost Raaphorst, Michelle Klein Twennaar, Marianne de Visser, Ben A. Schmand, and Rob J. de Haan. 2016. "The Cognitive Profile of ALS: A Systematic Review and Meta-Analysis Update." *Journal of Neurology, Neurosurgery, and Psychiatry* 87 (6): 611–19.
- Benyamin, Beben, Ji He, Qiongyi Zhao, Jacob Gratten, Fleur Garton, Paul J. Leo, Zhijun Liu, et al. 2017. "Cross-Ethnic Meta-Analysis Identifies Association of the GPX3-TNIP1 Locus with Amyotrophic Lateral Sclerosis." *Nature Communications* 8 (1): 611.
- Berg, Jeremy J., Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, et al. 2019. "Reduced Signal for Polygenic Adaptation of Height in UK Biobank." *ELife* 8 (March). <https://doi.org/10.7554/eLife.39725>.
- Berisa, Tomaz, and Joseph K. Pickrell. 2016. "Approximately Independent Linkage Disequilibrium Blocks in Human Populations." *Bioinformatics* 32 (2): 283–85.
- Bhatia, Gaurav, Nicholas A. Furlotte, Po-Ru Loh, Xuanyao Liu, Hilary Kiyu Finucane, Alexander Gusev, and Alkes Price. 2016. "Correcting Subtle Stratification in Summary Association Statistics." *BioRxiv*. <https://doi.org/10.1101/076133>.
- Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics* 2008 (10): P10008.
- Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86.
- Broce, Iris J., Chun C. Fan, Nicholas T. Olney, Catherine Lomen-Hoerth, Steve Finkbeiner, Nazem Atassi, Merit E. Cudkowicz, et al. 2018. "Partitioning the Genetic Architecture of Amyotrophic Lateral Sclerosis." *BioRxiv*. <https://doi.org/10.1101/505693>.
- Browning, Brian L., and Sharon R. Browning. 2011. "A Fast, Powerful Method for Detecting Identity by Descent." *American Journal of Human Genetics* 88 (2): 173–82.
- . 2013a. "Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data." *Genetics* 194 (2): 459–71.
- . 2013b. "Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data." *American Journal of Human Genetics* 93 (5): 840–51.
- Browning, Sharon R., and Brian L. Browning. 2007. "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering." *American Journal of Human Genetics* 81 (5): 1084–97.
- . 2015. "Accurate Non-Parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent." *American Journal of Human Genetics* 97 (3): 404–18.
- Browning, Sharon R., Brian L. Browning, Martha L. Daviglus, Ramon A. Durazo-Arvizu, Neil Schneiderman, Robert C. Kaplan, and Cathy C. Laurie. 2018. "Ancestry-Specific Recent Effective Population Size in the Americas." *PLoS Genetics* 14 (5): e1007385.
- Bulik-Sullivan, Brendan, Hilary K. Finucane, Verner Anttila, Alexander Gusev, Felix R. Day, Po-Ru Loh, ReproGen Consortium, et al. 2015. "An Atlas of Genetic Correlations across Human Diseases and Traits." *Nature Genetics* 47 (11): 1236–41.
- Bulik-Sullivan, Brendan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." *Nature Genetics* 47 (3): 291–95.
- Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James

- Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12.
- Bycroft, Clare, Ceres Fernandez-Rozadilla, Clara Ruiz-Ponte, Inés Quintela, Ángel Carracedo, Peter Donnelly, and Simon Myers. 2019. "Patterns of Genetic Differentiation and the Footprints of Historical Migrations in the Iberian Peninsula." *Nature Communications* 10 (1): 551.
- Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9.
- Byrne, Ross P., Rui Martiniano, Lara M. Cassidy, Matthew Carrigan, Garrett Hellenthal, Orla Hardiman, Daniel G. Bradley, and Russell L. McLaughlin. 2018. "Insular Celtic Population Structure and Genomic Footprints of Migration." *PLoS Genetics* 14 (1): e1007152.
- Byrne, Ross P., Wouter van Rheenen, Leonard H. van den Berg, Jan H. Veldink, Russell L. McLaughlin, and Project MinE ALS GWAS Consortium. 2020. "Dutch Population Structure across Space, Time and GWAS Design." *Nature Communications* 11 (1): 4556.
- Byrne, Susan, Mark Heverin, Marwa Elamin, Peter Bede, Catherine Lynch, Kevin Kenna, Russell MacLaughlin, Cathal Walsh, Ammar Al Chalabi, and Orla Hardiman. 2013. "Aggregation of Neurologic and Neuropsychiatric Disease in Amyotrophic Lateral Sclerosis Kindreds: A Population-Based Case-Control Cohort Study of Familial and Sporadic Amyotrophic Lateral Sclerosis." *Annals of Neurology* 74 (5): 699–708.
- Campbell, Catarina D., Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, Leif C. Groop, David Altshuler, Kristin G. Ardlie, and Joel N. Hirschhorn. 2005. "Demonstrating Stratification in a European American Population." *Nature Genetics* 37 (8): 868–72.
- Carter, C. O., and K. A. Evans. 1969. "Inheritance of Congenital Pyloric Stenosis." *Journal of Medical Genetics* 6 (3): 233–54.
- Caspi, Avshalom, Renate M. Houts, Daniel W. Belsky, Sidra J. Goldman-Mellor, Honalee Harrington, Salomon Israel, Madeline H. Meier, et al. 2014. "The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders?" *Clinical Psychological Science* 2 (2): 119–37.
- Cassidy, Lara M., Rui Martiniano, Eileen M. Murphy, Matthew D. Teasdale, James Mallory, Barrie Hartwell, and Daniel G. Bradley. 2016. "Neolithic and Bronze Age Migration to Ireland and Establishment of the Insular Atlantic Genome." *Proceedings of the National Academy of Sciences* 113 (2): 368–73.
- Chacón-Duque, Juan-Camilo, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuña-Alonzo, Rodrigo Barquera, Mirsha Quinto-Sánchez, et al. 2018. "Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance." *Nature Communications* 9 (1): 5388.
- Chahrour, Maria, Sung Yun Jung, Chad Shaw, Xiaobo Zhou, Stephen T. C. Wong, Jun Qin, and Huda Y. Zoghbi. 2008. "MeCP2, a Key Contributor to Neurological Disease, Activates and Represses Transcription." *Science* 320 (5880): 1224–29.
- Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (February): 7.
- Chia, Ruth, Adriano Chiò, and Bryan J. Traynor. 2018. "Novel Genes Associated with Amyotrophic Lateral Sclerosis: Diagnostic and Clinical Implications." *Lancet Neurology* 17 (1): 94–102.
- Chiò, Adriano, A. Ilardi, S. Cammarosano, C. Moglia, A. Montuschi, and A. Calvo. 2012. "Neurobehavioral Dysfunction in ALS Has a Negative Effect on Outcome and Use of PEG and NIV." *Neurology* 78 (14): 1085–89.

- Chiò, Adriano, Letizia Mazzini, Sandra D'Alfonso, Lucia Corrado, Antonio Canosa, Cristina Moglia, Umberto Manera, et al. 2018. "The Multistep Hypothesis of ALS Revisited: The Role of Genetic Mutations." *Neurology* 91 (7): e635–42.
- Coady, Tristan H., and James L. Manley. 2015. "ALS Mutations in TLS/FUS Disrupt Target Gene Expression." *Genes & Development* 29 (16): 1696–1706.
- Commission, Eurostat-European, and Others. 2011. "Regions in the European Union. Nomenclature of Territorial Units for Statistics." NUTS 2010/EU-27. Luxembourg: Publications Office of the European Union.
- Cook, James P., Anubha Mahajan, and Andrew P. Morris. 2020. "Fine-Scale Population Structure in the UK Biobank: Implications for Genome-Wide Association Studies." *Human Molecular Genetics*, July. <https://doi.org/10.1093/hmg/ddaa157>.
- CoreTeam, R. 2015. "R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015."
- Cronin, Simon, Stephen Berger, Jinhui Ding, Jennifer C. Schymick, Nicole Washecka, Dena G. Hernandez, Matthew J. Greenway, Daniel G. Bradley, Bryan J. Traynor, and Orla Hardiman. 2008. "A Genome-Wide Association Study of Sporadic ALS in a Homogenous Irish Population." *Human Molecular Genetics* 17 (5): 768–74.
- Csardi, Gabor, Tamas Nepusz, and Others. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal, Complex Systems* 1695 (5): 1–9.
- Dandine-Roulland, Claire, Céline Bellenguez, Stéphanie Debette, Philippe Amouyel, Emmanuelle Génin, and Hervé Perdry. 2016. "Accuracy of Heritability Estimations in Presence of Hidden Population Stratification." *Scientific Reports* 6 (May): 26471.
- Davies, Gail, Max Lam, Sarah E. Harris, Joey W. Trampush, Michelle Luciano, W. David Hill, Saskia P. Hagenaars, et al. 2018. "Study of 300,486 Individuals Identifies 148 Independent Genetic Loci Influencing General Cognitive Function." *Nature Communications* 9 (1): 2098.
- Deacon, Bernard. 2007. *A Concise History of Cornwall*. University of Wales Press-Hi.
- Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. 2011. "A Linear Complexity Phasing Method for Thousands of Genomes." *Nature Methods* 9 (2): 179–81.
- Demontis, Ditte, Raymond K. Walters, Joanna Martin, Manuel Mattheisen, Thomas D. Als, Esben Agerbo, Gísli Baldursson, et al. 2019. "Discovery of the First Genome-Wide Significant Risk Loci for Attention Deficit/Hyperactivity Disorder." *Nature Genetics* 51 (1): 63–75.
- Devlin, B., and K. Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004.
- Devlin, B., K. Roeder, and L. Wasserman. 2001. "Genomic Control, a New Approach to Genetic-Based Association Studies." *Theoretical Population Biology* 60 (3): 155–66.
- Diaz-Papkovich, Alex, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. 2019. "UMAP Reveals Cryptic Population Structure and Phenotype Heterogeneity in Large Genomic Cohorts." *PLoS Genetics* 15 (11): e1008432.
- Duffy, Seán. 2000. *The Concise History of Ireland*. Gill & Macmillan.
- . 2012. *Atlas of Irish History*. Gill & Macmillan.
- Duncan, L. E., A. Ratanatharathorn, A. E. Aiello, L. M. Almli, A. B. Amstadter, A. E. Ashley-Koch, D. G. Baker, et al. 2018. "Largest GWAS of PTSD (N=20 070) Yields Genetic Overlap with Schizophrenia and Sex Differences in Heritability." *Molecular Psychiatry* 23 (3): 666–73.
- Duncan, L. E., H. Shen, B. Gelaye, J. Meijssen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. 2019. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications* 10 (1): 3328.
- Durbin, Richard. 2014. "Efficient Haplotype Matching and Storage Using the Positional Burrows-Wheeler Transform (PBWT)." *Bioinformatics* 30 (9): 1266–72.
- Eijk, Ruben P. A. van, Ashley R. Jones, William Sproviero, Aleksey Shatunov, Pamela J. Shaw, P. Nigel Leigh, Carolyn A. Young, et al. 2017. "Meta-Analysis of

- Pharmacogenetic Interactions in Amyotrophic Lateral Sclerosis Clinical Trials.” *Neurology* 89 (18): 1915–22.
- Es, Michael A. van, Orla Hardiman, Adriano Chio, Ammar Al-Chalabi, R. Jeroen Pasterkamp, Jan H. Veldink, and Leonard H. van den Berg. 2017. “Amyotrophic Lateral Sclerosis.” *The Lancet* 390 (10107): 2084–98.
- Euesden, Jack, Cathryn M. Lewis, and Paul F. O’Reilly. 2015. “PRSice: Polygenic Risk Score Software.” *Bioinformatics* 31 (9): 1466–68.
- Evans, Luke M., Rasool Tahmasbi, Matt Jones, Scott I. Vrieze, Gonçalo R. Abecasis, Sayantan Das, Douglas W. Bjelland, et al. 2018. “Narrow-Sense Heritability Estimation of Complex Traits Using Identity-by-Descent Information.” *Heredity* 121 (6): 616–30.
- Fadista, João, Alisa K. Manning, Jose C. Florez, and Leif Groop. 2016. “The (in)Famous GWAS P-Value Threshold Revisited and Updated for Low-Frequency Variants.” *European Journal of Human Genetics: EJHG* 24 (8): 1202–5.
- Falconer, D. S. 1967. “The Inheritance of Liability to Diseases with Variable Age of Onset, with Particular Reference to Diabetes Mellitus.” *Annals of Human Genetics* 31 (1): 1–20.
- Fang, Fang, Freya Kamel, Paul Lichtenstein, Rino Bellocco, Pär Sparén, Dale P. Sandler, and Weimin Ye. 2009. “Familial Aggregation of Amyotrophic Lateral Sclerosis.” *Annals of Neurology* 66 (1): 94–99.
- Felsenstein, J. 1971. “Inbreeding and Variance Effective Numbers in Populations with Overlapping Generations.” *Genetics* 68 (4): 581–97.
- Finucane, Hilary K., Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verner Anttila, et al. 2015. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics.” *Nature Genetics* 47 (11): 1228–35.
- Finucane, Hilary K., Yakir A. Reshef, Verner Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. “Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types.” *Nature Genetics* 50 (4): 621–29.
- Fisher, R. A. 1918. “The Correlation Between Relatives on the Supposition of Mendelian Inheritance.” *Proceedings of the Royal Society Edinburgh* 52.  
<https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15097/1/9.pdf>.
- Gazal, Steven, Hilary K. Finucane, Nicholas A. Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, et al. 2017. “Linkage Disequilibrium-Dependent Architecture of Human Complex Traits Shows Action of Negative Selection.” *Nature Genetics* 49 (10): 1421–27.
- Genome of the Netherlands Consortium. 2014. “Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population.” *Nature Genetics* 46 (8): 818–25.
- Gilbert, Edmund, Seamus O’Reilly, Michael Merrigan, Darren McGettigan, Anne M. Molloy, Lawrence C. Brody, Walter Bodmer, et al. 2017. “The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland.” *Scientific Reports* 7 (1): 17199.
- Gilbert, Edmund, Seamus O’Reilly, Michael Merrigan, Darren McGettigan, Veronique Vitart, Peter K. Joshi, David W. Clark, et al. 2019. “The Genetic Landscape of Scotland and the Isles.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (38): 19064–70.
- Gonzales, Michael L., and Janine M. LaSalle. 2010. “The Role of MeCP2 in Brain Development and Neurodevelopmental Disorders.” *Current Psychiatry Reports* 12 (2): 127–34.
- Gratten, Jacob, Naomi R. Wray, Matthew C. Keller, and Peter M. Visscher. 2014. “Large-Scale Genomics Unveils the Genetic Architecture of Psychiatric Disorders.” *Nature Neuroscience* 17 (6): 782–90.
- Grotzinger, Andrew D., Mijke Rhemtulla, Ronald de Vlaming, Stuart J. Ritchie, Travis T.

- Mallard, W. David Hill, Hill F. Ip, et al. 2019. "Genomic Structural Equation Modelling Provides Insights into the Multivariate Genetic Architecture of Complex Traits." *Nature Human Behaviour* 3 (5): 513–25.
- Gusev, Alexander, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. 2009. "Whole Population, Genome-Wide Mapping of Hidden Relatedness." *Genome Research* 19 (2): 318–26.
- Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522 (7555): 207–11.
- Hagenaars, Saskia P., Ratko Radaković, Christopher Crockford, Chloe Fawns-Ritchie, International FTD-Genomics Consortium (IFGC), Sarah E. Harris, Catharine R. Gale, and Ian J. Deary. 2018. "Genetic Risk for Neurodegenerative Disorders, and Its Overlap with Cognitive Ability and Physical Function." *PloS One* 13 (6): e0198187.
- Han, Eunjung, Peter Carbonetto, Ross E. Curtis, Yong Wang, Julie M. Granka, Jake Byrnes, Keith Noto, et al. 2017. "Clustering of 770,000 Genomes Reveals Post-Colonial Population Structure of North America." *Nature Communications* 8 (February): 14238.
- Hardiman, Orla, Leonard H. van den Berg, and Matthew C. Kiernan. 2011. "Clinical Diagnosis and Management of Amyotrophic Lateral Sclerosis." *Nature Reviews. Neurology* 7 (11): 639–49.
- Haworth, Simon, Ruth Mitchell, Laura Corbin, Kaitlin H. Wade, Tom Dudding, Ashley Budu-Aggrey, David Carslake, et al. 2019. "Apparent Latent Structure within the UK Biobank Sample Has Implications for Epidemiological Analysis." *Nature Communications* 10 (1): 333.
- Hellenthal, Garrett, George B. J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. 2014. "A Genetic Atlas of Human Admixture History." *Science* 343 (6172): 747–751.
- Hellwege, Jacklyn N., Jacob M. Keaton, Ayush Giri, Xiaoyi Gao, Digna R. Velez Edwards, and Todd L. Edwards. 2017. "Population Stratification in Genetic Association Studies." *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 95 (October): 1.22.1-1.22.23.
- Herlihy, David. 1997. *The Black Death and the Transformation of the West*. Harvard University Press.
- Hill, E. W., M. A. Jobling, and D. G. Bradley. 2000. "Y-Chromosome Variation and Irish Origins." *Nature* 404 (6776): 351–52.
- Holsinger, Kent E., and Bruce S. Weir. 2009. "Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting F(ST)." *Nature Reviews. Genetics* 10 (9): 639–50.
- Hosler, B. A., T. Siddique, P. C. Sapp, W. Sailor, M. C. Huang, A. Hossain, J. R. Daube, et al. 2000. "Linkage of Familial Amyotrophic Lateral Sclerosis with Frontotemporal Dementia to Chromosome 9q21-Q22." *JAMA: The Journal of the American Medical Association* 284 (13): 1664–69.
- Hujoel, Margaux L. A., Steven Gazal, Farhad Hormozdiari, Bryce van de Geijn, and Alkes L. Price. 2019. "Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species." *American Journal of Human Genetics* 104 (4): 611–24.
- Hulst, Egberdina-Józefa van der, Thomas H. Bak, and Sharon Abrahams. 2015. "Impaired Affective and Cognitive Theory of Mind and Behavioural Change in Amyotrophic Lateral Sclerosis." *Journal of Neurology, Neurosurgery, and Psychiatry* 86 (11): 1208–15.
- Ingre, Caroline, Per M. Roos, Fredrik Piehl, Freya Kamel, and Fang Fang. 2015. "Risk Factors for Amyotrophic Lateral Sclerosis." *Clinical Epidemiology* 7 (February): 181–93.
- International HapMap 3 Consortium, David M. Altshuler, Richard A. Gibbs, Leena

- Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, et al. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58.
- International HapMap Consortium. 2005. "A Haplotype Map of the Human Genome." *Nature* 437 (7063): 1299–1320.
- International HapMap Consortium, Kelly A. Frazer, Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61.
- Jacquemont, Sébastien, Bradley P. Coe, Micha Hersch, Michael H. Duyzend, Niklas Krumm, Sven Bergmann, Jacques S. Beckmann, Jill A. Rosenfeld, and Evan E. Eichler. 2014. "A Higher Mutational Burden in Females Supports a 'Female Protective Model' in Neurodevelopmental Disorders." *The American Journal of Human Genetics* 94 (3): 415–25.
- Johnston, Clare A., Biba R. Stanton, Martin R. Turner, Rebecca Gray, Ashley Hay-Ming Blunt, David Butt, Mary-Ann Ampong, Christopher E. Shaw, P. Nigel Leigh, and Ammar Al-Chalabi. 2006. "Amyotrophic Lateral Sclerosis in an Urban Setting: A Population Based Study of Inner City London." *Journal of Neurology* 253 (12): 1642–43.
- Jong, Sonja de, Mark Huisman, Nadia Sutedja, Anneke van der Kooi, Marianne de Visser, Jurgen Schelhaas, Yvonne van der Schouw, Jan Veldink, and Leonard van den Berg. 2013. "Endogenous Female Reproductive Hormones and the Risk of Amyotrophic Lateral Sclerosis." *Journal of Neurology* 260 (2): 507–12.
- Karch, Celeste M., Natalie Wen, Chun C. Fan, Jennifer S. Yokoyama, Naomi Kouri, Owen A. Ross, Gunter Höglinger, et al. 2018. "Selective Genetic Overlap Between Amyotrophic Lateral Sclerosis and Diseases of the Frontotemporal Dementia Spectrum." *JAMA Neurology*, April. <https://doi.org/10.1001/jamaneurol.2018.0372>.
- Kenna, Kevin P., Perry T. C. van Doormaal, Annelot M. Dekker, Nicola Ticozzi, Brendan J. Kenna, Frank P. Diekstra, Wouter van Rheenen, et al. 2016. "NEK1 Variants Confer Susceptibility to Amyotrophic Lateral Sclerosis." *Nature Genetics* 48 (9): 1037–42.
- Kerminen, Sini, Aki S. Havulinna, Garrett Hellenthal, Alicia R. Martin, Antti-Pekka Sarin, Markus Perola, Aarno Palotie, et al. 2017. "Fine-Scale Genetic Structure in Finland." *G3* 7 (10): 3459–68.
- Kerminen, Sini, Alicia R. Martin, Jukka Koskela, Sanni E. Ruotsalainen, Aki S. Havulinna, Ida Surakka, Aarno Palotie, et al. 2019. "Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland." *American Journal of Human Genetics* 104 (6): 1169–81.
- Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations." *Nature Genetics* 50 (9): 1219–24.
- Khramtsova, Ekaterina A., Lea K. Davis, and Barbara E. Stranger. 2019. "The Role of Sex in the Genomics of Human Complex Traits." *Nature Reviews. Genetics* 20 (3): 173–90.
- Kichaev, Gleb, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika K. Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. 2019. "Leveraging Polygenic Functional Enrichment to Improve GWAS Power." *American Journal of Human Genetics* 104 (1): 65–75.
- Kittles, Rick A., Weidong Chen, Ramesh K. Panguluri, Chiledum Ahaghotu, Aaron Jackson, Clement A. Adebamowo, Robin Griffin, et al. 2002. "CYP3A4-V and Prostate Cancer in African Americans: Causal or Confounding Association Because of Population Stratification?" *Human Genetics* 110 (6): 553–60.
- Knowler, W. C., R. C. Williams, D. J. Pettitt, and A. G. Steinberg. 1988. "Gm3;5,13,14 and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic

- Admixture." *American Journal of Human Genetics* 43 (4): 520–26.
- Kosoy, Roman, Rami Nassir, Chao Tian, Phoebe A. White, Lesley M. Butler, Gabriel Silva, Rick Kittles, et al. 2009. "Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America." *Human Mutation* 30 (1): 69–78.
- Lawson, Daniel John, Neil Martin Davies, Simon Haworth, Bilal Ashraf, Laurence Howe, Andrew Crawford, Gibran Hemani, George Davey Smith, and Nicholas John Timpson. 2019. "Is Population Structure in the Genetic Biobank Era Irrelevant, a Challenge, or an Opportunity?" *Human Genetics*, April. <https://doi.org/10.1007/s00439-019-02014-8>.
- Lawson, Daniel John, and Daniel Falush. 2012. "Population Identification Using Genetic Data." *Annual Review of Genomics and Human Genetics* 13 (June): 337–61.
- Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. "Inference of Population Structure Using Dense Haplotype Data." Edited by Gregory P. Copenhaver. *PLoS Genetics* 8 (1): e1002453.
- Lazaridis, Iosif, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, et al. 2016. "Genomic Insights into the Origin of Farming in the Ancient Near East." *Nature* 536 (7617): 419–24.
- Lee, James J., Matt McGue, William G. Iacono, and Carson C. Chow. 2018. "The Accuracy of LD Score Regression as an Estimator of Confounding and Genetic Correlations in Genome-Wide Association Studies." *Genetic Epidemiology* 42 (8): 783–95.
- Lee, James J., Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghziyan, Meghan Zacher, Tuan Anh Nguyen-Viet, et al. 2018. "Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals." *Nature Genetics* 50 (8): 1112–21.
- Lee, Sang Hong, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher. 2011. "Estimating Missing Heritability for Disease from Genome-Wide Association Studies." *American Journal of Human Genetics* 88 (3): 294–305.
- Leeuw, Christiaan A. de, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. 2015. "MAGMA: Generalized Gene-Set Analysis of GWAS Data." *PLoS Computational Biology* 11 (4): e1004219.
- Leslie, Stephen, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, et al. 2015. "The Fine Scale Genetic Structure of the British Population." *Nature* 519 (7543): 309–14.
- Li, Na, and Matthew Stephens. 2003. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data." *Genetics* 165 (4): 2213–33.
- Liley, James, and Chris Wallace. 2015. "A Pleiotropy-Informed Bayesian False Discovery Rate Adapted to a Shared Control Design Finds New Disease Associations from GWAS Summary Statistics." *PLoS Genetics* 11 (2): e1004926.
- Liu, Xiaolu, Ji He, Fen-Biao Gao, Aaron D. Gitler, and Dongsheng Fan. 2018. "The Epidemiology and Genetics of Amyotrophic Lateral Sclerosis in China." *Brain Research* 1693 (Pt A): 121–26.
- Liu, Xuanyao, Yang I. Li, and Jonathan K. Pritchard. 2019. "Trans Effects on Gene Expression Can Drive Omnigenic Inheritance." *Cell* 177 (4): 1022–1034.e6.
- Longinetti, Elisa, Daniela Mariosa, Henrik Larsson, Weimin Ye, Caroline Ingre, Catarina Almqvist, Paul Lichtenstein, Fredrik Piehl, and Fang Fang. 2017. "Neurodegenerative and Psychiatric Diseases among Families with Amyotrophic Lateral Sclerosis." *Neurology* 89 (6): 578–85.
- Maaten, Laurens. 2009. "Learning a Parametric Embedding by Preserving Local Structure." In *Artificial Intelligence and Statistics*, 384–91.
- Maher, Brendan. 2008. "Personal Genomes: The Case of the Missing Heritability." *Nature* 456 (7218): 18–21.
- Majounie, Elisa, Alan E. Renton, Kin Mok, Elise G. P. Dopper, Adrian Waite, Sara



- Rollinson, Adriano Chiò, et al. 2012. "Frequency of the C9orf72 Hexanucleotide Repeat Expansion in Patients with Amyotrophic Lateral Sclerosis and Frontotemporal Dementia: A Cross-Sectional Study." *Lancet Neurology* 11 (4): 323–30.
- Manjaly, Zita R., Kirsten M. Scott, Kumar Abhinav, Lokesh Wijesekera, Jeban Ganesalingam, Laura H. Goldstein, Anna Janssen, et al. 2010. "The Sex Ratio in Amyotrophic Lateral Sclerosis: A Population Based Study." *Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* 11 (5): 439–42.
- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53.
- Marchini, Jonathan, Lon R. Cardon, Michael S. Phillips, and Peter Donnelly. 2004. "The Effects of Human Population Structure on Large Genetic Association Studies." *Nature Genetics* 36 (5): 512–17.
- Marigorta, Urko M., Juan Antonio Rodríguez, Greg Gibson, and Arcadi Navarro. 2018. "Replicability and Prediction: Lessons and Challenges from GWAS." *Trends in Genetics: TIG* 34 (7): 504–17.
- Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. 2017. "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations." *American Journal of Human Genetics* 100 (4): 635–49.
- Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4): 584–91.
- Martin, Alicia R., Konrad J. Karczewski, Sini Kerminen, Mitja I. Kurki, Antti-Pekka Sarin, Mykyta Artomov, Johan G. Eriksson, et al. 2018. "Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland." *American Journal of Human Genetics* 102 (5): 760–75.
- Martin, Joanna, Ekaterina A. Khramtsova, Slavina B. Goleva, Gabriëlla A. M. Blokland, Michela Traglia, Raymond K. Walters, Christopher Hübel, et al. 2020. "Examining Sex-Differentiated Genetic Effects across Neuropsychiatric and Behavioral Traits." *BioRxiv*. <https://doi.org/10.1101/2020.05.04.076042>.
- Martiniano, Rui, Anwen Caffell, Malin Holst, Kurt Hunter-Mann, Janet Montgomery, Gundula Müldner, Russell L. McLaughlin, et al. 2016. "Genomic Signals of Migration and Continuity in Britain before the Anglo-Saxons." *Nature Communications* 7 (January): 10326.
- Mathieson, Iain, and Gil McVean. 2012. "Differential Confounding of Rare and Common Variants in Spatially Structured Populations." *Nature Genetics* 44 (3): 243–46.
- McEvoy, Brian, Claire Brady, Laoise T. Moore, and Daniel G. Bradley. 2006. "The Scale and Nature of Viking Settlement in Ireland from Y-Chromosome Admixture Analysis." *European Journal of Human Genetics: EJHG* 14 (12): 1288–94.
- McEvoy, Brian, Katharine Simms, and Daniel G. Bradley. 2008. "Genetic Investigation of the Patrilineal Kinship Structure of Early Medieval Ireland." *American Journal of Physical Anthropology* 136 (4): 415–22.
- McLaughlin, Rowan, Emma Hannah, and Lisa Coyle-McClung. 2018. "Frequency Analyses of Historical and Archaeological Datasets Reveal the Same Pattern of Declining Sociocultural Activity in 9th to 10th Century CE Ireland." *Cliodynamics* 9 (1). <https://doi.org/10.21237/C7clio9136654>.
- McLaughlin, Russell L., Kevin P. Kenna, Alice Vajda, Mark Heverin, Susan Byrne, Colette G. Donaghy, Simon Cronin, Daniel G. Bradley, and Orla Hardiman. 2015. "Homozygosity Mapping in an Irish ALS Case-Control Cohort Describes Local Demographic Phenomena and Points towards Potential Recessive Risk Loci." *Genomics* 105 (4): 237–41.

- McLaughlin, Russell L., Dick Schijven, Wouter van Rheenen, Kristel R. van Eijk, Margaret O'Brien, René S. Kahn, Roel A. Ophoff, et al. 2017. "Genetic Correlation between Amyotrophic Lateral Sclerosis and Schizophrenia." *Nature Communications* 8 (March): 14774.
- Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. "Synthetic Maps of Human Gene Frequencies in Europeans." *Science* 201 (4358): 786–92.
- Montinaro, Francesco, George B. J. Busby, Vincenzo L. Pascali, Simon Myers, Garrett Hellenthal, and Cristian Capelli. 2015. "Unravelling the Hidden Ancestry of American Admixed Populations." *Nature Communications* 6 (March): 6596.
- Moore, Laoise T., Brian McEvoy, Eleanor Cape, Katharine Simms, and Daniel G. Bradley. 2006. "A Y-Chromosome Signature of Hegemony in Gaelic Ireland." *American Journal of Human Genetics* 78 (2): 334–38.
- Moorjani, Priya, Nick Patterson, Joel N. Hirschhorn, Alon Keinan, Li Hao, Gil Atzmon, Edward Burns, Harry Ostrer, Alkes L. Price, and David Reich. 2011. "The History of African Gene Flow into Southern Europeans, Levantines, and Jews." *PLoS Genetics* 7 (4): e1001373.
- Morgan, Sarah, Aleksey Shatunov, William Sproviero, Ashley R. Jones, Maryam Shoai, Deborah Hughes, Ahmad Al Khleifat, et al. 2017. "A Comprehensive Analysis of Rare Genetic Variation in Amyotrophic Lateral Sclerosis in the UK." *Brain: A Journal of Neurology* 140 (6): 1611–18.
- Nagai, Akiko, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, et al. 2017. "Overview of the BioBank Japan Project: Study Design and Profile." *Journal of Epidemiology / Japan Epidemiological Association* 27 (3S): S2–8.
- Nakazawa, Minato. 2018. "Fmsb: Functions for Medical Statistics Book with Some Demographic Data, 2014." *R Package*.
- Nalls, Michael A., Javier Simon-Sanchez, J. Raphael Gibbs, Coro Paisan-Ruiz, Jose Tomas Bras, Toshiko Tanaka, Mar Matarin, et al. 2009. "Measures of Autozygosity in Decline: Globalization, Urbanization, and Its Implications for Medical Genetics." *PLoS Genetics* 5 (3): e1000415.
- Naseri, Ardalan, Xiaoming Liu, Kecong Tang, Shaojie Zhang, and Degui Zhi. 2019. "RaPID: Ultra-Fast, Powerful, and Accurate Detection of Segments Identical by Descent (IBD) in Biobank-Scale Cohorts." *Genome Biology* 20 (1): 143.
- Need, Anna C., and David B. Goldstein. 2009. "Next Generation Disparities in Human Genomics: Concerns and Remedies." *Trends in Genetics: TIG* 25 (11): 489–94.
- Nicolas, Aude, Kevin P. Kenna, Alan E. Renton, Nicola Ticozzi, Faraz Faghri, Ruth Chia, Janice A. Dominov, et al. 2018. "Genome-Wide Analyses Identify KIF5A as a Novel ALS Gene." *Neuron* 97 (6): 1268-1283.e6.
- Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, et al. 2008. "Genes Mirror Geography within Europe." *Nature* 456 (7218): 98–101.
- Novembre, John, and Benjamin M. Peter. 2016. "Recent Advances in the Study of Fine-Scale Population Structure in Humans." *Current Opinion in Genetics & Development* 41 (December): 98–105.
- O'Connor, Luke J., and Alkes L. Price. 2018. "Distinguishing Genetic Correlation from Causation across 52 Diseases and Complex Traits." *Nature Genetics* 50 (12): 1728–34.
- O'Dushlaine, Colm T., Derek Morris, Valentina Moskvina, George Kirov, International Schizophrenia Consortium, Michael Gill, Aiden Corvin, James F. Wilson, and Gianpiero L. Cavalleri. 2010. "Population Structure and Genome-Wide Patterns of Variation in Ireland and Britain." *European Journal of Human Genetics: EJHG* 18 (11): 1248–54.
- Otowa, T., K. Hek, M. Lee, E. M. Byrne, S. S. Mirza, M. G. Nivard, T. Bigdeli, et al. 2016. "Meta-Analysis of Genome-Wide Association Studies of Anxiety Disorders." *Molecular Psychiatry* 21 (10): 1485.

- Palamara, Pier Francesco. 2014. "Population Genetics of Identity by Descent." *ArXiv [q-Bio.PE]*. arXiv. <http://arxiv.org/abs/1403.4987>.
- Palamara, Pier Francesco, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. 2012. "Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History." *American Journal of Human Genetics* 91 (5): 809–22.
- Palamara, Pier Francesco, and Itsik Pe'er. 2013. "Inference of Historical Migration Rates via Haplotype Sharing." *Bioinformatics* 29 (13): i180–8.
- Pang, Shirley Yin-Yu, Jacob Shujui Hsu, Kay-Cheong Teo, Yan Li, Michelle H. W. Kung, Kathryn S. E. Cheah, Danny Chan, et al. 2017. "Burden of Rare Variants in ALS Genes Influences Survival in Familial and Sporadic ALS." *Neurobiology of Aging* 58 (October): 238.e9–238.e15.
- Patterson, Nick, Alkes L. Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2 (12): e190.
- Petkova, Desislava, John Novembre, and Matthew Stephens. 2016. "Visualizing Spatial Population Structure with Estimated Effective Migration Surfaces." *Nature Genetics* 48 (1): 94–100.
- Phukan, Julie, Marwa Elamin, Peter Bede, Norah Jordan, Laura Gallagher, Susan Byrne, Catherine Lynch, Niall Pender, and Orla Hardiman. 2012. "The Syndrome of Cognitive Impairment in Amyotrophic Lateral Sclerosis: A Population-Based Study." *Journal of Neurology, Neurosurgery, and Psychiatry* 83 (1): 102–8.
- Pierre, Aude Saint, Joanna Giemza, Isabel Alves, Matilde Karakachoff, Marinna Gaudin, Philippe Amouyel, Jean-François Dartigues, et al. 2020. "The Genetic History of France." *European Journal of Human Genetics: EJHG*, February, 1–13.
- Pirastu, N., M. Cordioli, P. Nandakumar, and G. Mignogna. 2020. "Genetic Analyses Identify Widespread Sex-Differential Participation Bias." *Biorxiv*. <https://www.biorxiv.org/content/10.1101/2020.03.22.001453v1.abstract>.
- Platzer, Alexander. 2013. "Visualization of SNPs with T-SNE." *PloS One* 8 (2): e56883.
- Popejoy, Alice B., and Stephanie M. Fullerton. 2016. "Genomics Is Failing on Diversity." October 12, 2016. <https://doi.org/10.1038/538161a>.
- Price, Alkes L., Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L. Pascali, Francesca Scarnicci, Andres Ruiz-Linares, et al. 2008. "Discerning the Ancestry of European Americans in Genetic Association Studies." *PLoS Genetics* 4 (1): e236.
- Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9.
- Price, Alkes L., Michael E. Weale, Nick Patterson, Simon R. Myers, Anna C. Need, Kevin V. Shianna, Dongliang Ge, et al. 2008. "Long-Range LD Can Confound Genome Scans in Admixed Populations." *American Journal of Human Genetics*.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59.
- Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly. 2000. "Association Mapping in Structured Populations." *American Journal of Human Genetics* 67 (1): 170–81.
- Randall, Joshua C., Thomas W. Winkler, Zoltán Kutalik, Sonja I. Berndt, Anne U. Jackson, Keri L. Monda, Tuomas O. Kilpeläinen, et al. 2013. "Sex-Stratified Genome-Wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits." *PLoS Genetics* 9 (6): e1003500.
- Raveane, A., S. Aneli, F. Montinaro, G. Athanasiadis, S. Barlera, G. Birolo, G. Boncoraglio, et al. 2019. "Population Structure of Modern-Day Italians Reveals Patterns of Ancient and Archaic Ancestries in Southern Europe." *Science Advances* 5 (9): eaaw3492.
- Relethford, J. H. 1983. "Genetic Structure and Population History of Ireland: A Comparison of Blood Group and Anthropometric Analyses." *Annals of Human Biology* 10 (4): 321–33.

- Rheenen, Wouter van, Sara L. Pulit, Annelot M. Dekker, Ahmad Al Khleifat, William J. Brands, Alfredo Iacoangeli, Kevin P. Kenna, et al. 2018. "Project MinE: Study Design and Pilot Analyses of a Large-Scale Whole-Genome Sequencing Study in Amyotrophic Lateral Sclerosis." *European Journal of Human Genetics: EJHG* 26 (10): 1537–46.
- Rheenen, Wouter van, Aleksey Shatunov, Annelot M. Dekker, Russell L. McLaughlin, Frank P. Diekstra, Sara L. Pulit, Rick A. A. van der Spek, et al. 2016. "Genome-Wide Association Analyses Identify New Risk Variants and the Genetic Architecture of Amyotrophic Lateral Sclerosis." *Nature Genetics* 48 (9): 1043–48.
- Ripatti, Samuli, Emmi Tikkanen, Marju Orho-Melander, Aki S. Havulinna, Kaisa Silander, Amitabh Sharma, Candace Guiducci, et al. 2010. "A Multilocus Genetic Risk Score for Coronary Heart Disease: Case-Control and Prospective Cohort Analyses." *The Lancet* 376 (9750): 1393–1400.
- Rippon, Gregory A., Nikolaos Scarmeas, Paul H. Gordon, Peregrine L. Murphy, Steven M. Albert, Hiroshi Mitsumoto, Karen Marder, Lewis P. Rowland, and Yaakov Stern. 2006. "An Observational Study of Cognitive Impairment in Amyotrophic Lateral Sclerosis." *Archives of Neurology* 63 (3): 345–52.
- Robinson, Matthew R., Geoffrey English, Gerhard Moser, Luke R. Lloyd-Jones, Marcus A. Triplett, Zhihong Zhu, Ilja M. Nolte, et al. 2017. "Genotype-Covariate Interaction Effects and the Heritability of Adult Body Mass Index." *Nature Genetics* 49 (8): 1174–81.
- Rooney, James P. K., Isabella Fogh, Henk-Jan Westeneng, Alice Vajda, Russell McLaughlin, Mark Heverin, Ashley Jones, et al. 2017. "C9orf72 Expansion Differentially Affects Males with Spinal Onset Amyotrophic Lateral Sclerosis." *Journal of Neurology, Neurosurgery, and Psychiatry* 88 (4): 281.
- Rooney, James P. K., Anne E. Visser, Fabrizio D'Ovidio, Roel Vermeulen, Ettore Beghi, Adriano Chio, Jan H. Veldink, et al. 2017. "A Case-Control Study of Hormonal Exposures as Etiologic Factors for ALS in Women: Euro-MOTOR." *Neurology* 89 (12): 1283–90.
- Roosen, Joris, and Daniel R. Curtis. 2019. "The 'Light Touch' of the Black Death in the Southern Netherlands: An Urban Trick?" *The Economic History Review* 72 (1): 32–56.
- Rosen, D. R., T. Siddique, D. Patterson, D. A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J. P. O'Regan, and H. X. Deng. 1993. "Mutations in Cu/Zn Superoxide Dismutase Gene Are Associated with Familial Amyotrophic Lateral Sclerosis." *Nature* 362 (6415): 59–62.
- Ruderfer, Douglas M., Stephan Ripke, Andrew McQuillin, James Boocock, Eli A. Stahl, Jennifer M. Whitehead Pavlides, Niamh Mullins, et al. 2018. "Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes." *Cell* 173 (7): 1705-1715.e16.
- Ryan, Marie, Mark Heverin, Mark A. Doherty, Nicola Davis, Emma M. Corr, Alice Vajda, Niall Pender, Russell McLaughlin, and Orla Hardiman. 2018. "Determining the Incidence of Familiality in ALS: A Study of Temporal Trends in Ireland from 1994 to 2016." *Neurology. Genetics* 4 (3): e239.
- Ryan, Marie, Mark Heverin, Russell L. McLaughlin, and Orla Hardiman. 2019. "Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis." *JAMA Neurology*, July. <https://doi.org/10.1001/jamaneurol.2019.2044>.
- Sakaue, Saori, Jun Hirata, Masahiro Kanai, Ken Suzuki, Masato Akiyama, Chun Lai Too, Thurayya Arayssi, et al. 2020. "Dimensionality Reduction Reveals Fine-Scale Structure in the Japanese Population with Consequences for Polygenic Risk Prediction." *Nature Communications* 11 (1): 1–11.
- Sawcer, Stephen, Garrett Hellenthal, Matti Pirinen, Chris C. A. Spencer, Nikolaos A. Patsopoulos, Loukas Moutsianas, Alexander Dilthey, et al. 2011. "Genetic Risk and a Primary Role for Cell-Mediated Immune Mechanisms in Multiple Sclerosis." *Nature* 476 (7359): 214–19.

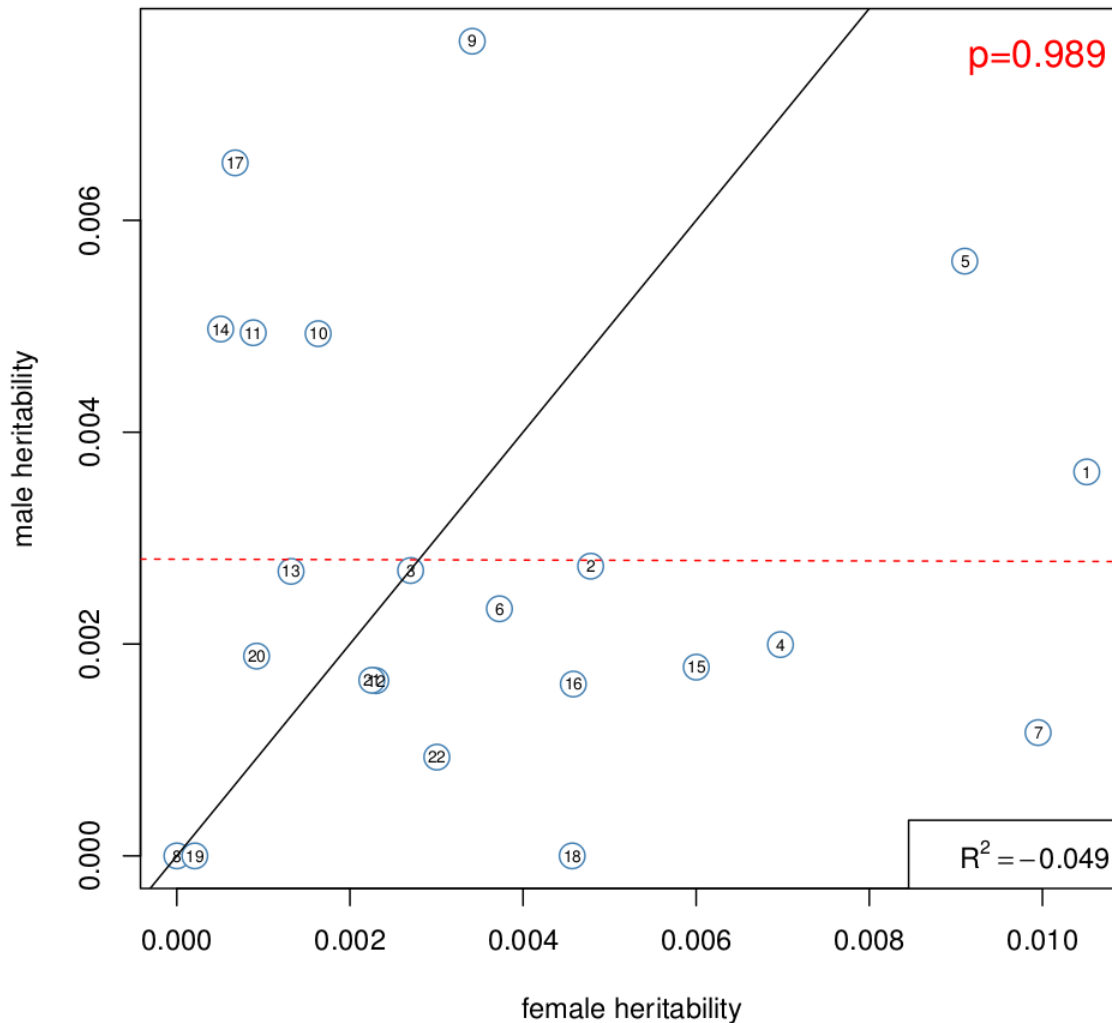
- Schiffels, Stephan, Wolfgang Haak, Pirita Paajanen, Bastien Llamas, Elizabeth Popescu, Louise Loe, Rachel Clarke, et al. 2016. "Iron Age and Anglo-Saxon Genomes from East England Reveal British Migration History." *Nature Communications* 7 (January): 10408.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium, Stephan Ripke, James T. R. Walters, and Michael C. O'Donovan. 2020. "Mapping Genomic Loci Prioritises Genes and Implicates Synaptic Biology in Schizophrenia." *Genetic and Genomic Medicine*. medRxiv. <https://doi.org/10.1101/2020.09.12.20192922>.
- Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289.
- Shi, Huwenbo, Gleb Kichaev, and Bogdan Pasaniuc. 2016. "Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data." *American Journal of Human Genetics* 99 (1): 139–53.
- Shi, Huwenbo, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. 2017. "Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits." *American Journal of Human Genetics* 101 (5): 737–51.
- Smith, Malcolm T., and Donald M. Macrauld. 2009. "Nineteenth-Century Population Structure of Ireland and of the Irish in England and Wales: An Analysis by Isonymy." *American Journal of Human Biology: The Official Journal of the Human Biology Council* 21 (3): 283–89.
- Sohail, Mashaal, Robert M. Maier, Andrea Ganna, Alex Bloemendal, Alicia R. Martin, Michael C. Turchin, Charleston Wk Chiang, et al. 2019. "Polygenic Adaptation on Height Is Overestimated Due to Uncorrected Stratification in Genome-Wide Association Studies." *ELife* 8 (March). <https://doi.org/10.7554/eLife.39702>.
- Speed, Doug, Na Cai, UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. 2017. "Reevaluation of SNP Heritability in Complex Human Traits." *Nature Genetics* 49 (7): 986–92.
- Spek, Rick A. A. van der, Wouter van Rheenen, Sara L. Pulit, Kevin P. Kenna, Leonard H. van den Berg, Jan H. Veldink, and Project MinE ALS Sequencing Consortium¶. 2019. "The Project MinE Databrowser: Bringing Large-Scale Whole-Genome Sequencing in ALS to Researchers and the Public." *Amyotrophic Lateral Sclerosis & Frontotemporal Degeneration* 20 (5–6): 432–40.
- Sreedharan, Jemeen, Ian P. Blair, Vineeta B. Tripathi, Xun Hu, Caroline Vance, Boris Rogelj, Steven Ackerley, et al. 2008. "TDP-43 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis." *Science* 319 (5870): 1668–72.
- Takada, Leonel T. 2015. "The Genetics of Monogenic Frontotemporal Dementia." *Dementia & Neuropsychologia* 9 (3): 219–29.
- Takeuchi, Fumihiko, Tomohiro Katsuya, Ryosuke Kimura, Toru Nabika, Minoru Isomura, Takayoshi Ohkubo, Yasuharu Tabara, et al. 2017. "The Fine-Scale Genetic Structure and Evolution of the Japanese Population." *PloS One* 12 (11): e0185487.
- Timpson, Nicholas J., Celia M. T. Greenwood, Nicole Soranzo, Daniel J. Lawson, and J. Brent Richards. 2018. "Genetic Architecture: The Shape of the Genetic Contribution to Human Traits and Disease." *Nature Reviews. Genetics* 19 (2): 110–24.
- Trojsi, Francesca, Giulia D'Alvano, Simona Bonavita, and Gioacchino Tedeschi. 2020. "Genetics and Sex in the Pathogenesis of Amyotrophic Lateral Sclerosis (ALS): Is There a Link?" *International Journal of Molecular Sciences* 21 (10). <https://doi.org/10.3390/ijms21103647>.
- Tropf, Felix C., S. Hong Lee, Renske M. Verweij, Gert Stulp, Peter J. van der Most, Ronald de Vlaming, Andrew Bakshi, et al. 2017. "Hidden Heritability Due to Heterogeneity across Seven Populations." *Nature Human Behaviour* 1 (10): 757–65.
- Turley, Patrick, Raymond K. Walters, Omeed Maghziyan, Aysu Okbay, James J. Lee, Mark

- Alan Fontana, Tuan Anh Nguyen-Viet, et al. 2018. "Multi-Trait Analysis of Genome-Wide Association Summary Statistics Using MTAG." *Nature Genetics* 50 (2): 229–37.
- Turner, Martin R., Raph Goldacre, Kevin Talbot, and Michael J. Goldacre. 2016. "Psychiatric Disorders Prior to Amyotrophic Lateral Sclerosis." *Annals of Neurology* 80 (6): 935–38.
- Vilhjálmsón, Bjarni J., Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *American Journal of Human Genetics* 97 (4): 576–92.
- Vivekananda, Umesh, Zita-Rose Manjalay, Jeban Ganesalingam, Jacqueline Simms, Christopher E. Shaw, P. Nigel Leigh, Martin R. Turner, and Ammar Al-Chalabi. 2011. "Low Index-to-Ring Finger Length Ratio in Sporadic ALS Supports Prenatally Defined Motor Neuronal Vulnerability." *Journal of Neurology, Neurosurgery, and Psychiatry* 82 (6): 635–37.
- Wainschtein, Pierrick, Deepti P. Jain, Loic Yengo, Zhili Zheng, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, L. Adrienne Cupples, et al. 2019. "Recovery of Trait Heritability from Whole Genome Sequence Data." *BioRxiv*. <https://doi.org/10.1101/588020>.
- Wangkumhang, Pongsakorn. 2020. "Fast and Efficient Statistical Methods for Detecting Genetic Admixture Events and Its Applications in Large-Scale Data Cohorts." Edited by G. Hellenthal. Doctoral, UCL (University College London). <https://discovery.ucl.ac.uk/id/eprint/10094120/>.
- Watanabe, Kyoko, Sven Stringer, Oleksandr Frei, Maša Umičević Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. 2019. "A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits." *Nature Genetics* 51 (9): 1339–48.
- Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. "Functional Mapping and Annotation of Genetic Associations with FUMA." *Nature Communications* 8 (1): 1826.
- Winkler, Thomas W., Anne E. Justice, Mariaelisa Graff, Llilda Barata, Mary F. Feitosa, Su Chu, Jacek Czajkowski, et al. 2015. "The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study." *PLoS Genetics* 11 (10): e1005378.
- Wojcik, Genevieve L., Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, et al. 2019. "Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits." *Nature* 570 (7762): 514–18.
- Wray, Naomi R., Michael E. Goddard, and Peter M. Visscher. 2007. "Prediction of Individual Genetic Risk to Disease from Genome-Wide Association Studies." *Genome Research* 17 (10): 1520–28.
- Wray, Naomi R., and Robert Maier. 2014. "Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability." *Current Epidemiology Reports* 1 (4): 220–27.
- Wray, Naomi R., Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M. Byrne, Abdel Abdellaoui, Mark J. Adams, et al. 2018. "Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression." *Nature Genetics* 50 (5): 668–81.
- Xiao, Xiao, Fanfan Zheng, Hong Chang, Yina Ma, Yong-Gang Yao, Xiong-Jian Luo, and Ming Li. 2018. "The Gene Encoding Protocadherin 9 (PCDH9), a Novel Risk Factor for Major Depressive Disorder." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 43 (5): 1128–37.
- Yang, Jian, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A. E. Vinkhuyzen, Sang

- Hong Lee, Matthew R. Robinson, et al. 2015. "Genetic Variance Estimation with Imputed Variants Finds Negligible Missing Heritability for Human Height and Body Mass Index." *Nature Genetics* 47 (10): 1114–20.
- Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–69.
- Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A Tool for Genome-Wide Complex Trait Analysis." *American Journal of Human Genetics* 88 (1): 76–82.
- Yang, Jian, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. 2014. "Advantages and Pitfalls in the Application of Mixed-Model Association Methods." *Nature Genetics* 46 (2): 100–106.
- Yang, Jian, Jian Zeng, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2017. "Concepts, Estimation and Interpretation of SNP-Based Heritability." *Nature Genetics* 49 (9): 1304–10.
- Zaidi, Arslan A., and Iain A. Mathieson. 2020. "Demographic History Impacts Stratification in Polygenic Scores." *BioRxiv*. <https://doi.org/10.1101/2020.07.20.212530>.
- Zeng, Jian, Ronald de Vlaming, Yang Wu, Matthew R. Robinson, Luke R. Lloyd-Jones, Loic Yengo, Chloe X. Yap, et al. 2018. "Signatures of Negative Selection in the Genetic Architecture of Human Complex Traits." *Nature Genetics*, April. <https://doi.org/10.1038/s41588-018-0101-4>.
- Zheng, Jie, A. Mesut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, 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* 33 (2): 272–79.
- Zhu, Xianjun, Richard T. Libby, Wilhelmine N. de Vries, Richard S. Smith, Dana L. Wright, Roderick T. Bronson, Kevin L. Seburn, and Simon W. M. John. 2012. "Mutations in a P-Type ATPase Gene Cause Axonal Degeneration." *PLoS Genetics* 8 (8): e1002853.
- Zhu, Zhihong, Andrew Bakshi, Anna A. E. Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, et al. 2015. "Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits." *American Journal of Human Genetics* 96 (3): 377–85.

## Appendix - Supporting material

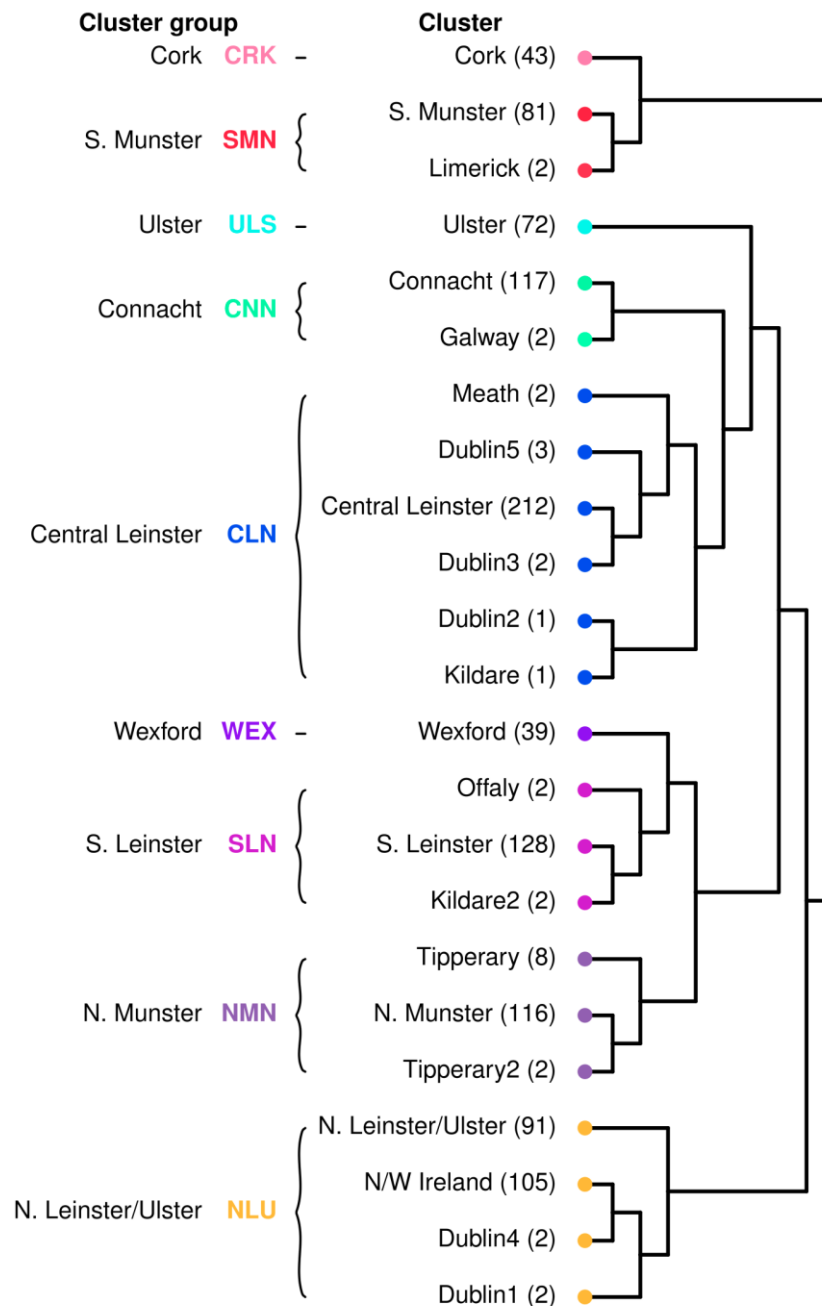
### Appendix material for Chapter 2



**Appendix Figure 2.1: Comparing male and female per chromosome heritability estimates from GREML shows strong differences in distribution of genetic effects.** GREML estimates of heritability in males and females show a non-significant linear relationship (red dashed line:  $p=0.98$ ;  $r^2=-0.049$ ). Several chromosomes show strong deviation from the 1:1 relationship line (black line) for heritability in males and females such as chromosome 9 which harbours the male specific hit in *PIP5K1B* and chromosome 17 which harbours male specific hits in *UNK*, *FBF1* and *SARM1* both of which show strong inflation in heritability in males compared to females.



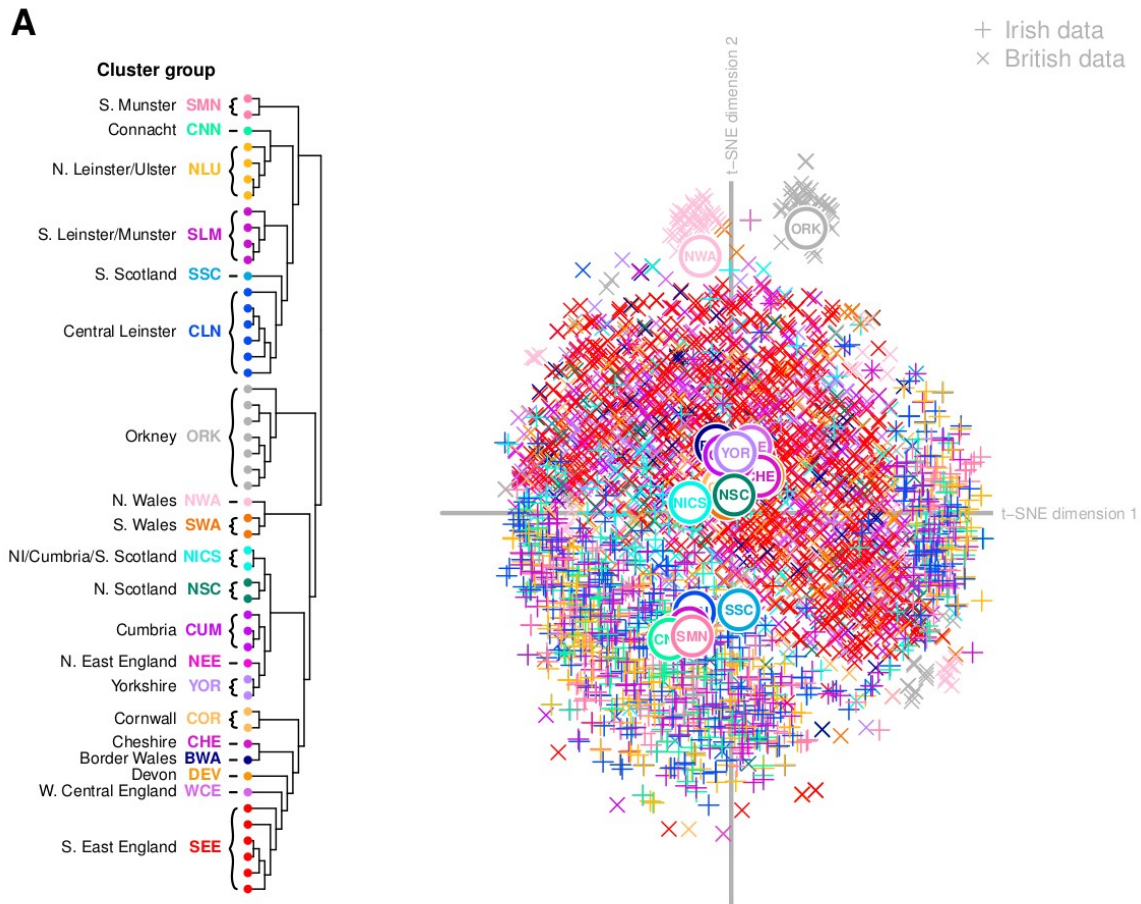
## Appendix material for Chapter 3



### Appendix Figure 3.1: Irish fineSTRUCTURE tree cluster details.

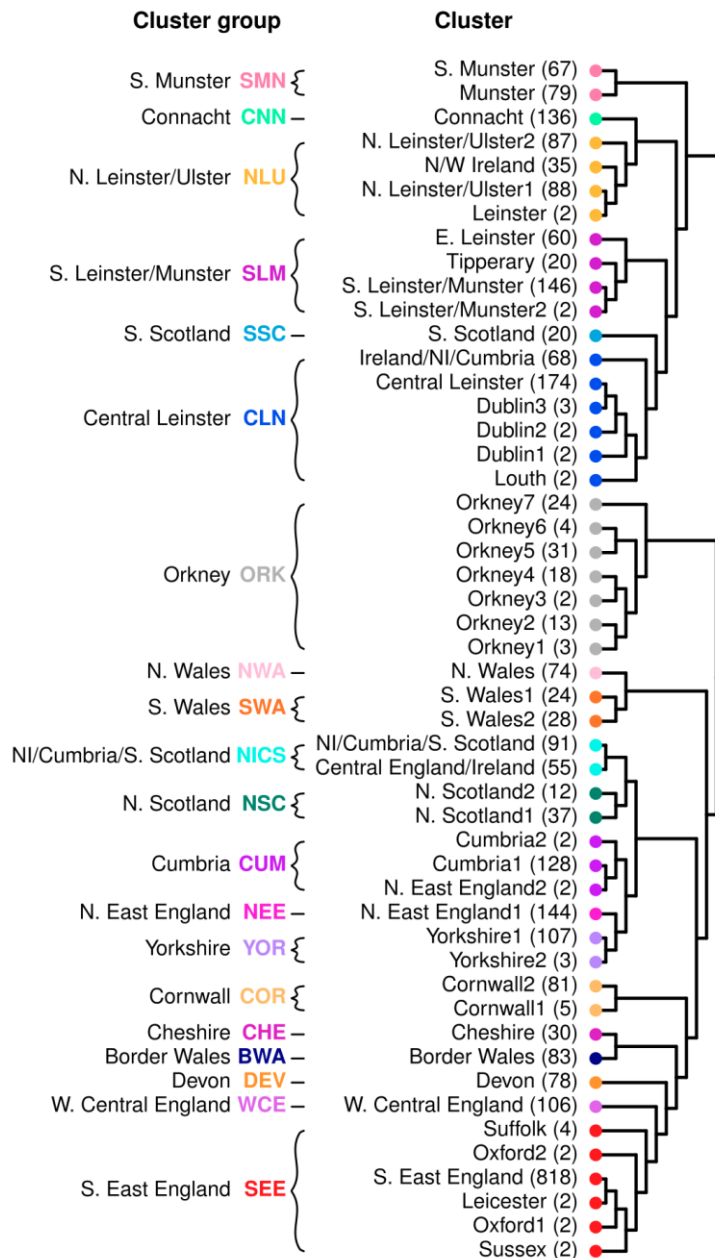
The fineSTRUCTURE tree presented in Figure 3.1 for Irish clusters with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups are identical to those defined in Figure 3.1.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))



**Appendix Figure 3.2: t-SNE projection of British and Irish SNP data.**

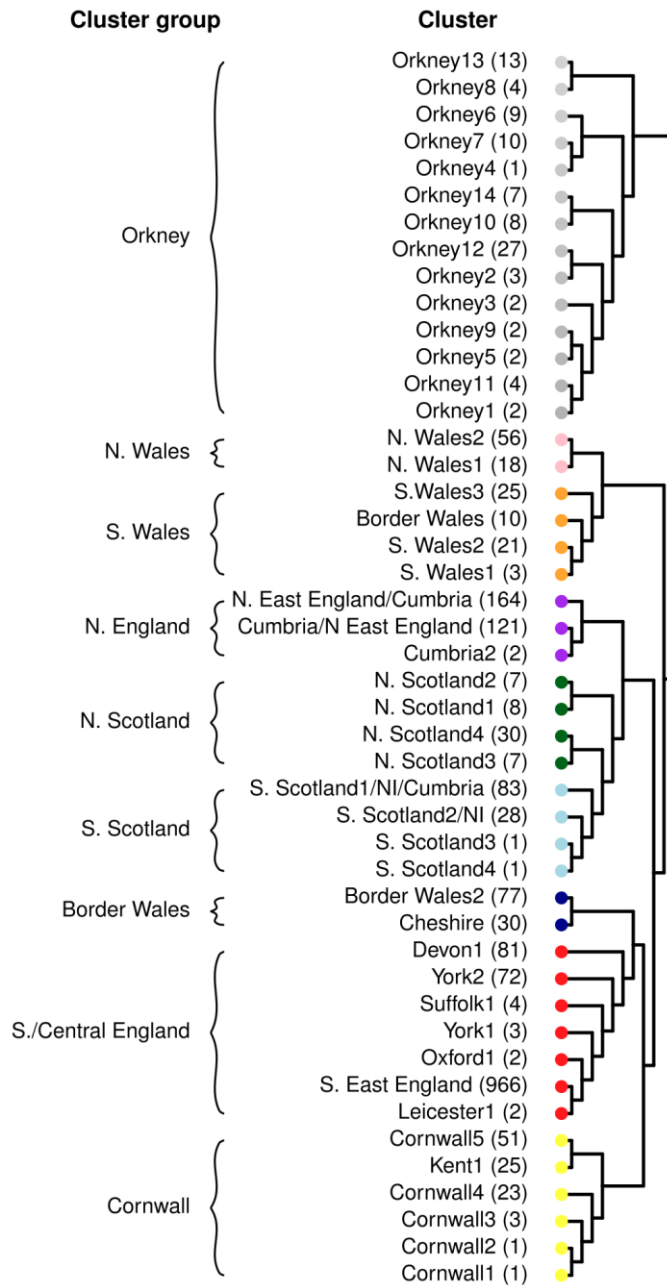
t-SNE projection of the hamming distance matrix for samples from the combined PoBI/Irish dataset demonstrates that less global and local structure is contained within independent SNPs than in the ChromoPainter coancestry matrix (Compare to Figure 3.6). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))



**Appendix Figure 3.3: PoBI/Irish fineSTRUCTURE tree cluster details.**

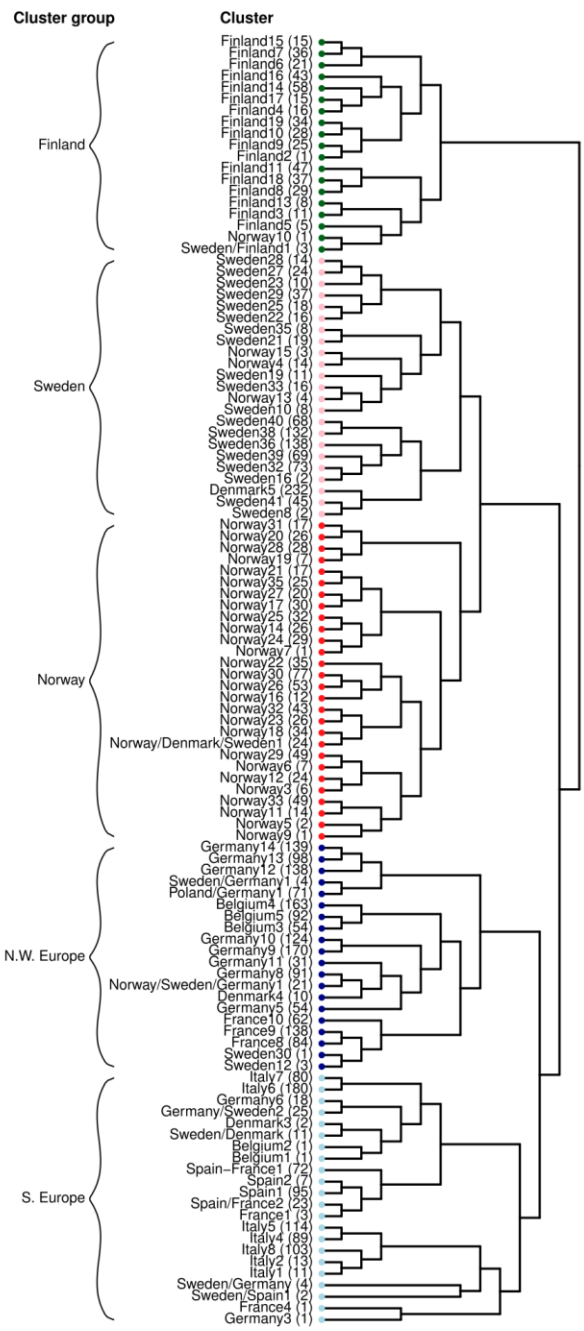
The fineSTRUCTURE tree presented in figure 3.2, 3.3 and 3.6 for British and Irish clusters with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups are identical to those defined in Figure 3.2.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

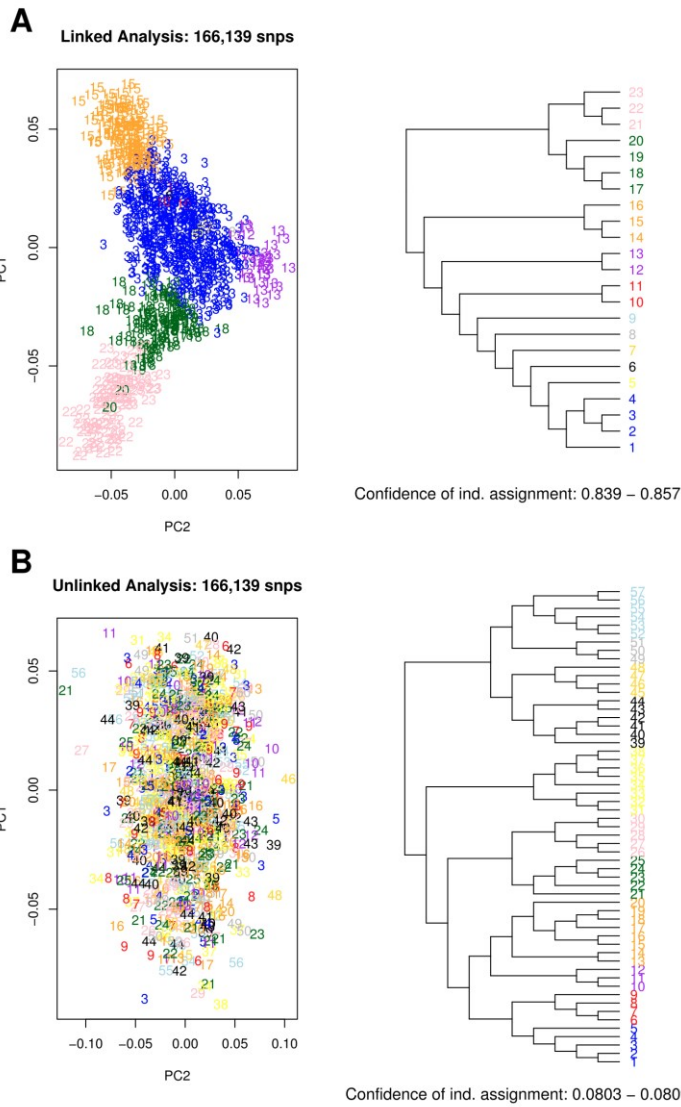


**Appendix Figure 3.4: PoBI maximum concordance fineSTRUCTURE tree cluster details.**

The fineSTRUCTURE maximum concordance tree for British clusters used in GLOBETROTTER analysis with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups describe clusters which are neighbouring in the tree and geographically adjacent. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))



**Appendix Figure 3.5: European maximum concordance fineSTRUCTURE tree cluster details.** The fineSTRUCTURE maximum concordance tree for European clusters used in GLOBETROTTER analysis with detailed breakdown of individual clusters. Additional individuals from WTCCC exclusion list have been removed post fineSTRUCTURE clustering but prior to GLOBETROTTER analysis and the tree updated to reflect this. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups describe clusters which are neighbouring in the tree and geographically adjacent. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))



**Appendix Figure 3.6: Comparison of Linked vs Unlinked fineSTRUCTURE in Ireland at 166,139 SNPs.**

Displays ChromoPainter PC1 and PC2 alongside a fineSTRUCTURE Maximum Concordance clustering dendrogram for fineSTRUCTURE A.) Linked and B.) Unlinked analysis for 991 Irish individuals at the 166,139 SNP positions used for our European GLOBETROTTER run. Trees and PCA are coloured at a  $k = 11$  split for ease of visualisation. Considerably more structure is apparent in the PCA of the Linked analysis indicating that linkage information defines meaningful haplotypes even at this number of SNPs. We report “Confidence of ind. assignment” for each method. This metric is the confidence of individual assignment to their final cluster based on their assignment across. This was on average 84.8% (95% CI: 83.9–85.7%) for the Linked analysis, while in the Unlinked analysis this was only 8.06% (95% CI: 8.03–8.09%), suggesting that the final clustering assignment in the unlinked mode is extremely uncertain and variable. This demonstrates that use of linkage information is informative even at the lowest SNP density in this Chapter 3. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2018))

**Appendix Table 3.1: Europe GLOBETROTTER table.**

Cluster	Fit.quality. 2events	Conclusion	Minor Source	Major Source	Date Estimate	Lower	Upper	Prop	maxR2	P No Admixture
All Ireland	0.997	One-date	Southern Europe (FRA(8))	Northern Europe (NOR-SWE-GER(1))	949.473	823.099	1131.356	0.36	0.711	0.000
NW Ulster	0.998	One-date	Southern Europe (FRA(9))	Northern Europe (NOR-SWE-GER(1))	1399.838	1141.77	1693.182	0.42	0.172	0.000
N Leinster/Ulster	0.986	One-date-multiway	Southern Europe (FRA(8))	Northern Europe (NOR-SWE-GER(1))	770.882	367.116	1324.096	0.41	0.166	0.000
Connaught	0.997	One-date	Southern Europe (FRA(8))	Northern Europe (NOR-SWE-GER(1))	950.909	567.449	1382.689	0.50	0.263	0.000
Central Leinster	0.996	One-date	Southern Europe (FRA(8))	Northern Europe (NOR-SWE-GER(1))	1029.596	686.242	1353.035	0.37	0.261	0.000
S Leinster/Munster	0.998	One-date-multiway	Southern Europe (FRA(8))	Northern Europe (NOR-SWE-GER(1))	1029.596	815.532	1235.877	0.45	0.584	0.000
Wexford	0.996	One-date	Northern Europe (SWE(41))	Southern Europe (FRA(10))	1051.455	198.910	1926.223	0.24	0.097	0.000
N Munster	0.992	One-date	Northern Europe (NOR-SWE-GER(1))	Southern Europe (FRA(9))	1216.368	838.765	1836.107	0.47	0.163	0.000
SW Munster	0.990	One-date	Northern Europe (NOR-SWE-GER(1))	Southern Europe (FRA(9))	1083.725	397.966	1821.872	0.34	0.084	0.000
Cork	0.978	No-Admix	Northern Europe (NOR-SWE-GER(1))	Southern Europe (FRA(9))	1125.783	347.566	2378.011	0.31	0.088	0.020

Table describing the model fit of GLOBETROTTER for admixture events into Irish clusters from Europe (Figure 3.7 and 3.8).

(Table reprinted from Byrne et al. (R. P. Byrne et al. 2018))

**Appendix Table 3.2: British GLOBETROTTER table.**

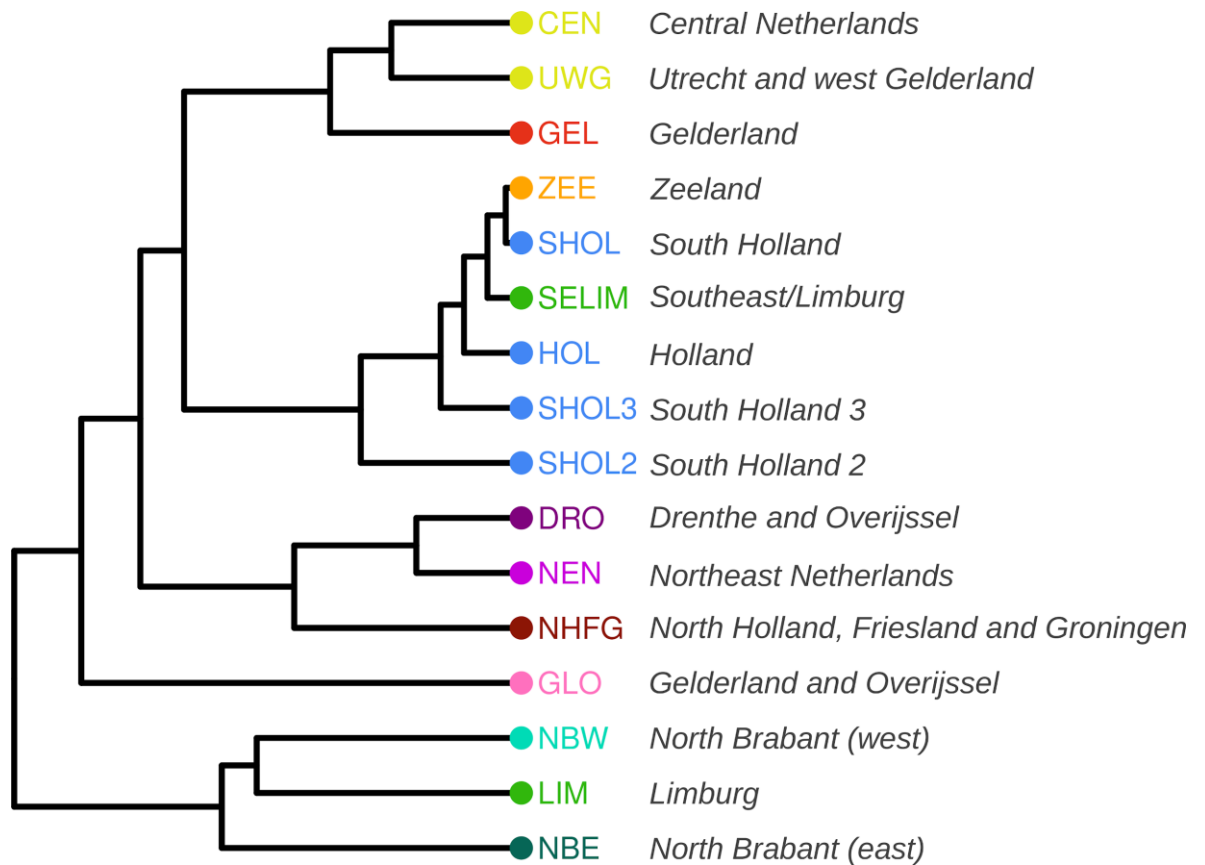
Cluster	fit.quality. 2events	Conclusion	Minor Source	Major Source	Date Estimate	Lower	Upper	Prop	maxR2	P No Admixture
All Ireland	0.850	Uncertain	England (SE_ENG)	Scotland (S_SCOT)	1325.243	1191.224	1458.975	0.380	0.615	0.000
NW Ulster	0.973	Uncertain	Scotland (S_SCOT)	Scotland (S_SCOT)	1538.944	1341.243	1780.356	0.320	0.365	0.000
N Leinster/Ulster	0.860	No-Admix	England (SE_ENG)	Scotland (S_SCOT)	1882.963	1861.926	2235.159	0.370	0.090	0.446
Connaught	0.986	Uncertain	Scotland (S_SCOT)	England (SE_ENG)	1146.397	486.185	1547.760	0.420	0.149	0.000
Central Leinster	0.976	Uncertain	Scotland (S_SCOT)	Scotland (S_SCOT)	1328.593	753.186	1616.627	0.300	0.346	0.010
S Leinster/Munster	0.964	Uncertain	England (SE_ENG)	Scotland (S_SCOT)	1430.666	1105.012	1605.811	0.470	0.371	0.000
Wexford	0.952	No-Admix	England (CHE)	Scotland (S_SCOT)	1713.425	1522.850	2242.600	0.440	0.085	0.743
N Munster	0.893	No-Admix	England (SE_ENG)	Scotland (S_SCOT)	1374.476	905.532	1686.670	0.460	0.088	0.040
SW Munster	0.864	No-Admix	England (SE_ENG)	Scotland (S_SCOT)	1701.700	1499.400	2432.485	0.330	0.190	0.376
Cork	0.928	No-Admix	England (SE_ENG)	Scotland (N_SCOT)	1904.000	1904.000	2826.637	0.430	0.049	0.644

Table describing the model fit of GLOBETROTTER for admixture events into Irish clusters from Britain (Figure 3.7 and 3.8).

(Table reprinted from Byrne et al. (R. P. Byrne et al. 2018))

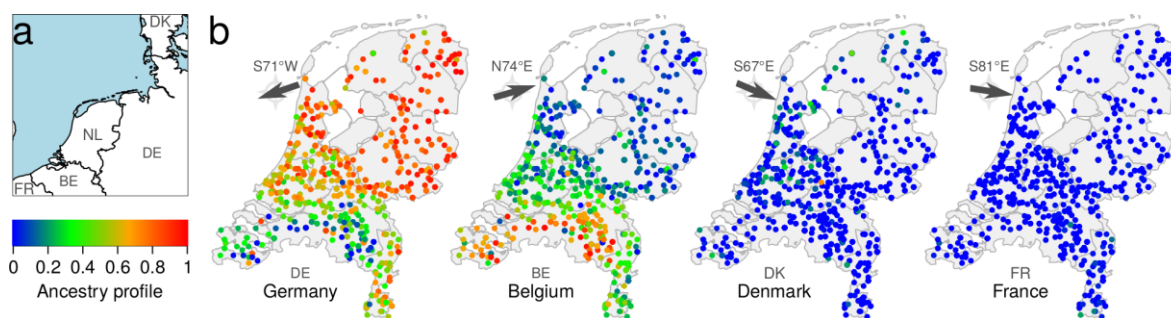


## Appendix material for Chapter 4



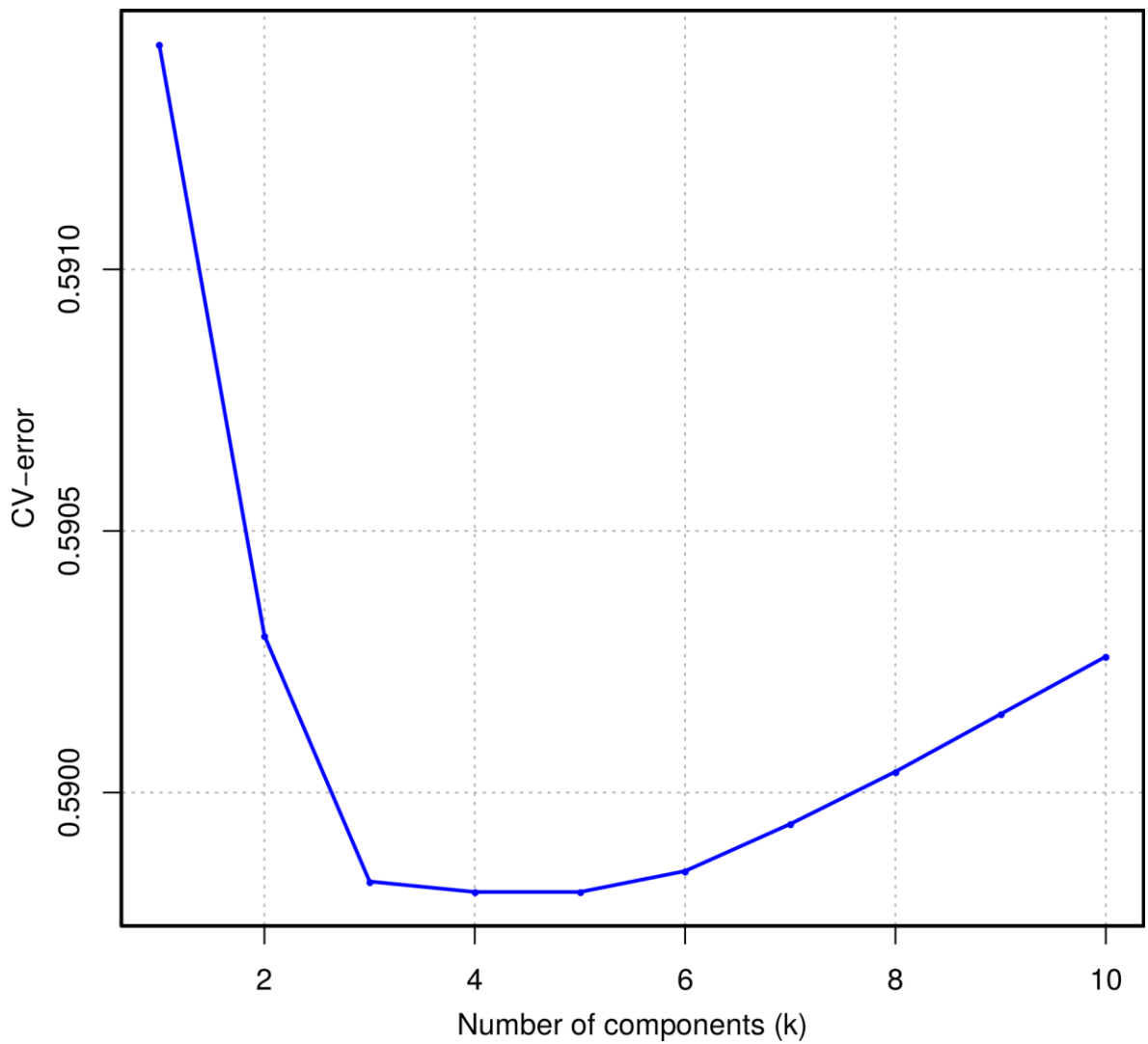
**Appendix Figure 4.1: Total variation distance (TVD) tree for  $k=16$  split in the Netherlands.**

TVD tree analysis provides an alternative view of the relationship between fineSTRUCTURE clusters based on their pairwise TVD scores (Figure 4.1a) which should be robust to differences in cluster size. Notably this tree prioritises the split between NBE/NBW/LIM with the rest of the Netherlands, which may reflect the geographic boundary seen in Figure 4.6. This tree is based on mean sharing between clusters and may thus miss subtle nuanced relationships where within cluster variation in sharing is non-zero. Clusters are coloured and labelled according to scheme in Figure 4.1. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



**Appendix Figure 4.2: SOURCEFIND ancestry gradients.**

(a) The Netherlands and its geographical relationship to neighbouring lands. (b) German, Belgian, Danish and French haplotypic ancestry profiles estimated using the alternative method SOURCEFIND for 1,422 Dutch individuals. Arrows indicate the predominant directions along which the ancestry gradients are arranged across the Netherlands. Major ancestry sources from the NNLS method (Figure 4.3) are strongly correlated with these estimates ( $r^2_{DE}=0.92$ ;  $r^2_{BE}=0.97$ ;  $r^2_{DK}=0.71$ ). Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>) and Natural Earth. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

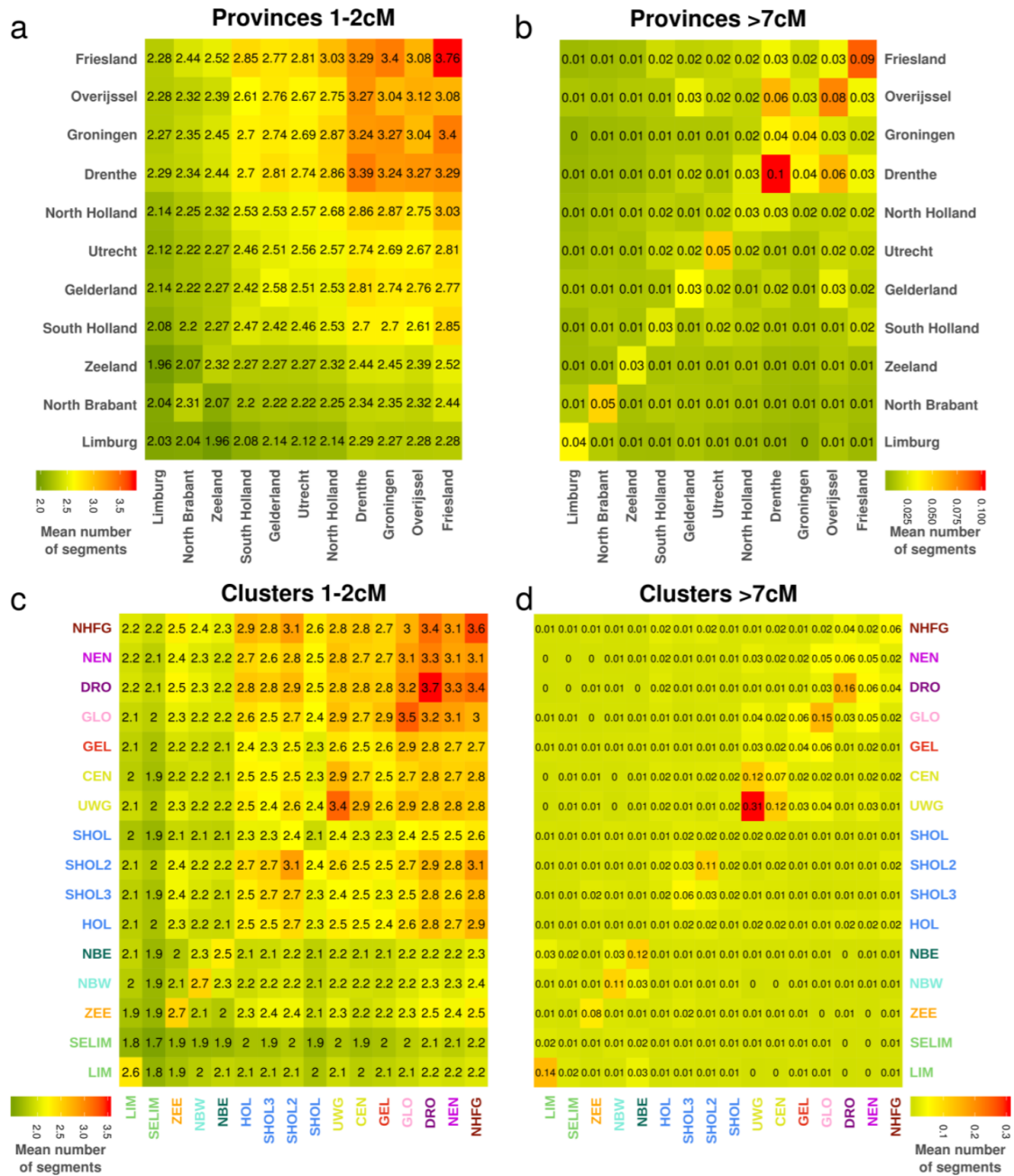


**Appendix Figure 4.3: Dutch and European ADMIXTURE CV-error plot.**

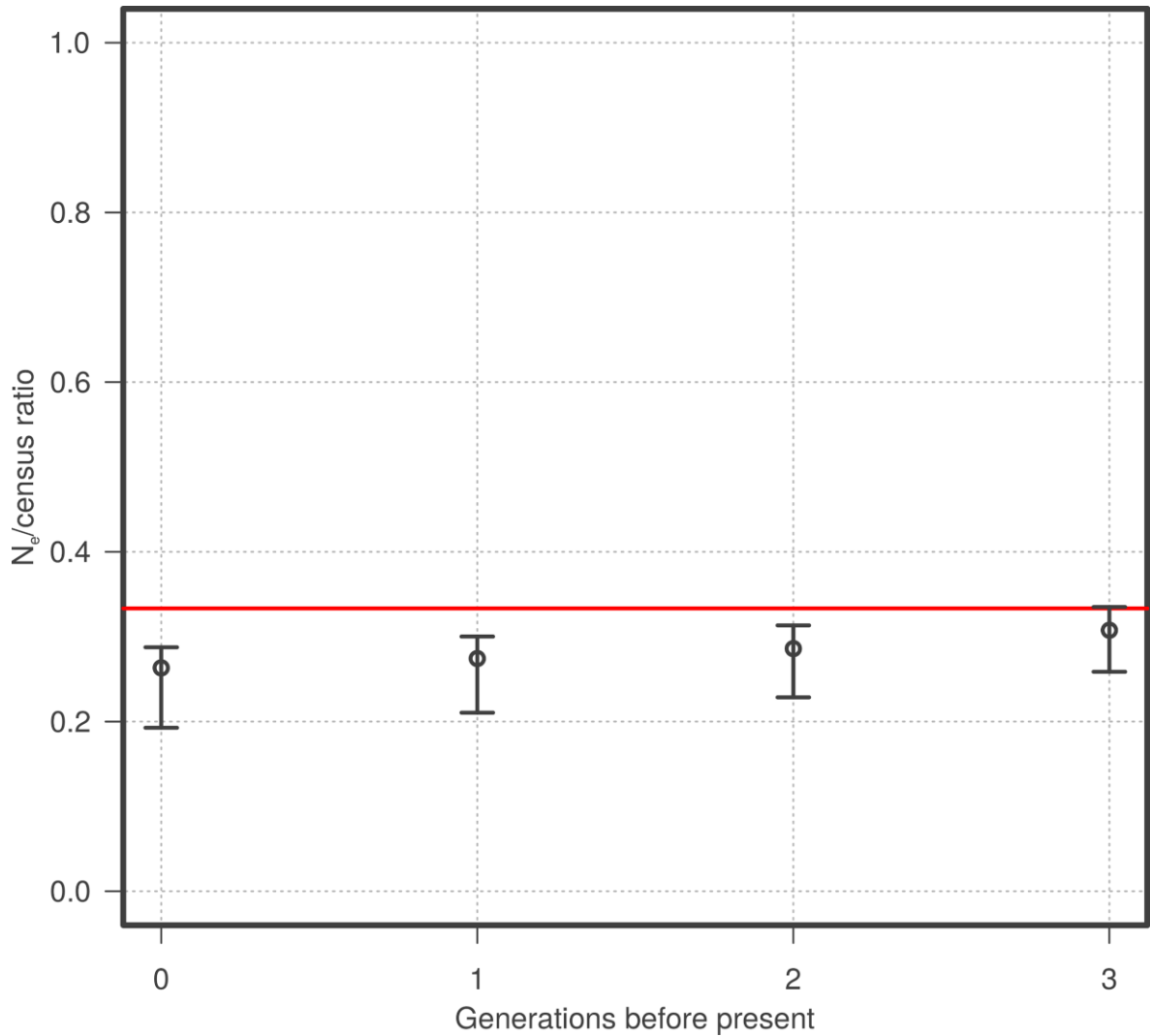
Displays the cross-validation (CV) error for ADMIXTURE run with 1-10 components.

Models with 4 and 5 components are tied for lowest CV error, suggesting choice of either is suitable.

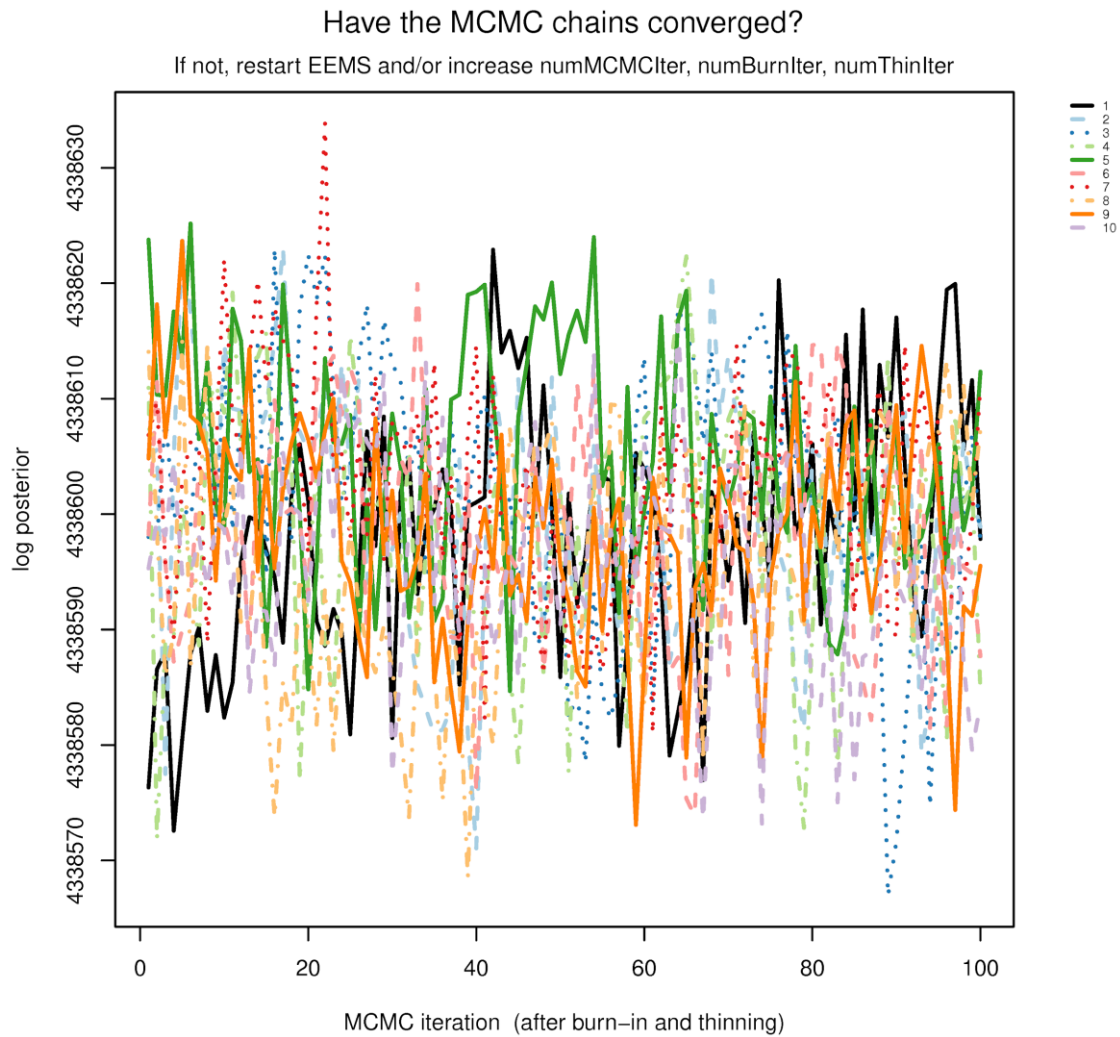
(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



**Appendix Figure 4.4: Old and recent IBD sharing per province and per cluster.** Average sharing of IBD segments between provinces and clusters respectively is described for old (short) segments (1-2 cM) in (a) provinces and (c) clusters and for recent (long) segments (>7cM) in (b) provinces and (d) clusters. Average sharing of old (short) segments is enriched in northern provinces and clusters (a and c). Average sharing of recent (long) segments is higher on average within clusters than within provinces, indicating haplotypic clustering captures marginally more recent ancestry. (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

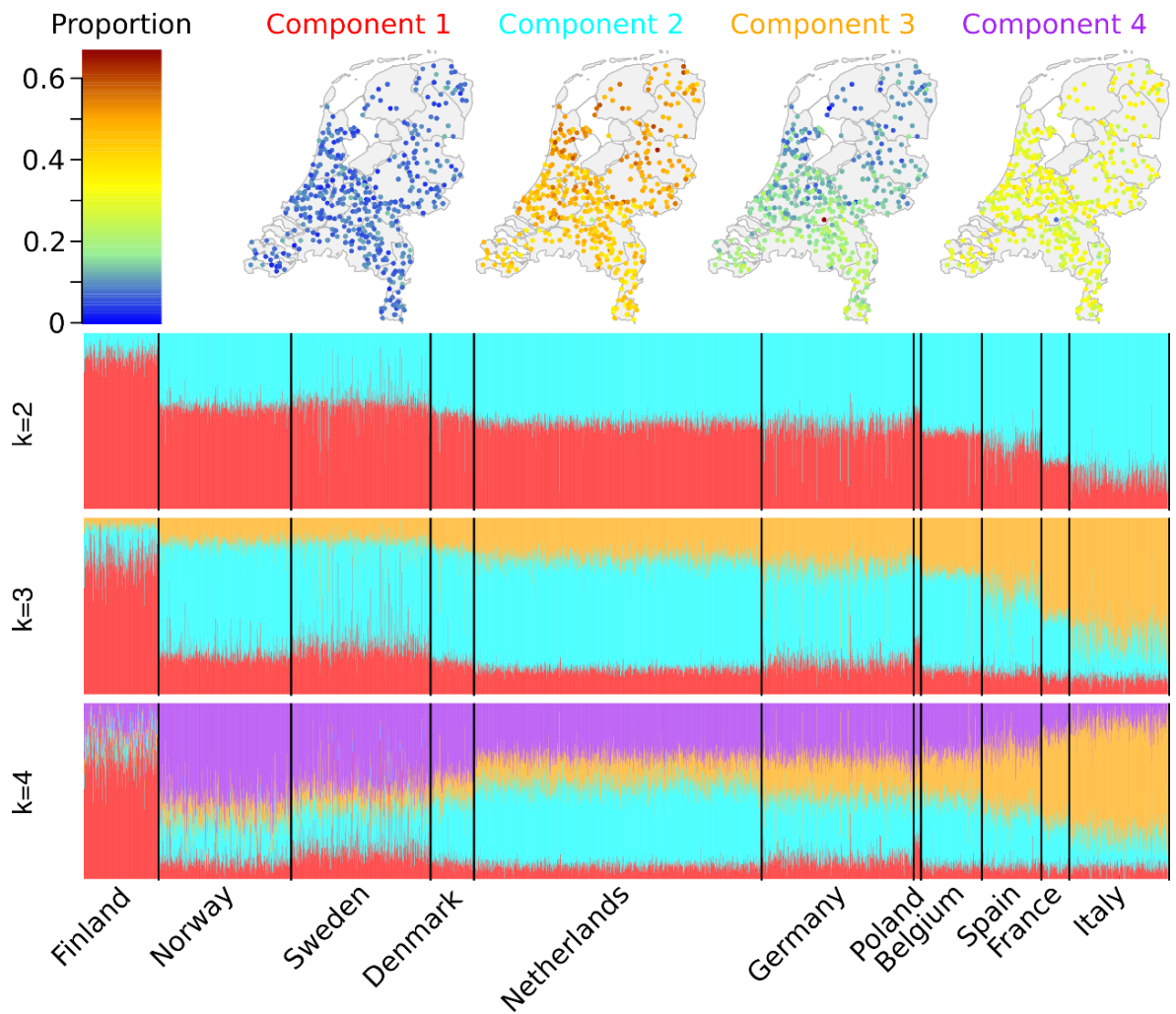


**Appendix Figure 4.5: Ratio of estimated  $N_e$ /Census is stable over the past 3 generations.** The red line at 0.33 corresponds to the expected ratio of  $N_e$  to census if lifespan is 3 times the generation time. Point estimates represent the ratio of estimated  $N_e$  to census value for a given generation in the full Netherlands dataset, while error bars represent 95% confidence interval for this value calculated using 80 bootstrap resamples in IBDNe (note this is not necessarily symmetric on the point estimate). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



**Appendix Figure 4.6: Convergence of MCMC chains for EEMS run in The Netherlands.**

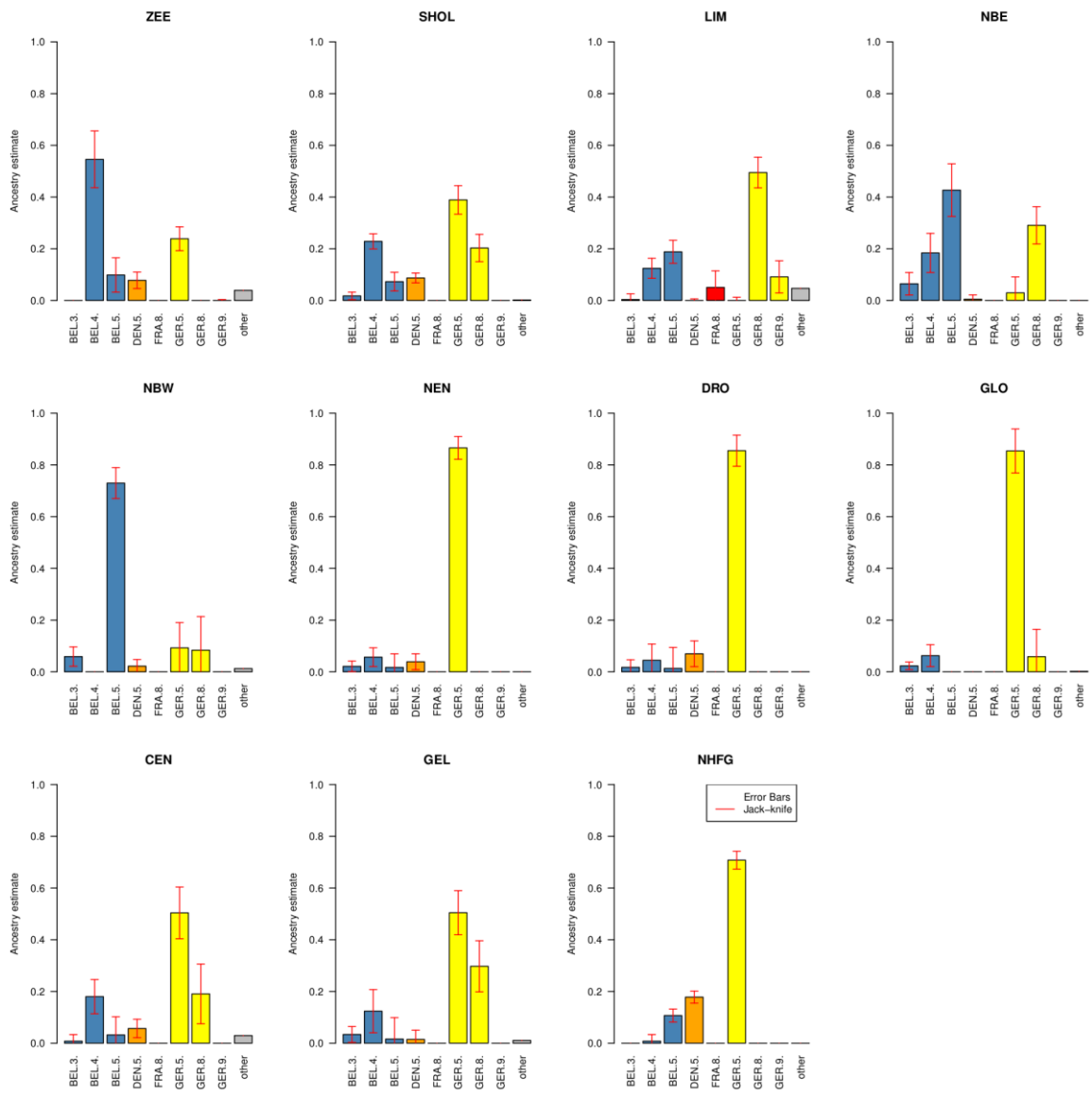
10 independently seeded MCMC chains reach approximate convergence.  
(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



**Appendix Figure 4.7: ADMIXTURE modelling for Dutch and European samples.**

Maps depict the regional breakdown of ADMIXTURE components for  $k=4$  split. Dutch samples have a high value for admixture component 2, which is next highest in Germany and Belgium. Components 2 and 3 show opposing north-south gradients in the Netherlands, with component 2 highest in the north and component 3 highest in the south. Component 3 is best represented in southern European countries such as Italy. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

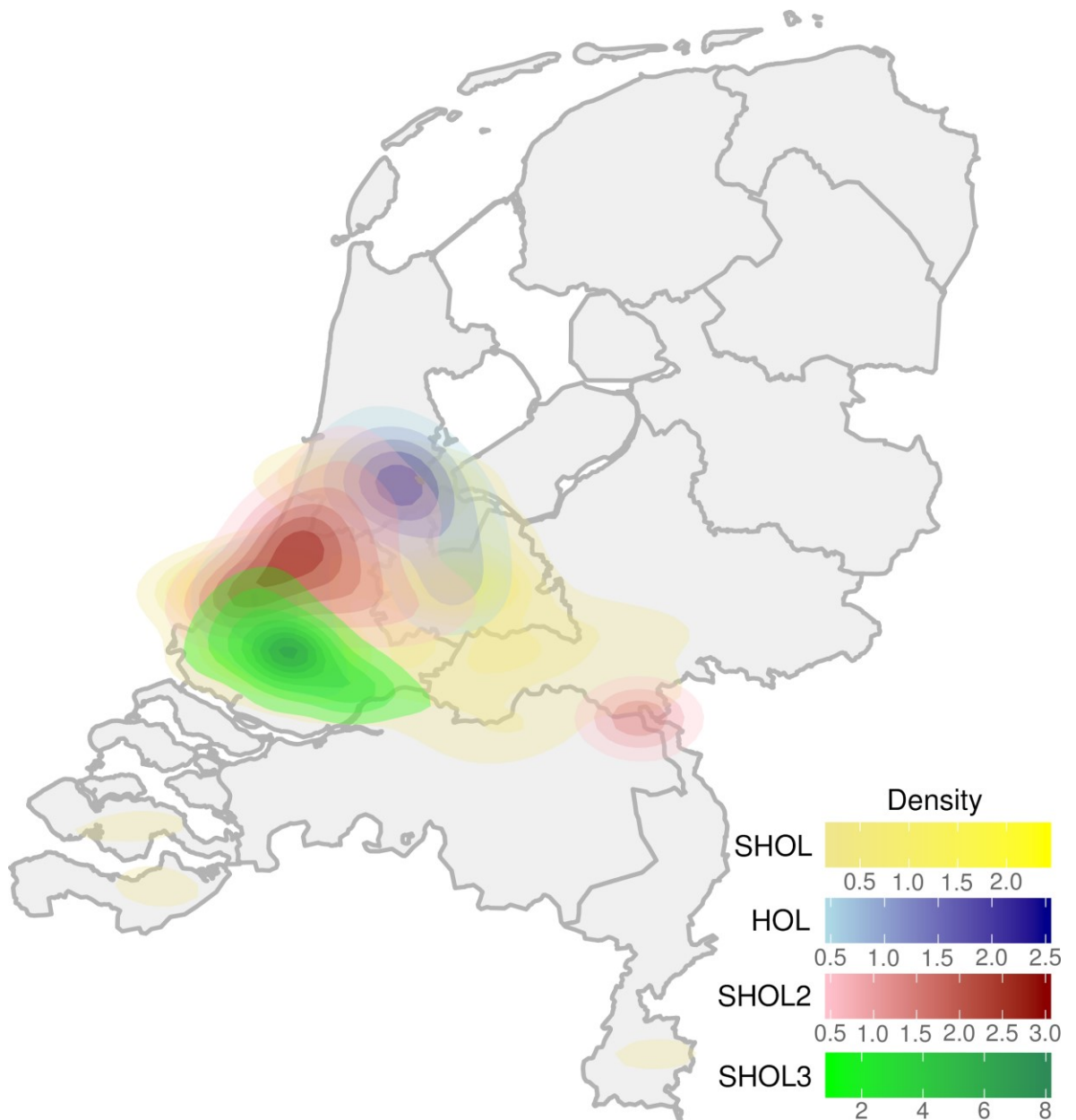


**Appendix Figure 4.8: Ancestry profile per Dutch cluster group.**

Bar charts displaying the GLOBETROTTER estimated European ancestry contribution profile for Dutch cluster groups (Defined in Figure 4.1) from clusters of 4,514 European samples (Appendix Figure 3.5). Only donors that make at least a 5% contribution to at least one Dutch cluster are displayed with the remaining proportions subsumed into the “other” category. Error bars represent a jack-knife approach (Montinaro et al. 2015) leaving one chromosome out (22 resamples). Label abbreviations: BEL, Belgium; Den, Denmark; FRA, France; GER, Germany.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



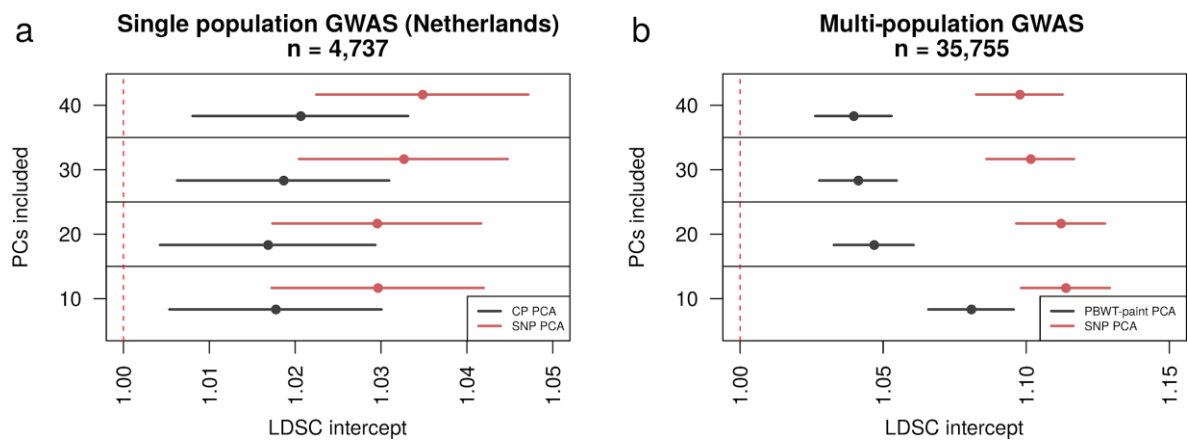


**Appendix Figure 4.9: Geographic distribution of South Holland clusters from the SHOL cluster group.**

Plotted are the 2D kernel density estimates for the geographic spread of samples from clusters SHOL (yellow), HOL (blue), SHOL2 (red), and SHOL3 (green) which form the SHOL cluster group in Figure 4.1. Kernel density estimates were calculated using the `stat_density2d` function in `ggplot2` (R version 3.2.3) with default settings. >80% of samples are contained within plotted polygons for each cluster. Notably, although overlapping, three of the four clusters show quite distinct geographic ranges. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>). (Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))



## Appendix material for Chapter 5



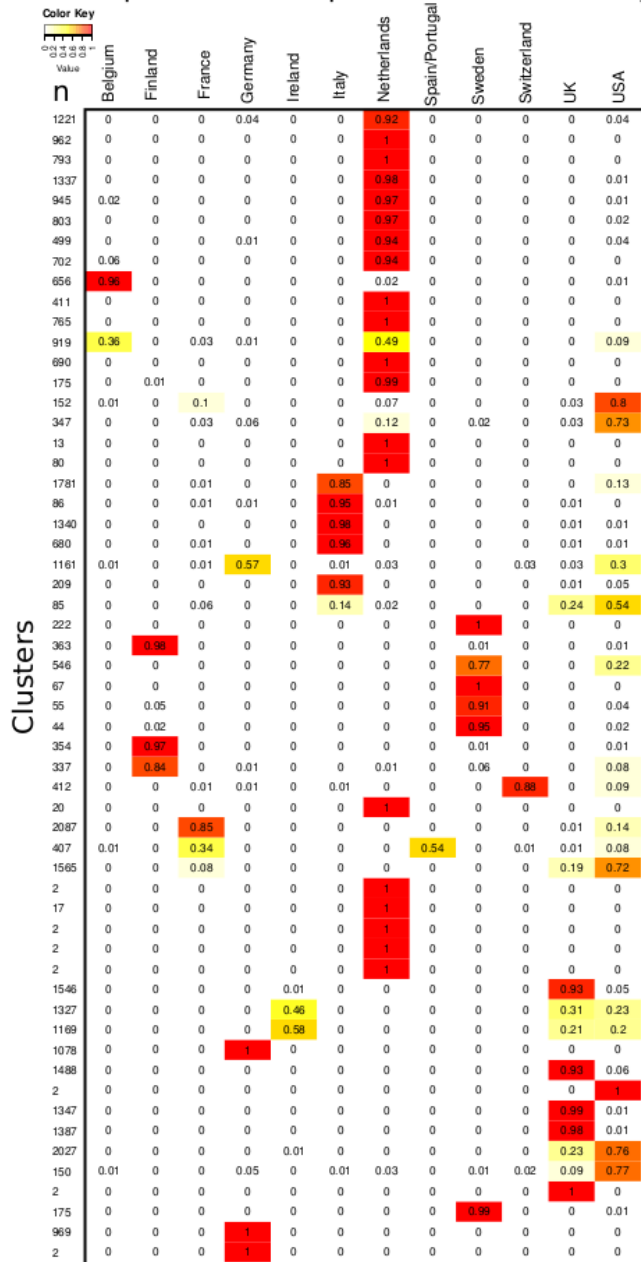
**Appendix Figure 5.1: LDSC intercepts from ALS GWAS using haplotype PCs vs SNP PCs.**

Displays LDSC intercepts (points) and 95% confidence intervals (whiskers) for (a) a single population GWAS and (b) a multi-population GWAS of ALS, fitting PCs calculated from haplotype sharing matrices (black) and SNPs (red) as covariates. Using PCs from haplotype sharing matrices reduces the LDSC intercept relative to using SNP PCs, suggesting haplotype sharing matrices correct for confounding not captured by SNP PCA. Error bars represent 95% confidence intervals centred on the LDSC intercepts. Abbreviations: CP PCA, ChromoPainter PCA; LDSC, LD score regression; PBWT-paint, positional Burrows-Wheeler transform-paint.

NB: This analysis was run with a small number of individuals with relatedness over 0.075 removed (reviewer request). It is clear from comparison with Table 5.1 and Figure 5.11 for the 20 PC run that this exclusion has no effect on the results, hence the analysis is expected to be unchanged by this exclusion.

(Figure reprinted from Byrne et al. (R. P. Byrne et al. 2020))

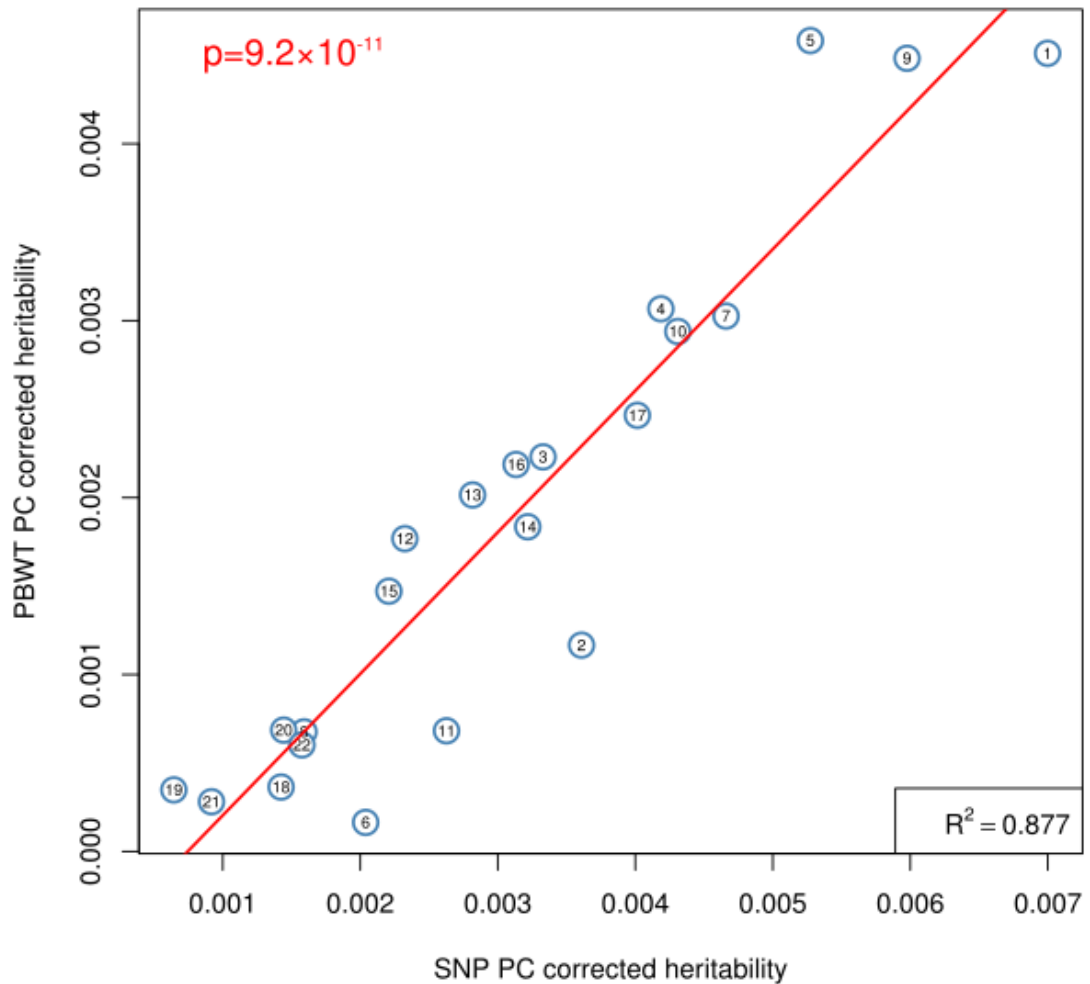
### Louvain clustering iteration 3: Proportion of samples from each country



**Appendix Figure 5.2: Louvain clustering third iteration breakdown.**

A breakdown of proportions of each cluster from the third iteration of Louvain community detection on the multi-population dataset (n=35,985). Each row represents a single cluster, with proportions of samples in that cluster from each country recorded. Notably most clusters are composed almost entirely of samples from a single country or two related countries with known immigration between them (e.g. Cluster on row 12 - Belgium and the Netherlands). Several sub country clusters are detected in the Netherlands, Italy, Finland, Sweden and the UK, many of which have recorded examples of finescale population structure from single population studies.

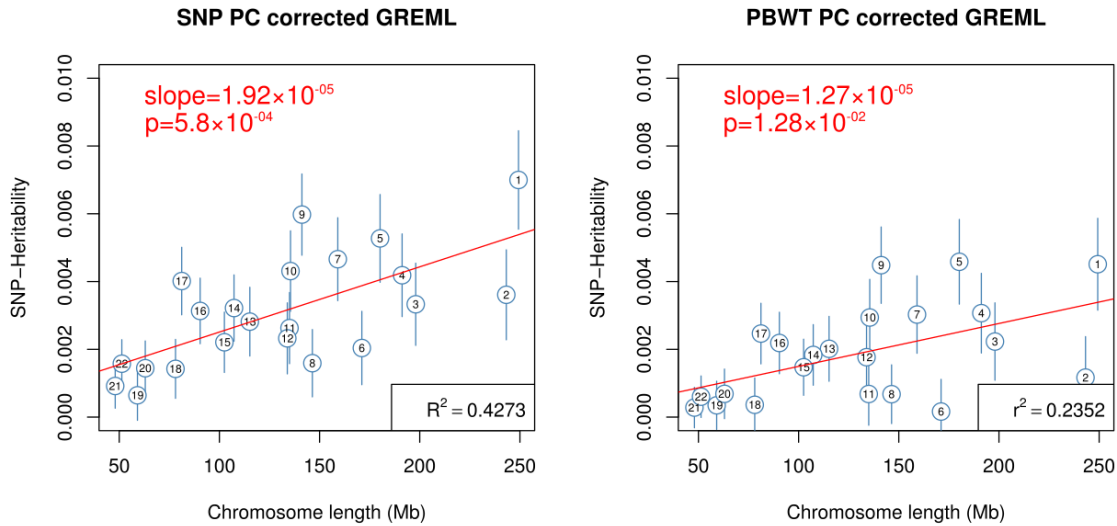
### PBWT-paint PC corrected vs SNP PC corrected GREML



**Appendix Figure 5.3: Correlation of per chromosome ALS heritability estimates corrected using PBWT-paint PCs and SNP PCs.**

Per chromosome heritability estimates for ALS calculated using GREML with 20 PBWT PCs as covariates (Y-axis) are regressed against estimates with 20 SNP PCs as covariates (X-axis). Despite consistently lower estimates from the PBWT-paint corrected model, there is a strong correlation between the two methods, indicating that there are few major differences estimated in the distribution of heritability from the two methods.

# Heritability vs chromosome length



**Appendix Figure 5.4: Heritability estimates corrected with PBWT-paint retain evidence of polygenicity.**

Per chromosome heritability estimates (GREML) corrected using 20 SNP PCs as covariates (left) or 20 PBWT-paint PCs (right) are plotted against chromosome length to investigate the distribution of heritability across the genome from each method. Both methods demonstrate a significant linear relationship between chromosome length and heritability, suggesting heritability is driven by many loci relatively evenly spread across the genome (i.e. polygenicity). Slopes from the two methods are practically identical, indicating there is equal evidence of polygenic signal in ALS from both the SNP PC corrected model (left) and PBWT-paint PC corrected model (right).

**Appendix Table 5.1: Sample breakdown for the 2016 ALS GWAS dataset.**

<b>Stratum</b>	<b>Country</b>	<b>Cases</b>	<b>Controls</b>	<b>Total</b>	<b>Platform</b>
sBE1	Belgium	299	317	616	Illumina370K
sBE2	Belgium	205	242	447	IlluminaOmniExpress
sFIN1	Finland	378	378	756	Illumina370K
sFIN2	Finland	135	97	232	Illumina1M+IlluminaOmniExpress
sFR1	France	155	654	809	Illumina317K
sFR2	France	327	1005	1332	Illumina550K+IlluminaOmniExpress
sGER1	Germany	518	258	776	Illumina550K
sGER2	Germany	1399	648	2047	Illumina550K+IlluminaCoreExome
sIR1	Ireland	308	331	639	Illumina550K+Illumina610K
sIR2	Ireland	264	443	707	IlluminaOmniExpress
sIT1	Italy	382	244	626	Illumina610K+Illumina550K
sIT2	Italy	290	93	383	IlluminaOmniExpress
sIT3	Italy	1715	1075	2790	Illumina660W
sNL1	Netherlands	423	420	843	Illumina317K
sNL2	Netherlands	145	4882	5027	Illumina370K+Illumina550K
sNL3	Netherlands	952	1829	2781	IlluminaOmniExpress
sNL4	Netherlands	596	533	1129	Illumina2.5M
sIB1	Portugal+Spain	126	99	225	IlluminaOmniExpress
sSW1	Sweden	288	268	556	Illumina370K
sSW2	Sweden	232	235	467	IlluminaOmniExpress
sSWISS1	Switzerland	203	221	424	IlluminaOmniExpress
sUK1	United Kingdom	168	159	327	Illumina317K
sUK2	United Kingdom	614	2687	3301	Illumina550K+Illumina1M
sUK3	United Kingdom	1032	2502	3534	Illumina1M+IlluminaOmniExpress
sUS1	United states	598	1339	1937	Illumina317K+Illumina370K
sUS2	United states	266	513	779	Illumina550K
sUS3	United states	559	2003	2562	Illumina1M+IlluminaOmniExpress

Breakdown of sample numbers, Country of origin and genotyping platform for each stratum from the 2016 ALS GWAS dataset (n=36,052) (van Rheenen et al. 2016) used in Chapter 5. Table compiled from Supplementary tables 1,4 and 5 from the source paper (van Rheenen et al. 2016)

Appendix Table 5.2: Geographic clustering of SNP and cp-PCs for Dutch only dataset.

PC	SNP Pcs		ChromoPainter Pcs	
	Moran's I	p	Moran's I	p
1	3.03E-01	<b>1.00E-04</b>	3.10E-01	<b>1.00E-04</b>
2	1.76E-01	<b>1.00E-04</b>	2.39E-01	<b>1.00E-04</b>
3	1.22E-01	<b>1.00E-04</b>	1.73E-01	<b>1.00E-04</b>
4	7.81E-02	<b>1.00E-04</b>	2.12E-01	<b>1.00E-04</b>
5	5.37E-02	<b>1.00E-04</b>	1.38E-01	<b>1.00E-04</b>
6	6.65E-03	5.08E-02	1.40E-01	<b>1.00E-04</b>
7	2.16E-04	5.36E-01	1.08E-01	<b>1.00E-04</b>
8	5.02E-03	9.47E-02	1.42E-01	<b>1.00E-04</b>
9	6.40E-03	5.56E-02	1.08E-01	<b>1.00E-04</b>
10	1.09E-02	7.00E-03	1.08E-01	<b>1.00E-04</b>
11	-2.17E-03	9.19E-01	1.05E-01	<b>1.00E-04</b>
12	2.26E-03	2.87E-01	4.23E-02	<b>1.00E-04</b>
13	4.62E-03	1.21E-01	9.66E-02	<b>1.00E-04</b>
14	5.80E-05	5.57E-01	6.91E-02	<b>1.00E-04</b>
15	2.55E-03	2.81E-01	9.45E-02	<b>1.00E-04</b>
16	7.21E-03	3.77E-02	7.07E-02	<b>1.00E-04</b>
17	9.73E-04	4.37E-01	7.99E-02	<b>1.00E-04</b>



18	-5.15E-03	5.81E-01	7.19E-02	<b>1.00E-04</b>
19	-3.70E-03	8.20E-01	9.60E-02	<b>1.00E-04</b>
20	9.29E-03	1.50E-02	3.49E-02	<b>1.00E-04</b>

Geographic clustering of SNP and ChromoPainter PCs for 1,352 samples from the Netherlands measured using Moran's I. Significant instances of clustering passing the Bonferroni adjusted significance threshold ( $p < 0.0025$ ) are highlighted in bold. Notably there is significant evidence of clustering for at least 20 PCs when using ChromoPainter PCs, but only 5 PCs when using SNP based PCs.

**Appendix Table 5.3: GREML per chromosome heritability estimates from 2016 ALS GWAS under SNP and PBWT-paint PC corrections.**

Chr	SNP PC corrected		PBWT PC corrected		p difference
	h <sup>2</sup>	SE	h <sup>2</sup>	SE	
1	7.00E-03	1.45E-03	4.51E-03	1.36E-03	7.65E-01
2	3.61E-03	1.33E-03	1.17E-03	1.21E-03	5.19E-01
3	3.33E-03	1.21E-03	2.23E-03	1.14E-03	7.84E-01
4	4.19E-03	1.22E-03	3.07E-03	1.18E-03	8.29E-01
5	5.27E-03	1.30E-03	4.58E-03	1.25E-03	9.21E-01
6	2.04E-03	1.08E-03	1.64E-04	9.50E-04	3.59E-01
7	4.66E-03	1.22E-03	3.03E-03	1.14E-03	7.69E-01
8	1.59E-03	9.93E-04	6.76E-04	8.66E-04	5.97E-01
9	5.98E-03	1.20E-03	4.48E-03	1.13E-03	8.42E-01
10	4.31E-03	1.19E-03	2.94E-03	1.13E-03	7.93E-01
11	2.63E-03	1.05E-03	6.82E-04	9.21E-04	4.74E-01
12	2.32E-03	1.05E-03	1.77E-03	1.01E-03	8.49E-01
13	2.82E-03	1.01E-03	2.02E-03	9.57E-04	8.17E-01
14	3.22E-03	9.78E-04	1.84E-03	8.94E-04	7.09E-01
15	2.21E-03	8.90E-04	1.47E-03	8.28E-04	7.81E-01
16	3.13E-03	9.67E-04	2.19E-03	9.07E-04	8.04E-01
17	4.02E-03	9.96E-04	2.47E-03	8.94E-04	7.42E-01

18	1.42E-03	8.69E-04	3.64E-04	7.91E-04	4.71E-01
19	6.44E-04	7.38E-04	3.48E-04	7.11E-04	6.86E-01
20	1.44E-03	8.04E-04	6.86E-04	7.31E-04	6.36E-01
21	9.18E-04	6.55E-04	2.81E-04	5.96E-04	5.07E-01
22	1.58E-03	7.08E-04	6.02E-04	6.12E-04	5.64E-01

Estimates of SNP heritability for ALS from the 2016 ALS GWAS dataset partitioned by chromosome, using 20 SNP PCs or 20 PBWT-paint PCs as covariates in the model. No chromosomes show significant differences ( $p$  difference) in heritability estimates across methods.

# Appendix 2 – Publications

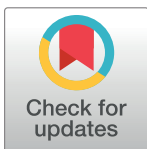
RESEARCH ARTICLE

# Insular Celtic population structure and genomic footprints of migration

Ross P. Byrne<sup>1\*</sup>, Rui Martiniano<sup>2,3</sup>, Lara M. Cassidy<sup>2</sup>, Matthew Carrigan<sup>4</sup>, Garrett Hellenthal<sup>5</sup>, Orla Hardiman<sup>6</sup>, Daniel G. Bradley<sup>2</sup>, Russell L. McLaughlin<sup>1\*</sup>

**1** Complex Trait Genomics Laboratory, Smurfit Institute of Genetics, School of Genetics and Microbiology, Trinity College Dublin, College Green, Dublin, Republic of Ireland, **2** Population Genetics Laboratory, Smurfit Institute of Genetics, School of Genetics and Microbiology, Trinity College Dublin, College Green, Dublin, Republic of Ireland, **3** Wellcome Trust Sanger Institute, Cambridge, United Kingdom, **4** Ocular Genetics Unit, Smurfit Institute of Genetics, School of Genetics and Microbiology, Trinity College Dublin, College Green, Dublin, Republic of Ireland, **5** UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, **6** Academic Unit of Neurology, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Republic of Ireland

\* [rbyrne5@tcd.ie](mailto:rbyrne5@tcd.ie) (RPB); [mclaugr@tcd.ie](mailto:mclaugr@tcd.ie) (RLM)



**OPEN ACCESS**

**Citation:** Byrne RP, Martiniano R, Cassidy LM, Carrigan M, Hellenthal G, Hardiman O, et al. (2018) Insular Celtic population structure and genomic footprints of migration. *PLoS Genet* 14(1): e1007152. <https://doi.org/10.1371/journal.pgen.1007152>

**Editor:** Daniel Falush, Max Planck Institute for Evolutionary Anthropology, GERMANY

**Received:** September 14, 2017

**Accepted:** December 13, 2017

**Published:** January 25, 2018

**Copyright:** © 2018 Byrne et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Irish data is fully available from the European Genome-phenome Archive (EGA; <https://ega-archive.org/>) under accession number EGAS00001002769. The UK PoBI and European WTCCC MS datasets were generated by third parties (see original papers) and are available from the European Genome-phenome Archive (EGA; <https://ega-archive.org/>) under accession numbers EGAD00010000632 and EGAD00000000120 respectively.

## Abstract

Previous studies of the genetic landscape of Ireland have suggested homogeneity, with population substructure undetectable using single-marker methods. Here we have harnessed the haplotype-based method fineSTRUCTURE in an Irish genome-wide SNP dataset, identifying 23 discrete genetic clusters which segregate with geographical provenance. Cluster diversity is pronounced in the west of Ireland but reduced in the east where older structure has been eroded by historical migrations. Accordingly, when populations from the neighbouring island of Britain are included, a west-east cline of Celtic-British ancestry is revealed along with a particularly striking correlation between haplotypes and geography across both islands. A strong relationship is revealed between subsets of Northern Irish and Scottish populations, where discordant genetic and geographic affinities reflect major migrations in recent centuries. Additionally, Irish genetic proximity of all Scottish samples likely reflects older strata of communication across the narrowest inter-island crossing. Using GLOBETROTTER we detected Irish admixture signals from Britain and Europe and estimated dates for events consistent with the historical migrations of the Norse-Vikings, the Anglo-Normans and the British Plantations. The influence of the former is greater than previously estimated from Y chromosome haplotypes. In all, we paint a new picture of the genetic landscape of Ireland, revealing structure which should be considered in the design of studies examining rare genetic variation and its association with traits.

## Author summary

A recent genetic study of the UK (People of the British Isles; PoBI) expanded our understanding of population history of the islands, using newly-developed, powerful techniques that harness the rich information embedded in chunks of genetic code called haplotypes. These methods revealed subtle regional diversity across the UK, and, using genetic data

**Funding:** This study received support from the Motor Neurone Disease Association of England, Wales and Northern Ireland (957-799; <https://www.mndassociation.org/>), Science Foundation Ireland (15/SPP/3244; <http://www.sfi.ie/>). RM has been supported by the Marie Curie ITN (289966; [https://ec.europa.eu/research/mariecurieactions/about/innovative-training-networks\\_en](https://ec.europa.eu/research/mariecurieactions/about/innovative-training-networks_en)) and by the Wellcome Trust (098051; <https://wellcome.ac.uk/>). GH is jointly funded by the Wellcome Trust and Royal Society (098386/Z/12/Z; <https://wellcome.ac.uk/> and <https://royalsociety.org/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** OH has received speaking honoraria from Janssen Cilag, Biogen Idec, Sanofi Aventis, Novartis, and Merck-Serono. She has been a member of advisory panels for Biogen Idec, Allergan, Ono Pharmaceuticals, Novartis, Cytokinetics, and Sanofi Aventis. She serves as Editor-in-Chief of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. All other authors have declared that no competing interests exist.

alone, timed key migration events into southeast England and Orkney. We have extended these methods to Ireland, identifying regional differences in genetics across the island that adhere to geography at a resolution not previously reported. Our study reveals relative western diversity and eastern homogeneity in Ireland owing to a history of settlement concentrated on the east coast and longstanding Celtic diversity in the west. We show that Irish Celtic diversity enriches the findings of PoBI; haplotypes mirror geography across Britain and Ireland, with relic Celtic populations contributing greatly to haplotypic diversity. Finally, we used genetic information to date migrations into Ireland from Europe and Britain consistent with historical records of Viking and Norman invasions, demonstrating the signatures of these migrations on the modern Irish genome. Our findings demonstrate that genetic structure exists in even small isolated populations, which has important implications for population-based genetic association studies.

## Introduction

Situated at the northwestern edge of Europe, Ireland is the continent's third largest island, with a modern-day population of approximately 6.4 million. The island is politically partitioned into the Republic of Ireland and Northern Ireland, with the latter forming part of the United Kingdom (UK) alongside the neighbouring island of Britain. Alternative divisions separate Ireland into four provinces reflecting early historical divisions: Ulster to the north, including Northern Ireland; Leinster (east); Munster (south) and Connacht (west). Humans have continuously inhabited Ireland for around 10,000 years [1], though it is not until after the demographic upheavals of the Early Bronze Age (circa 2200 BCE), that strong genetic continuity between ancient and modern Irish populations is observed [2]. Linguistically, the island's earliest attested language forms part of the Insular Celtic family, specifically the Gaelic branch, whose historic range also extended to include many regions of Scotland, via maritime connections with Ulster [3,4]. A second branch of Insular Celtic, the Brittonic languages, had been spoken across much of Britain up until the introduction of Anglo-Saxon in the 5th and 6th centuries, by which time they were diversifying into Cornish, Welsh and Cumbric dialects [5].

Since the establishment of written history, numerous settlements and invasions of Ireland from the neighbouring island of Britain and continental Europe have been recorded. This includes Norse-Vikings (9th-12th century), especially in east Leinster, and Anglo-Normans (12th-14th century), who invaded through Wexford in the southeast and established English rule mainly from an area later called the Pale in northeast Leinster [6]. There has also been continuous movement of people from Britain, in particular during the 16-17th century Plantation periods during which Gaelic and Norman lands were systematically colonized by English and Scottish settlers. These events had a particularly enduring impact in Ulster in comparison with other planted regions such as Munster. As with the previous Norman invasion, the less fertile west of the country (Connacht) remained largely untouched during this period.

The genetic contributions of these migratory events cannot be considered mutually independent, given that they derive from either related Germanic populations (such as the Vikings and their purported Norman descendants) or from other Celtic populations inhabiting Britain, which had themselves been subjected to mass Germanic influx from Anglo-Saxon migrations and later Viking and Norman invasions [7]. Moreover, each movement of people originated from northern Europe, a region which had witnessed a mass homogenizing of genetic variation during the migrations of the Early Bronze Age, possibly linked to Indo-European

language spread. [8,9]. However, each event had a geographic and temporal focal point on the island, which may be detectable in local population structure.

Previous genome-wide surveys have detected little to no structure in Ireland using methods such as principal component analysis (PCA) on independent markers, concluding that the Irish population is genetically homogenous [10]. However, runs of homozygosity are relatively long and frequent in Ireland [10] and correlate negatively with population density and diversity of grandparental origins [11], suggesting that low ancestral mobility may have preserved regional genetic legacies within Ireland, which may be detectable in modern genomes as local population structure embedded within haplotypes. This is further supported by the restricted regional distributions of certain Y chromosome haplotypes [12,13].

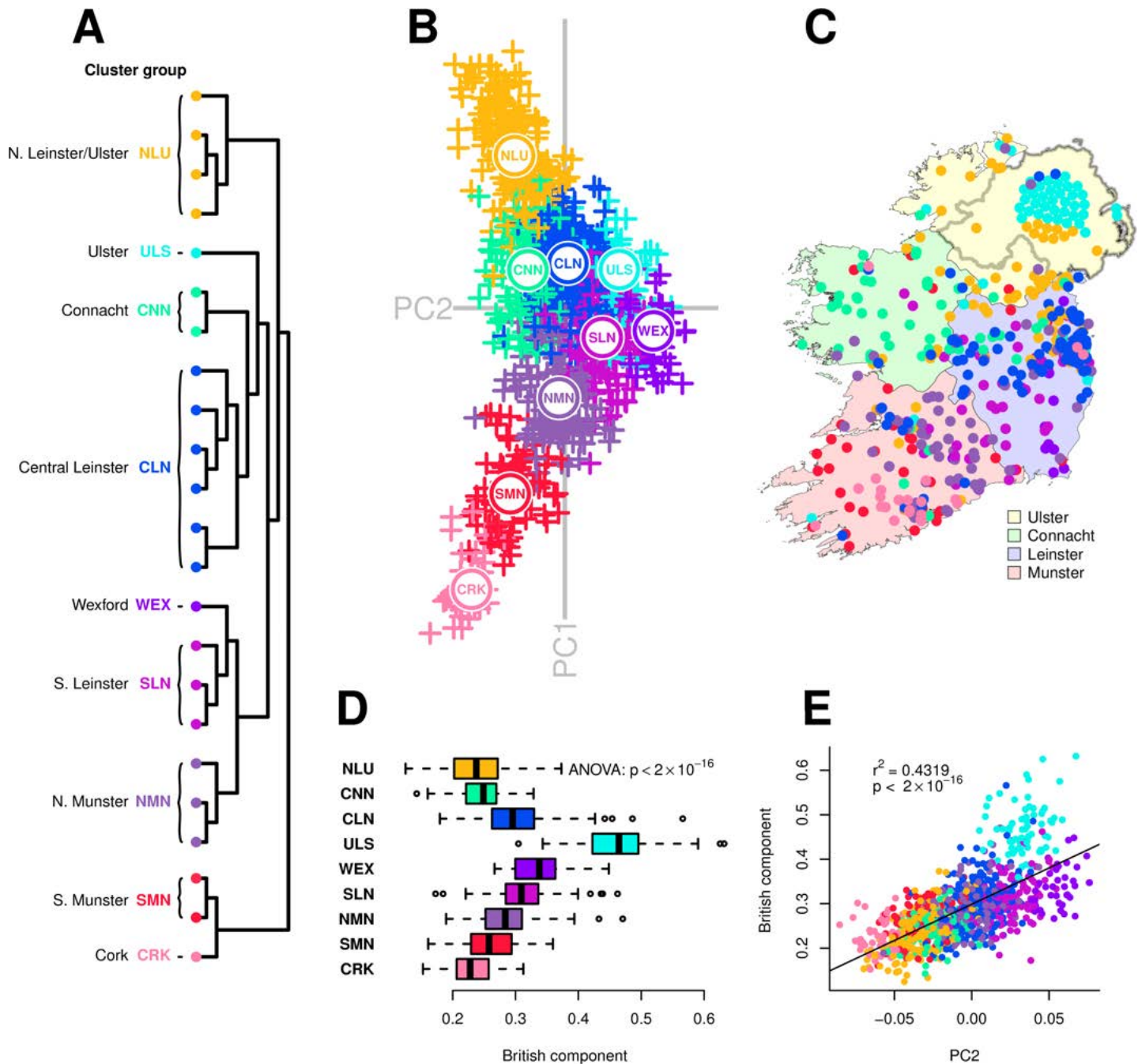
The haplotype-based methods ChromoPainter and fineSTRUCTURE [14] were recently used to uncover hidden genetic structure among the people of modern Britain [7]. These approaches exploit the rich information available within haplotypes (usually statistically phased) to identify clusters of genetically distinct individuals with a resolution that could not be attained using single-marker methods. In doing so, the People of the British Isles (PoBI) study was able to identify discrete genetic clusters of individuals that strongly segregate with geographical regions within Britain, though notably, structure was undetectable across a large southeastern portion of the island. However, although this study sampled over 2,000 individuals, only 44 were from Northern Ireland with none from the remainder of the island. Ireland was also excluded from admixture and ancestry analyses due to the confounding effects of the island acting as “a source and a sink for ancestry from the UK”. With this focus on a single island, the PoBI study has an obvious limit, despite its title.

Here, we have used the methods of the PoBI study to explore fine-grained Irish population substructure. We first investigate Ireland on its own, then we consider the genetic substructure observed on the island in the context of Britain and continental Europe. Using modern individuals from these two sources as surrogates for historical populations, we apply the GLOBE-TROTTER model to infer admixture events into Ireland and we consider these in the context of historically recorded invasions and migrations. Our inclusion of Irish data with previously-published data from Britain presents a more complete representation of genetic ancestry in the contemporary populations of the British Isles, providing a comprehensive population genetic perspective of the peopling of these islands.

## Results and discussion

### Celtic population structure in Ireland

We used ChromoPainter [14] to identify haplotypic similarities within a genome-wide single nucleotide polymorphism (SNP) dataset of individuals from the Republic of Ireland and Northern Ireland ( $n = 1,035$ , including 44 from the PoBI study), in which local geographic origin was known for a subset ( $n = 588$ ). Clustering the resulting coancestry matrix using fineSTRUCTURE identified 23 clusters, demonstrating local population structure within Ireland to a level not previously reported, with apparent geographical, sociopolitical and ancestral correlates (Fig 1). All clusters were robustly defined, with total variation distance (TVD)  $p$ -values less than 0.001 (S1 and S2 Tables). We projected the ChromoPainter coancestry matrix in lower-dimensional space using principal component analysis (PCA) and, to ease interpretation and for visual brevity with labels, we defined 9 cluster groups that formed higher order clades in the fineSTRUCTURE dendrogram, overlapped in PC space and were sampled from geographically contiguous regions. These cluster groups also showed robust definition by TVD analysis (S3 Table and S4 Table), suggesting they represent a meaningful grouping of the data. ChromoPainter PCA revealed a tight relationship between haplotypic similarity and



**Fig 1. Fine-grained population structure in Ireland.** (A) fineSTRUCTURE clustering dendrogram for 1,035 Irish individuals. Twenty-three clusters are defined, which are combined into cluster groups for clusters that are neighbouring in the dendrogram, overlapping in principal component space (B) and sampled from regions that are geographically contiguous. Details for each cluster in the dendrogram are provided in S1 Fig. (B) Principal components analysis (PCA) of haplotypic similarity, based on ChromoPainter coancestry matrix for Irish individuals. Points are coloured according to cluster groups defined in (A); the median location of each cluster group is plotted. (C) Map of Ireland showing the sampling location for a subset of 588 individuals analysed in (A) and (B), coloured by cluster group. Points have been randomly jittered within a radius of 5 km to preserve anonymity. Precise sampling location for 44 Northern Irish individuals from the People of the British Isles dataset was unknown; these individuals are plotted geometrically in a circle. The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>). (D) “British admixture component” (ADMIXTURE estimates;  $k = 2$ ) for Irish cluster groups. This component has the largest contribution in ancient Anglo-Saxons and the SEE cluster. (E) Linear regression of principal component 2 (B) versus British admixture component ( $r^2 = 0.43$ ;  $p < 2 \times 10^{-16}$ ). Points are coloured by cluster group. (Standard error for ADMIXTURE point estimates presented in S11 Fig.).

<https://doi.org/10.1371/journal.pgen.1007152.g001>



geographical proximity, with ChromoPainter principal component (PC) 1 roughly describing a north to south cline and PC2 largely describing an east to west cline (Fig 1B).

At a high level, both ChromoPainter PCA and fineSTRUCTURE clustering loosely separated the historical provinces of Ireland (Ulster, Leinster, Munster and Connacht) suggesting that these socially constructed territories may have had an impact on genetic structure within Ireland which is deeply embedded in time. Careful inspection of the tree ordering and the PCA revealed more nuanced relations between the provinces; for example south Leinster clusters share more haplotypes with those from north Munster than with their central and north Leinster counterparts. The geographical distribution of this deep subdivision of Leinster resembles pre-Norman territorial boundaries which divided Ireland into fiths (*cúige*), with north Leinster a kingdom of its own known as Meath (*Mide*) [15]. However interpreted, the firm implication of the observed clustering is that despite its previously reported homogeneity, the modern Irish population exhibits genetic structure that is subtly but detectably affected by ancestral population structure conferred by geographical distance and, possibly, ancestral social structure.

ChromoPainter PC1 demonstrated high diversity amongst clusters from the west coast, which may be attributed to longstanding residual ancient (possibly Celtic) structure in regions largely unaffected by historical migration. Alternatively, genetic clusters may also have diverged as a consequence of differential influence from outside populations, as this diversity between western genetic clusters cannot be explained in terms of geographic distance alone. South Munster (SMN) and Cork (CRK) clusters branch off first in the fineSTRUCTURE tree and show distinct separation from their neighbouring north Munster clusters (NMN), indicating that south Munster's haplotypic makeup is more distinct from its neighbouring regions and the remaining regions than any other cluster. TVD analysis supports this observation (S1 Table and S3 Table), with the Cork cluster in particular showing strong differentiation from other clusters. This may reflect the persistent isolating effects of the mountain ranges surrounding the south Munster counties of Cork and Kerry, restricting gene flow with the rest of Ireland and preserving older structure.

In contrast to the west of Ireland, eastern individuals exhibited relative homogeneity; a similar pattern was observed in the PoBI study [7], in which all samples in a large region in southeast England formed a single indivisible cluster of genetically similar individuals comprising almost half the dataset. However, while east coast clusters in Ireland are the largest and demonstrate strong cluster integrity, the largest of these (Central Leinster, CLN) comprises roughly a fifth of our dataset (S1 Fig), hence they are dwarfed proportionally in both number and geographical extent by the southeast England cluster (SEE), suggesting that deeper structure persists in eastern Ireland than in southeast England. The overall pattern of western diversity and eastern homogeneity in Ireland may be explained by increased gene flow and migration into and across the east coast of Ireland from geographically proximal regions, the closest of which is the neighbouring island of Britain.

To explore this, we estimated the extent of admixture per individual in the Irish dataset from Britain, using samples from the PoBI dataset as a reference [7], along with eighteen ancient British individuals from the Iron Age, Roman and Anglo-Saxon periods in northeast and southeast England [16,17]. Using an unsupervised ADMIXTURE analysis [18], we observed that one of the ADMIXTURE clusters ( $k = 2$ ) comprises the totality of ancestry of several Anglo-Saxon individuals and forms the largest proportion in British groups, with varying representation across Irish clusters (S8 Fig). For simplicity we will call this the British component, which was among the lowest for individuals falling in Irish west coast fineSTRUCTURE clusters, including the south Munster and Cork cluster groups (Fig 1D), supporting the interpretation that these regions differ in terms of restricted haplotypic contribution from Britain. Analysis of variance

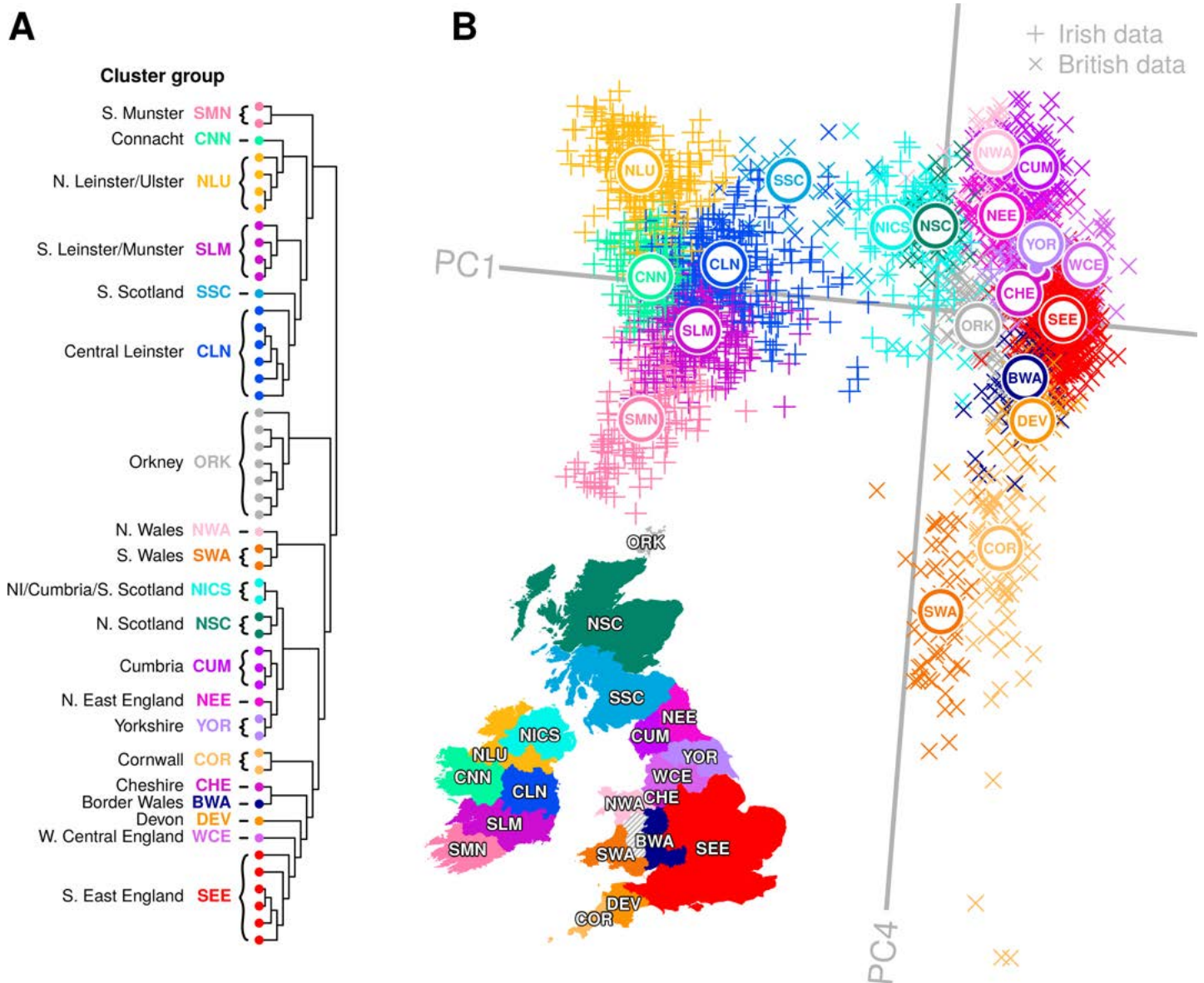
of the British admixture component in cluster groups showed a significant difference ( $p < 2 \times 10^{-16}$ ), indicating a role for British Anglo-Saxon admixture in distinguishing clusters, and ChromoPainter PC2 was correlated with the British component ( $p < 2 \times 10^{-16}$ ), explaining approximately 43% of the variance. PC2 therefore captures an east to west Anglo-Celtic cline in Irish ancestry. This may explain the relative eastern homogeneity observed in Ireland, which could be a result of the greater English influence in Leinster and the Pale during the period of British rule in Ireland following the Norman invasion, or simply geographic proximity of the Irish east coast to Britain. Notably, the Ulster cluster group harboured an exceptionally large proportion of the British component (Fig 1D and 1E), undoubtedly reflecting the strong influence of the Ulster Plantations in the 17th century and its residual effect on the ethnically British population that has remained.

### The genetic structure of the British Isles

The genetic substructure observed in Ireland is consistent with long term geographic diversification of Celtic populations and the continuity shown between modern and Early Bronze Age Irish people [2]. However, this diversity is weaker on the east coast in a manner that correlates with British admixture, suggesting a role for recent migrations in eroding this structure. We therefore further investigated the relationship between Ireland and Britain by generating a ChromoPainter coancestry matrix for all Irish and PoBI data combined ( $n = 3,008$ ). Clustering with fineSTRUCTURE revealed 50 distinct clusters that segregated geographically, both on a cohort-wide and local level (Fig 2). Projecting this coancestry matrix in PC space revealed a striking concordance between haplotypes and geography (sampling regions were defined using Nomenclature of Territorial Units for Statistics 2010; [19]) for ChromoPainter PCs 1 and 4, reminiscent of previous observations for single marker-based summaries of genetic variation within European populations [20].

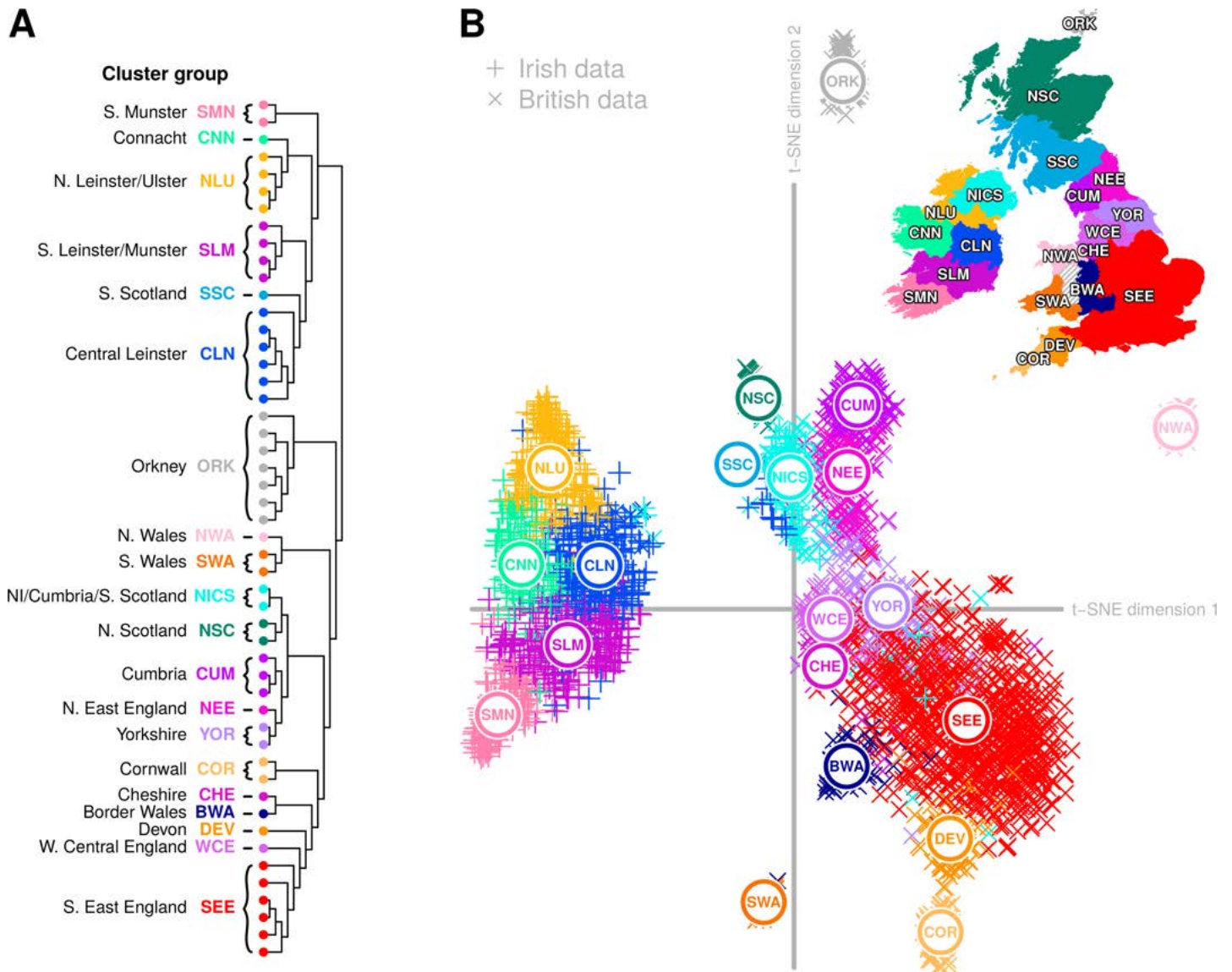
The principal split in the combined Irish and British data defined two genetic islands, both in the fineSTRUCTURE tree and in ChromoPainter PC1 (Fig 2). This distinction between Irish and British genetic data was particularly pronounced when we applied t-distributed stochastic neighbour embedding (t-SNE) [21] to the ChromoPainter coancestry matrix (Fig 3). t-SNE is a nonlinear dimensionality reduction method that attempts to provide an optimal low-dimensional embedding of data by preserving both local and global structure, placing similar points close to each other and dissimilar points far apart. In principle, a two-dimensional t-SNE plot can therefore summarize more of the overall differences between groups than those described by any two principal components, although the relative group sizes, positions and distances on the plot are less straightforward to interpret. Applying t-SNE to the Irish and British coancestry matrix captured the salient structure described by ChromoPainter PCA, and particularly validates that observed in the plot of PC1 vs PC4. This clearly distinguishes the two islands, discerns their north-south and west-east genetic structure and places Orkney and north/south Wales, whose variation is captured in PCs 2 and 3 respectively (Fig 4), as independent entities from the bulk of the British data.

As observed in Fig 1, ChromoPainter PCA in Ireland and Britain (Fig 2) demonstrates eastern homogeneity for each island and relative diversity on the west coast. The southeast England (SEE) cluster group is centred at zero on PC4, representing a group with predominantly Anglo-Saxon-like ancestry (S8 Fig). Clusters representing Celtic populations harbouring less Anglo-Saxon influence separate out above and below SEE on PC4. Notably, northern Irish clusters (NLU), Scottish (NISC, SSC and NSC), Cumbria (CUM) and North Wales (NWA) all separate out at a mutually similar level, representing northern Celtic populations. The southern Celtic populations Cornwall (COR), south Wales (SWA) and south Munster



**Fig 2. Genes mirror geography in the British Isles.** (A) fineSTRUCTURE clustering dendrogram for combined Irish and British data. Data principally split into Irish and British groups before subdividing into a total of 50 distinct clusters, which are combined into cluster groups for clusters that formed clades in the dendrogram, overlapped in principal component space (B) and were sampled from regions that are geographically contiguous. Names and labels follow the geographical provenance for the majority of data within the cluster group. Details for each cluster in the dendrogram are provided in S2 Fig. (B) Principal component analysis (PCA) of haplotypic similarity based on the ChromoPainter coancestry matrix, coloured by cluster group with their median locations labelled. We have chosen to present PC1 versus PC4 here as these components capture new information regarding correlation between haplotypic variation across Britain and Ireland and geography, while PC2 and PC3 (Fig 4) capture previously reported splitting for Orkney and Wales, respectively, from Britain [7]. A map of Ireland and Britain is shown for comparison, coloured by sampling regions for cluster groups, the boundaries of which are defined based on the Nomenclature of Territorial Units for Statistics (NUTS 2010), with some regions combined. Sampling regions are coloured by the cluster group with the majority presence in the sampling region; some sampling regions have significant minority cluster group representations as well, for example the Northern Ireland sampling region (UKN0; NUTS 2010) is majorly explained by the NICS cluster group but also has significant representation from the NLU cluster group. The PCA plot has been rotated clockwise by 5 degrees to highlight its similarity with the geographical map of the Ireland and Britain. NI, Northern Ireland; PC, principal component. Cluster groups that share names with groups from Fig 1 (NLU; SMN; CLN; CNN) have an average of 80% of their samples shared with the initial cluster groups. The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>), note some boundaries have been subsumed or modified to better reflect sampling regions.

<https://doi.org/10.1371/journal.pgen.1007152.g002>

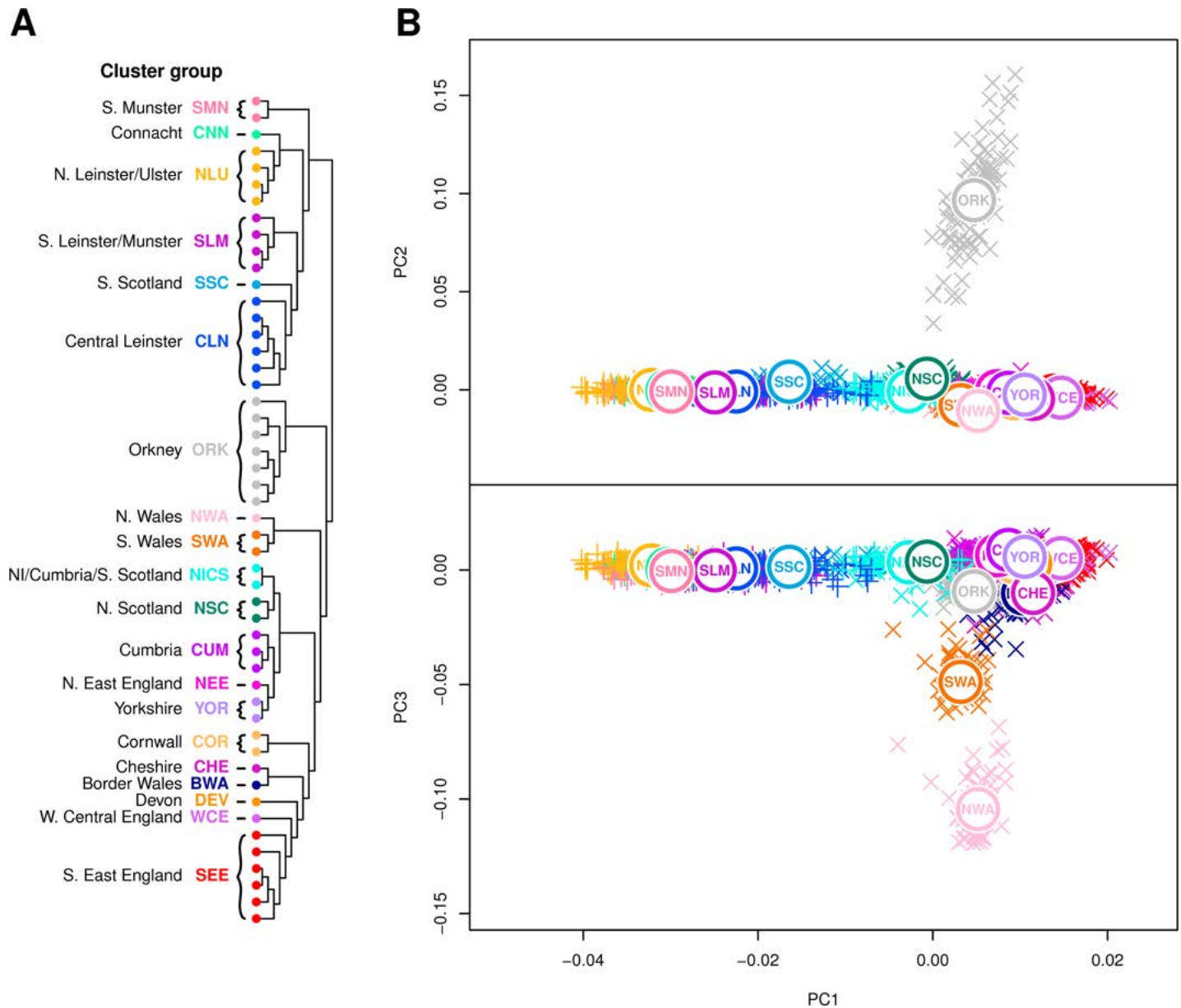


**Fig 3. t-distributed stochastic neighbour embedding (t-SNE) of Irish and British coancestry matrix.** (A) fineSTRUCTURE dendrogram with clusters and cluster groups defined as in Fig 2. (B) Two-dimensional t-SNE embedding of ChromoPainter coancestry matrix, with median locations for cluster groups plotted. As t-SNE is a stochastic method, different runs produce different solutions to the 2-dimensional embedding; shown here is a typical result. t-SNE performed significantly better with the ChromoPainter coancestry matrix than with Hamming distances (identity-by-state) computed over single SNP markers (S9 Fig). The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>), note some boundaries have been subsumed or modified to better reflect sampling regions.

<https://doi.org/10.1371/journal.pgen.1007152.g003>

(SMN) also separate out on similar levels, indicating some shared haplotypic variation between geographically proximate Celtic populations across both Islands. It is notable that after the split of the ancestrally divergent Orkney, successive PCs describe diversity in British populations where “Anglo-saxonization” was repelled [22]. PC3 is dominated by Welsh variation, while PC4 in turn splits North and South Wales significantly, placing south Wales adjacent to Cornwall and north Wales at the other extreme with Cumbria, all enclaves where Brittonic languages persisted.

Scotland is another region of Britain which successfully retained its Celtic language, however in contrast to Welsh and Cornish clusters, the majority of Scottish variation is described

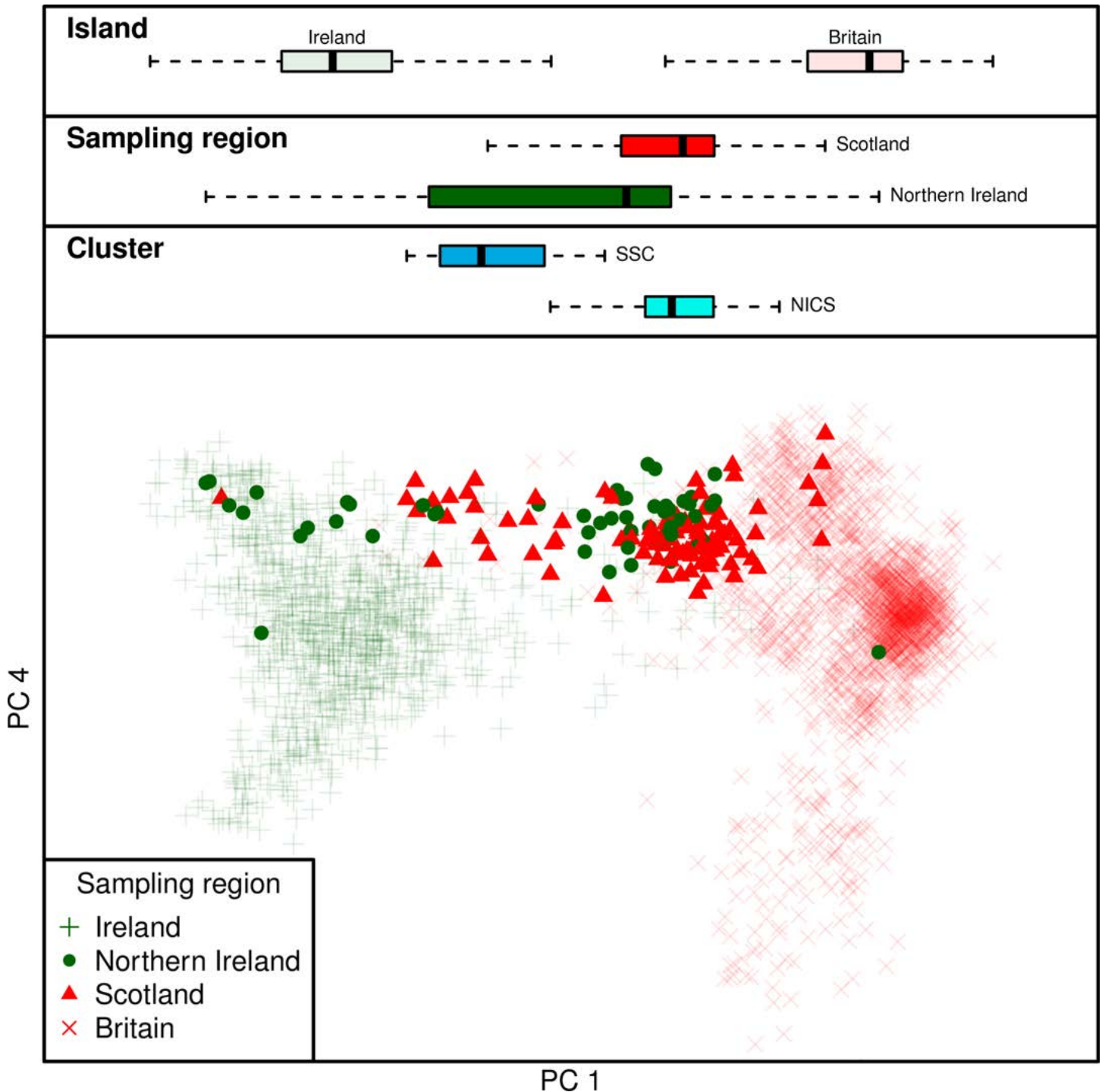


**Fig 4. Principal components 2 and 3 of combined Irish and British coancestry matrix.** (A) fineSTRUCTURE clustering dendrogram for combined Irish and British data, with cluster groups defined as in Fig 2. Immediately following the principal inter-island split, Orkney and Wales branch in sequence, consistent with previous observations. (B) Principal component analysis (PCA) of haplotypic similarity based on the ChromoPainter coancestry matrix, coloured by cluster group with their median locations labelled. PC2 captures an Orkney split, while PC3 captures a Welsh split.

<https://doi.org/10.1371/journal.pgen.1007152.g004>

by ChromoPainter PC1. The three definable Scottish groups do not drive any further components of variation (up to PC7 considered) and fall away from the bulk of British variation on PC1 towards Irish clusters. This is most strikingly observed for the southern Scottish cluster (SSC) which fell amongst Irish branches in the fineSTRUCTURE tree, overlapping with samples from the north of Ireland in PC space (Fig 2 and Fig 5). In an interesting symmetry, many Northern Irish samples clustered strongly with southern Scottish and northern English samples, defining the Northern Irish/Cumbrian/Scottish (NICS) cluster group. More generally, by modelling Irish genomes as a linear mixture of haplotypes from British clusters, we found that

Scottish and northern English samples donated more haplotypes to clusters in the north of Ireland than to the south, reflecting an overall correlation between Scottish/north English contribution and PC1 position in Fig 1 (Linear regression:  $p < 2 \times 10^{-16}$ ,  $r^2 = 0.24$ ).



**Fig 5. Inter-island exchange of haplotypes between the north of Ireland and northern Britain.** The boxplots show the distribution of individuals on principal component (PC) 1 for each island and for specific sampling regions (Scotland/Northern Ireland) and cluster groups (SSC and NICS; see Fig 2). A substantial proportion of Northern Irish individuals fall within the expected range for Scottish individuals in PC space and *vice versa*. This exchange is particularly pronounced for Northern Irish and Scottish individuals that fall within the NICS and SSC cluster groups (Fig 2), respectively.

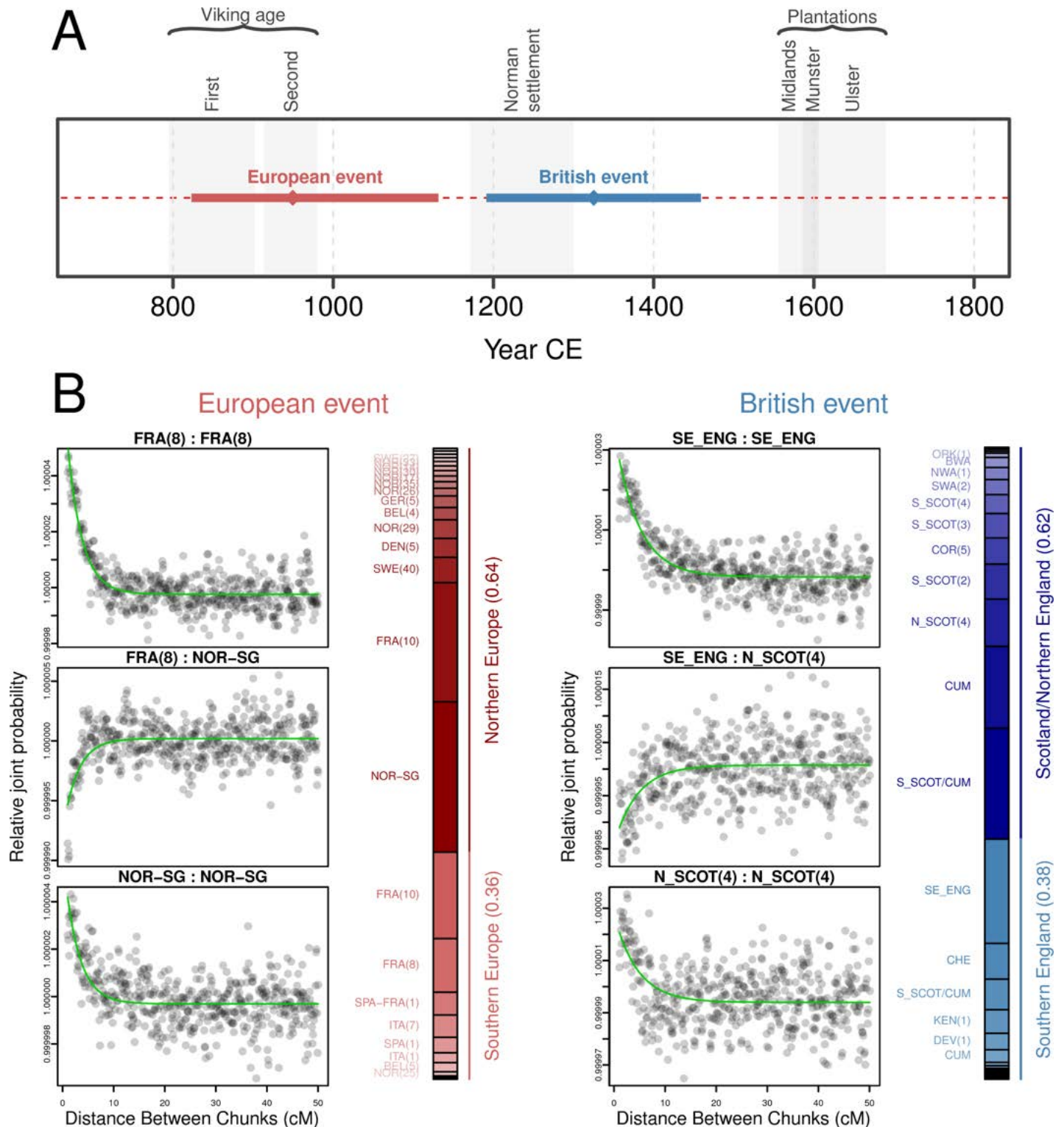
<https://doi.org/10.1371/journal.pgen.1007152.g005>

North to south variation in Ireland and Britain are therefore not independent, reflecting major gene flow between the north of Ireland and Scotland (Fig 5) which resonates with three layers of historical contacts. First, the presence of individuals with strong Irish affinity among the third generation PoBI Scottish sample can be plausibly attributed to major economic migration from Ireland in the 19th and 20th centuries [6]. Second, the large proportion of Northern Irish who retain genomes indistinguishable from those sampled in Scotland accords with the major settlements (including the Ulster Plantation) of mainly Scottish farmers following the 16th Century Elizabethan conquest of Ireland which led to these forming the majority of the Ulster population. Third, the suspected Irish colonisation of Scotland through the *Dál Riata* maritime kingdom, which expanded across Ulster and the west coast of Scotland in the 6th and 7th centuries, linked to the introduction and spread of Gaelic languages [3]. Such a migratory event could work to homogenise older layers of Scottish population structure, in a similar manner as noted on the east coasts of Britain and Ireland. Earlier communications and movements across the Irish Sea are also likely, which at its narrowest point separates Ireland from Scotland by approximately 20 km.

### Genomic footprints of migration into Ireland

To temporally anchor the major historical admixture events into Ireland we used GLOBETROTTER [23] with modern surrogate populations represented by 4,514 Europeans [24] and 1,973 individuals from the PoBI dataset [7], excluding individuals sampled from Northern Ireland. Of all the European populations considered, ancestral influence in Irish genomes was best represented by modern Scandinavians and northern Europeans, with a significant single-date one-source admixture event overlapping the historical period of the Norse-Viking settlements in Ireland ( $p < 0.01$ ; fit quality  $FQ_B > 0.985$ ; Fig 6). This was recapitulated to varying degrees in specific genetically- and geographically-defined groups within Ireland, with the strongest signals in south and central Leinster (the largest recorded Viking settlement in Ireland was *Dubh linn* in present-day Dublin), followed by Connacht and north Leinster/Ulster (S5 Fig; S6 Table). This suggests a contribution of historical Viking settlement to the contemporary Irish genome and contrasts with previous estimates of Viking ancestry in Ireland based on Y chromosome haplotypes, which have been very low [25]. The modern-day paucity of Norse-Viking Y chromosome haplotypes may be a consequence of drift with the small patrilineal effective population size, or could have social origins with Norse males having less influence after their military defeat and demise as an identifiable community in the 11th century, with persistence of the autosomal signal through recombination.

European admixture date estimates in northwest Ulster did not overlap the Viking age but did include the Norman period and the Plantations (S5 Fig). This may indicate limited Viking activity in Ulster, or, that due to the similarity in sources for the Viking and Anglo-Norman invasions and the Plantations, GLOBETROTTER failed to disentangle the earlier events from the later. This is not unexpected given the extent of the Plantations in Ulster [26], the relative timings of the invasions and the degree of Viking involvement in Britain and Europe. Indeed, when considering Britain as an admixing source using PoBI data, GLOBETROTTER date estimates for northwest Ulster overlapped the Plantations, although for other regions in Ireland (and for Ireland considered as a whole) these admixture events were less clearly defined, likely reflecting a history of continuous gene flow between the two islands in the prevailing years (Fig 6; S5 Fig and S7 Table). The all-Ireland point estimate for admixture from Britain spanned the Norman settlement instead of the Plantations, but GLOBETROTTER was unable to adequately resolve the model details for this event (fit quality  $FQ_B < 0.985$ ; Fig 6), indicating that this estimate is not a good reflection of the true timings and extent of admixture from Britain. As noted



**Fig 6. All-Ireland GLOBETROTTER admixture date estimates for European and British surrogate admixing populations.** A summary of the date estimates and 95% confidence intervals for inferred admixture events into Ireland from European and British admixing sources is shown in (A), with ancestry proportion estimates for each historical source population for the two events and example coancestry curves shown in (B). In the coancestry curves *Relative joint probability* estimates the pairwise probability that two haplotype chunks separated by a given genetic distance come from the two modelled source populations respectively (i.e. FRA(8) and NOR-SG); if a single admixture event occurred, these curves are expected to decay exponentially at a rate corresponding to the number of generations since the event. The green fitted line describes this GLOBETROTTER fitted exponential decay for the coancestry curve. If the sources come from the same ancestral group the slope of this curve will be negative (as with FRA(8) vs FRA(8)), while a positive slope indicates that sources come from different admixing groups (as with FRA(8) vs NOR-SG).



The adjacent bar plot shows the inferred genetic composition of the historical admixing sources modelled as a mixture of the sampled modern populations. A European admixture event was estimated by GLOBETROTTER corresponding to the historical record of the Viking age, with major contributions from sources similar to modern Scandinavians and northern Europeans and minor contributions from southern European-like sources. For admixture date estimates from British-like sources the influence of the Norman settlement and the Plantations could not be disentangled, with the point estimate date for admixture falling between these two eras and GLOBETROTTER unable to adequately resolve source and proportion details of admixture event (fit quality  $FQ_B < 0.985$ ). The relative noise of the coancestry curves reflects the uncertainty of the British event. Cluster labels (for the European clustering dendrogram, see S4 Fig; for the PoBI clustering dendrogram, see S3 Fig): FRA(8), France cluster 8; NOR-SG, Norway, with significant minor representations from Sweden and Germany; SE\_ENG, southeast England; N\_SCOT(4) northern Scotland cluster 4.

<https://doi.org/10.1371/journal.pgen.1007152.g006>

in the PoBI study, the overall influence of British admixture in Ireland (and *vice versa*) has involved extensive and constant gene flow before, during and after the major population movements detailed in Fig 6, with particular swells of peopling during the Plantations. The genetic legacies of the populations of Ireland and Britain are therefore extensively intertwined and, unlike admixture from northern Europe, too complex to model with GLOBETROTTER.

## Conclusions

Our results show that population structure is detectable on the island of Ireland and is consistent with a combination of the homogenizing effect of geographically punctuated admixture and diversification among Celtic subpopulations. The inclusion of Irish data with British samples from the PoBI study provides an anchor for Celtic ancestry in the British Isles, filling out the genetic landscape of the islands. It is also clear that historical migrations into Ireland have left a greater genomic footprint than previously anticipated; our consideration of autosomal data escapes the constraints of patrilineal genetics and has allowed us to detect a much greater Viking influence than previously estimated with Y chromosome data. Although the genetic imprint of the British Plantations is much harder to delineate, the inter-island exchange and clustering observed between present-day individuals from Northern Ireland and Scotland signals the enduring impact of these historical movements of people.

Unlike the PoBI study, Irish data were not specifically selected for longstanding pure ancestry in each geographic region (for example, having four grandparents in a location), but instead represent a repurposed medical dataset. Our data are therefore more representative of those that are typically used in population-based genome-wide surveys for trait-associated genetic variation; as these studies survey increasingly rare genetic variants in larger populations, the geospatial segregation of rare haplotypes and variants will become increasingly important, especially when environmental effects and interactions play a role [27]. Our observation that these haplotypes are intricately tied to geography in Ireland and Britain highlights the importance of considering fine-grained population structure in future studies.

## Methods

### Ethics statement

All Irish subjects provided written informed consent to participate in genetic research and the study was approved by the Beaumont Hospital Research Ethics Committee in Dublin, Ireland under approval number 05/49 following guidelines laid out at [www.beaumontethics.ie](http://www.beaumontethics.ie).

### Data and quality control

Our study included three datasets of genotype data: a population-based Irish ALS case-control dataset ( $n = 991$ ) incorporating existing [28] and newly-genotyped samples, the People of the British Isles dataset (EGA accession ID EGAD00010000632;  $n = 2,020$ ) [7] and a pan-European dataset derived from a genome-wide association study (GWAS) for multiple sclerosis

(MS; EGA accession ID EGAD00000000120;  $n = 4,514$ ) [24] (S1 Text: Populations). All Irish subjects provided written informed consent to participate in genetic research and the study was approved by the Beaumont Hospital Research Ethics Committee in Dublin, Ireland. We applied quality control to each dataset using PLINK 1.9 [29] and merged data as detailed in Supplementary Methods (S1 Text: Quality Control). Briefly, we excluded both infrequent and high-missingness SNPs; individuals with high missingness, excessive heterozygosity or cryptic relationships to other individuals in the data; and finally individuals who had been removed during QC carried out in the source papers.

As the European dataset included patients and controls from a GWAS for MS, we additionally removed SNPs in a 15 Mb region surrounding the strongly associated HLA locus on chromosome 6 (GRCh37 position chr6:22,915,594–37,945,593), as is consistent with previous studies using the data [7,30]. This was to avoid haplotypic bias arising from this association.

The final post-QC Irish ( $n = 991$ ), British ( $n = 2,020$ ) and European datasets ( $n = 4,514$ ) contained 407,750 SNPs, 521,883 SNPs and 363,396 SNPs at zero missingness, respectively. The final merge of British and Irish data ( $n = 3,008$ ) and European and Irish data ( $n = 5,506$ ) contained 214,632 SNPs and 166,139 SNPs respectively at zero missingness. Further details regarding samples and QC per dataset are described in Supplementary Methods (S1 Text: Populations and S1 Text: QC)

Geographic information was available for 544 of the 991 Irish samples in the form of home address. To preserve anonymity this was jittered in all maps containing patients (Fig 1 and S5 Fig). For all British and some Northern Irish data, sample location was supplied by the authors of PoBI [7] as membership of 35 sampling regions. Finally, for European data sampling country was available [24]. Full details of treatment of samples for mapping are available in Supplementary methods (S1 Text: Mapping.)

## Phasing

We phased autosomal genotypes in each dataset and merged dataset with SHAPEIT V2 [31] using the 1000 Genomes (Phase 3) as a reference panel [32]. A pre-phasing step was carried out (—check) to remove any SNPs which did not correctly align to the 1000 genomes reference panel. Samples were then split by chromosome and phased together using default settings and the GRCh37 build genetic map to estimate linkage disequilibrium.

## fineSTRUCTURE analysis

To detect population structure we performed ChromoPainter/fineSTRUCTURE analysis [14] on each of the population datasets (Irish, British and European) individually, and then separately on a merge of the Irish and British datasets. In brief, we used ChromoPainter to paint each individual using all other individuals (—a 0 0) using default settings with the exception that the number of “chunks” per region value was set to 50 (—k 50) for all analyses including Irish and British individuals to account for the longer haplotypes observed in these datasets, in keeping with previous studies [7,30]. The fineSTRUCTURE algorithm was then run on the resulting coancestry matrix to determine genetic clusters based on patterns of haplotype sharing. Further details are included in the Supplementary Methods (S1 Text: fineSTRUCTURE analysis).

## Cluster robustness

We assessed the robustness of Irish clusters by calculating total variation distance (TVD) as described in the PoBI study [7]. This metric compares the “copying vectors” of pair of clusters. Here we define the copying vector for a given cluster  $A$  as a vector of the average lengths of DNA donated by each cluster to individuals within cluster  $A$  under the ChromoPainter model.

Hence the magnitude of differences between copying vectors of two clusters reflects the distances between those clusters in terms of their haplotypic sharing with other clusters. TVD can therefore be used to determine whether fineSTRUCTURE clusters detect significant differences in haplotype sharing, and hence ancestry.

We tested whether the observed clustering performed better than chance by permuting (1,000 times) the individuals in each of our cluster pairings into clusters of the same size, and calculating the number of permutations that exceeded our original TVD score. If 1,000 unique permutations were not possible, the maximum number of unique permutations was used instead. P-values were calculated based on the number of permutations greater than or equal to the original TVD statistic. All p-values for Irish clusters were less than or equal to 0.001 indicating robust clustering (S1 Table and S2 Table). We also applied these methods to our Irish cluster groups (Fig 1) and observed that these are statistically distinct (S3 Table and S4 Table).

To provide an additional measure of population differentiation between “cluster groups” we calculated mean  $F_{ST}$  between groups using PLINK 1.9 [29] which is reported in S5 Table.

### Estimating admixture dates

We used the GLOBETROTTER method [23] to infer and date admixture events from Europe and Britain into Ireland separately. GLOBETROTTER uses output from ChromoPainter to estimate the pairwise likelihood of being painted by any two surrogate populations at a variety of genetic distances to generate coancestry curves. Assuming a single admixture event, these curves are expected to follow an exponential decay rate equal to the time in generations since admixture occurred [23]. As the true admixing sources are modelled as a linear mixture of surrogate sources rather than individual sources this method has the advantage of not requiring exactly sampled source populations.

For our analysis we ran GLOBETROTTER with default settings twice to detect simple admixture into the island of Ireland as a whole, as well as into individual genetic clusters from the Republic of Ireland (S5 Fig). European clusters (S4 Fig) and British clusters (S3 Fig) were used as surrogate populations to represent the admixing sources in two independent analyses. Target and donor clusters for this analysis were defined using the fineSTRUCTURE maximum concordance tree method described in PoBI [7] to ensure homogeneity (Supplementary methods S1 Text: fineSTRUCTURE analysis); hence, the Irish target clusters that were used differ slightly from those in Fig 1. Briefly, for each surrogate population separately (Europe and Britain) we applied ChromoPainter v2 to paint Ireland and the surrogate population using the surrogate population as donors and generated a copying matrix (chunk lengths) for all individuals, and also 10 painting samples for each Irish individual as recommended. GLOBETROTTER was then run for 5 mixing iterations twice, first using the null.ind:1 setting to test for any evidence of admixture and then null.ind:0 setting to infer dates and sources. We ran 100 bootstraps for admixture date and calculated the probability of a null model of no admixture as the proportion of nonsensical inferred dates (<1 or >400 generations) produced by the null.ind:1 model, as in the GLOBETROTTER study [23]. Confidence intervals for the date were calculated from the bootstraps for the standard model (null.ind: 0) using the empirical bootstrap method. (See S1 Text: Globetrotter analysis of Admixture Dates for further details). A generation time of 28 years was assumed as in previous studies of this nature [7,23] for conversion of all date estimates from generations to years.

### Ancestry proportion estimation

We assessed the ancestral make up of Ireland in terms of Europe and Britain for each Republic of Ireland cluster (see Estimating admixture dates) to explore variation in ancestry across

Ireland. To do so we modelled each cluster's average genome as a linear mixture of the European and British donor populations using the method described in the PoBI study [7] and implemented in GLOBETROTTER (num.mixing.iterations: 0). This approach uses the ChromoPainter chunk length output to estimate the proportion of DNA which most closely coalesces with each individual from the donor populations, correcting for noise caused by similarities between donor populations whose splits may have occurred after the coalescence event. This is achieved through a multiple linear regression of the form  $Y_p = B_1X_1 + B_2X_2 + \dots + B_gX_g$ , where  $Y_p$  is a vector of the averaged length (cM) of DNA that individuals across cluster P copy from each donor group, normalised to sum to 1 across all donor groups, and  $X_g$  is the vector describing the average proportion of DNA that individuals in donor group g copy from other donor groups including their own. The coefficients of this equation  $B_1 \dots B_g$  are thus interpreted as the "cleaned" proportions of the genome ancestral to each donor group. The equation is solved using a non-negative-least squares function such that  $B_g \geq 0$  and the sum of proportions across groups equals 1.

To assess uncertainty of these ancestry proportion estimates we again follow PoBI [7] and resample from the ChromoPainter chunk length output to generate  $N_p$  pseudo individuals for each cluster P. We achieve this by randomly sampling each of the autosomal chromosome pairs 1–22 with replacement  $N_p$  times from the pool of all autosomal chromosomes pairs 1–22 across all individuals within that cluster, and then randomly summing sets of 22 of these chromosome pairs to generate each pseudo individual. We then use these  $N_p$  pseudo individuals as a bootstrap for  $Y_p$  above and solve for  $B_g$ . We resampled 1,000 times per cluster and used the inner 95% quantiles of this sampling distribution to estimate confidence intervals for the sample.

For comparison we implemented an alternative delete one chromosome jack-knife approach as in Montinaro et al. [33], and estimated the s.e. as in ref. [34] (S6 Fig and S7 Fig).

We also used this linear regression model to determine per-individual ancestry proportion estimates from different British clusters across Ireland, treating each individual as a cluster to enable us to assess whether gene flow from northern Britain had a gradient across Ireland.

## ADMIXTURE

To estimate the proportion of British admixture into Irish clusters, ADMIXTURE [18] was run on the combined PoBI and Irish datasets, alongside eighteen ancient individuals from the Iron Age, Roman and Anglo-Saxon periods of northeast and southeast England [16,17]. Pseudo-haploid genotypes were generated for the ancient genomes at the relevant variant sites, as is standard for low coverage data, and subsequently merged with the modern diploid dataset. Data were then pruned for linkage disequilibrium between SNPs using PLINK 1.9 ( $r^2 > 0.25$  in a sliding window of 1000 SNPs advancing 50 SNPs each time) resulting in 86,481 remaining SNPs. No missingness was allowed for modern individuals, with a range of 33,643–85,553 sites used for ancient samples. Following ADMIXTURE estimation, cross-validation error was calculated using the—cv flag for 5 iterations to determine the K value for which the model has the best predictive accuracy ( $K = 2$ ). Additionally 200 bootstraps of the data were run to estimate the standard error of the parameters using the—B flag. This British admixture component was regressed against PC2 of the Irish ChromoPainter coancestry matrix to determine the role of British ancestry in the differentiation of PC2 in Ireland. We also performed analysis of variance (ANOVA) on British admixture component per cluster group to identify if cluster by cluster differences existed.

## PCA and t-SNE

ChromoPainter coancestry matrices were projected in lower-dimensional space using principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE)

[21]. PCA was run using the default approach provided as part of the fineSTRUCTURE R tools [14] (<http://www.paintmychromosomes.com>). The R package Rtsne (<https://github.com/jkrijthe/Rtsne>) was used to construct a 2-dimensional embedding of the ChromoPainter coancestry matrix over 5,000 iterations using a perplexity of 30, a learning rate of 200 and an initial PCA calculated over 100 dimensions. Several t-SNE runs were performed to assess concordance between embedding solutions.

### Other statistical analyses

All linear regressions and ANOVA tests were carried out in base statistics package in R version 3.2.3 [35].

### Supporting information

**S1 Fig. Irish fineSTRUCTURE tree cluster details.** The fineSTRUCTURE tree presented in Fig 1 for Irish clusters with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups are identical to those defined in Fig 1.

(PDF)

**S2 Fig. PoBI/Irish fineSTRUCTURE tree cluster details.** The fineSTRUCTURE tree presented in Fig 2 for British and Irish clusters with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups are identical to those defined in Fig 2.

(PDF)

**S3 Fig. PoBI maximum concordance fineSTRUCTURE tree cluster details.** The fineSTRUCTURE maximum concordance tree for British clusters used in GLOBETROTTER analysis with detailed breakdown of individual clusters. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups describe clusters which are neighbouring in the tree and geographically adjacent.

(PDF)

**S4 Fig. European maximum concordance fineSTRUCTURE tree cluster details.** The fineSTRUCTURE maximum concordance tree for European clusters used in GLOBETROTTER analysis with detailed breakdown of individual clusters. Additional individuals from WTCCC exclusion list have been removed post fineSTRUCTURE clustering but prior to GLOBETROTTER analysis and the tree updated to reflect this. The individual labels for the clusters describe the geographic location of the majority of samples and the numbers of individuals within those clusters are provided in brackets. Cluster groups describe clusters which are neighbouring in the tree and geographically adjacent.

(PDF)

**S5 Fig. GLOBETROTTER breakdown for clusters in the Republic of Ireland.** A summary of the date estimates and 95% confidence intervals for inferred admixture events into Irish clusters from European (red) and British (blue) admixing sources is shown in (A). Faded lines highlight clusters in which there was no significant evidence of admixture ( $P > 0.01$ ). (B) Summarises the fineSTRUCTURE maximum concordance tree cluster assignment for the 991 Irish samples used as target populations in GLOBETROTTER estimates in (A). We present the

fineSTRUCTURE clustering dendrogram, a PCA of the coancestry matrix coloured by cluster group and a map of Ireland showing the sampling location for a subset of 544 individuals for which locational information was available, coloured by cluster group. Points have been randomly jittered within a radius of 5 km to preserve anonymity. The map and administrative boundaries were produced using data from the database of Global Administrative Areas (GADM; <https://gadm.org>).

(PDF)

**S6 Fig. British ancestry profile in Irish clusters.** Bar charts displaying the GLOBETROTTER estimated British ancestry profile for Republic of Ireland clusters (Defined in S5 Fig; Only clusters with 35+ samples displayed) from British clusters inferred from 2,017 individuals using fineSTRUCTURE (Defined in S3 Fig). Individuals from Northern Ireland were excluded to prevent masking of ancestry leaving 1973 individuals. Only donors that make at least a 2.5% contribution to at least one Irish cluster are displayed with the remaining proportions subsumed into the “other” category. Error bars represent the bootstrapping procedure with 10000 resamples (Black) and a jack-knife approach using 22 resamples (Red). Label abbreviations: S\_SCOT, south Scotland; SE\_ENG, southeast England; CHE, Cheshire; KEN, Kent; BWA, border Wales; DEV, Devon; COR, Cornwall; N\_SCOT north Scotland; SWA, south Wales; NWA, north Wales.

(PDF)

**S7 Fig. European ancestry profile in Irish clusters.** Bar charts displaying the GLOBETROTTER estimated European ancestry profile for republic of Ireland clusters (Defined in S5 Fig; Only samples with 35+ samples displayed) from European clusters inferred from 4,514 individuals using fineSTRUCTURE (Defined in S4 Fig). Only donors that make at least a 2.5% contribution to at least one Irish cluster are displayed with the remaining proportions subsumed into the “other” category. Error bars represent the bootstrapping procedure with 10000 resamples (Black) and a jack-knife approach using 22 resamples (Red). Label abbreviations: NOR-SG, Norway, with significant minor representations from Sweden and Germany; FRA, France; NOR, Norway; BEL, Belgium.

(PDF)

**S8 Fig. ADMIXTURE analysis for PoBI/Irish cluster groups with ancient British samples.** ADMIXTURE component ( $k = 2$ ) for each cluster group in the PoBI/Irish fineSTRUCTURE tree (S2 Fig) and 18 Ancient British Samples from the Iron age (IA;  $n = 4$ ), Anglo-Saxon (AS;  $n = 8$ ) and Roman (RM;  $n = 6$ ) periods. Admixture proportions are averaged across each cluster group (left) for brevity of display, while individual proportions are plotted for ancient samples. The Anglo-Saxon individuals are best described by the red component. This component is high in British cluster groups from areas affected by the Anglo-Saxon invasion such as the large SEE cluster, while relatively low in Celtic populations such as Ireland, Scotland and Wales.

(PDF)

**S9 Fig. t-distributed stochastic neighbour embedding (t-SNE) of Irish and British genotypes.** A t-SNE solution for 2-dimensional embedding is displayed for Irish and British genotype data using Hamming distances (identity-by-state). As t-SNE is a stochastic method, different runs produce different solutions to the 2-dimensional embedding; shown here is a typical result. Clusters and cluster groups are defined as in Fig 2, with median locations for cluster groups plotted. t-SNE performed significantly worse with the Hamming distances (identity-by-state) computed over single SNP markers than with the fineSTRUCTURE coancestry matrix (Fig 3).

(PDF)

**S10 Fig. Comparison of Linked vs Unlinked fineSTRUCTURE in Ireland at 166,139 SNPs.** Displays ChromoPainter PC1 and PC2 alongside a fineSTRUCTURE Maximum Concordance clustering dendrogram for A.) Linked and B.) Unlinked analysis for 991 Irish individuals at the 166,139 SNP positions used for our European GLOBETROTTER run. Trees and PCA are coloured at a  $k = 11$  split for ease of visualisation. Considerably more structure is apparent in the PCA of the Linked analysis indicating that linkage information defines meaningful haplotypes even at this resolution. We report “Confidence of ind. assignment” for each method. This metric is the confidence of individual assignment to their final cluster based on their assignment across all MCMC samples defined in PoBI [7]. This was on average 84.8% (95% CI: 83.9–85.7%) for the Linked analysis, while in the Unlinked analysis this was only 8.06% (95% CI: 8.03–8.09%), suggesting that the final clustering assignment in the unlinked mode is extremely uncertain and variable.  
(PDF)

**S11 Fig. Bootstraps for British ADMIXTURE component estimates.** Standard error calculated using 200 bootstrap resamples for each point in linear regression in Fig 1 (E.) are plotted using error bars to show variability in ADMIXTURE point estimates.  
(PDF)

**S1 Table. TVD table for Irish clusters.** Total Variation Distance (TVD) matrix between Irish clusters described in Fig 1 and S1 Fig demonstrating the degree of differentiation between clusters.  
(ODS)

**S2 Table. TVD p-values for Irish clusters.** P-values that individuals are assigned randomly to pairs of clusters based on permutation testing using TVD statistic from S1 Table.  
(ODS)

**S3 Table. TVD table for Irish cluster groups.** Total Variation Distance (TVD) matrix between Irish cluster groups described in Fig 1 and S1 Fig demonstrating the degree of differentiation between Cluster Groups.  
(ODS)

**S4 Table. TVD p-values for Irish cluster groups.** P-values that individuals are assigned randomly to pairs of cluster groups based on permutation testing using TVD statistic from S3 Table.  
(ODS)

**S5 Table. FST table for Irish cluster groups.** Mean FST statistic between Irish cluster groups calculated using PLINK 1.9.  
(ODS)

**S6 Table. European GLOBETROTTER table.** Table describing the model fit of GLOBETROTTER for admixture events into Irish clusters from Europe in Fig 6 and S5 Fig.  
(ODS)

**S7 Table. British GLOBETROTTER table.** Table describing the model fit of GLOBETROTTER for admixture events into Irish clusters from Britain in Fig 6 and S5 Fig.  
(ODS)

**S1 Text. Supplementary methods.**  
(PDF)

## Acknowledgments

We gratefully acknowledge Dr T. Rowan McLaughlin for providing helpful insights and critical review of the manuscript. A subset of the Irish data was generated as part of Project MinE ([www.projectmine.com](http://www.projectmine.com)) and we acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support. Boundary data for Britain and Ireland were downloaded from the GADM database ([www.gadm.org](http://www.gadm.org)).

## Author Contributions

**Conceptualization:** Ross P. Byrne, Rui Martiniano, Lara M. Cassidy, Matthew Carrigan, Daniel G. Bradley, Russell L. McLaughlin.

**Data curation:** Ross P. Byrne, Russell L. McLaughlin.

**Formal analysis:** Ross P. Byrne, Rui Martiniano, Lara M. Cassidy, Russell L. McLaughlin.

**Funding acquisition:** Orla Hardiman, Russell L. McLaughlin.

**Investigation:** Ross P. Byrne, Lara M. Cassidy.

**Methodology:** Rui Martiniano, Matthew Carrigan, Garrett Hellenthal, Russell L. McLaughlin.

**Project administration:** Ross P. Byrne, Daniel G. Bradley, Russell L. McLaughlin.

**Resources:** Garrett Hellenthal, Orla Hardiman, Daniel G. Bradley, Russell L. McLaughlin.

**Software:** Ross P. Byrne, Rui Martiniano, Lara M. Cassidy, Garrett Hellenthal, Russell L. McLaughlin.

**Supervision:** Rui Martiniano, Orla Hardiman, Daniel G. Bradley, Russell L. McLaughlin.

**Validation:** Ross P. Byrne.

**Visualization:** Ross P. Byrne, Russell L. McLaughlin.

**Writing – original draft:** Ross P. Byrne, Lara M. Cassidy, Daniel G. Bradley, Russell L. McLaughlin.

**Writing – review & editing:** Ross P. Byrne, Rui Martiniano, Lara M. Cassidy, Matthew Carrigan, Garrett Hellenthal, Orla Hardiman, Daniel G. Bradley, Russell L. McLaughlin.

## References





1. Bayliss A, Woodman P. A New Bayesian Chronology for Mesolithic Occupation at Mount Sandel, Northern Ireland. *Proceedings of the Prehistoric Society*. 2009; 75: 101–123.
2. Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, et al. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc Natl Acad Sci U S A*. 2016; 113: 368–373. <https://doi.org/10.1073/pnas.1518445113> PMID: 26712024
3. Dillon M, Chadwick NK. *The Celtic Realms*. Weidenfeld & Nicolson; 1967.
4. Jones C. *The Edinburgh History of the Scots Language*. Edinburgh University Press; 1997.
5. Koch JT. *Celtic culture: a historical encyclopedia*. ABC-CLIO; 2006.
6. Duffy S. *The concise history of Ireland*. Gill & Macmillan; 2000.
7. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. *Nature*. 2015; 519: 309–314. <https://doi.org/10.1038/nature14230> PMID: 25788095
8. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration to the steppe was a source for Indo-European languages in Europe. *Nature*. 2015; 522: 207–211. <https://doi.org/10.1038/nature14317> PMID: 25731166



9. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*. 2016; 536: 419–424. <https://doi.org/10.1038/nature19310> PMID: 27459054
10. O'Dushlaine CT, Morris D, Moskvina V, Kirov G, International Schizophrenia Consortium, Gill M, et al. Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur J Hum Genet*. 2010; 18: 1248–1254. <https://doi.org/10.1038/ejhg.2010.87> PMID: 20571510
11. McLaughlin RL, Kenna KP, Vajda A, Heverin M, Byrne S, Donaghy CG, et al. Homozygosity mapping in an Irish ALS case-control cohort describes local demographic phenomena and points towards potential recessive risk loci. *Genomics*. 2015; 105: 237–241. <https://doi.org/10.1016/j.ygeno.2015.01.002> PMID: 25620680
12. Moore LT, McEvoy B, Cape E, Simms K, Bradley DG. A Y-chromosome signature of hegemony in Gaelic Ireland. *Am J Hum Genet*. 2006; 78: 334–338. <https://doi.org/10.1086/500055> PMID: 16358217
13. McEvoy B, Simms K, Bradley DG. Genetic investigation of the patrilineal kinship structure of early medieval Ireland. *Am J Phys Anthropol*. Wiley Subscription Services, Inc., A Wiley Company; 2008; 136: 415–422. <https://doi.org/10.1002/ajpa.20823> PMID: 18350585
14. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet*. 2012; 8: e1002453. <https://doi.org/10.1371/journal.pgen.1002453> PMID: 22291602
15. Duffy S. *Atlas of Irish History*. Gill & MacMillan; 2012.
16. Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G, et al. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat Commun*. 2016; 7: 10326. <https://doi.org/10.1038/ncomms10326> PMID: 26783717
17. Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, et al. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*. 2016; 7: 10408. <https://doi.org/10.1038/ncomms10408> PMID: 26783965
18. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009; 19: 1655–1664. <https://doi.org/10.1101/gr.094052.109> PMID: 19648217
19. Commission E-E, Others. *Regions in the European Union. Nomenclature of territorial units for statistics. NUTS 2010/EU-27*. Luxembourg: Publications Office of the European Union; 2011.
20. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008; 456: 98–101. <https://doi.org/10.1038/nature07331> PMID: 18758442
21. van der Maaten L. Learning a Parametric Embedding by Preserving Local Structure. *Proceedings of Machine Learning Research*. 2009; 5: 384–391.
22. Deacon B. *A Concise History of Cornwall*. University of Wales Press-Hi; 2007.
23. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. *Science*. 2014; 343: 747–751. <https://doi.org/10.1126/science.1243518> PMID: 24531965
24. International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium 2, Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*. 2011; 476: 214–219. <https://doi.org/10.1038/nature10251> PMID: 21833088
25. McEvoy B, Brady C, Moore LT, Bradley DG. The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. *Eur J Hum Genet*. 2006; 14: 1288–1294. <https://doi.org/10.1038/sj.ejhg.5201709> PMID: 16957681
26. McLaughlin R, Lyttleton J. *An Archaeology of Northern Ireland, 1600–1650*. Department for Communities; 2017.
27. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012; 44: 243–246. <https://doi.org/10.1038/ng.1074> PMID: 22306651
28. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat Genet*. 2016; 48: 1043–1048. <https://doi.org/10.1038/ng.3622> PMID: 27455348
29. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4: 7. <https://doi.org/10.1186/s13742-015-0047-8> PMID: 25722852
30. Gilbert E, Carmi S, Ennis S, Wilson JF, Cavalleri GL. Genomic insights into the population structure and history of the Irish Travellers. *Sci Rep*. 2017; 7: 42187. <https://doi.org/10.1038/srep42187> PMID: 28181990
31. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011; 9: 179–181. <https://doi.org/10.1038/nmeth.1785> PMID: 22138821

32. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature*. 2015; 526: 68–74. <https://doi.org/10.1038/nature15393> PMID: [26432245](https://pubmed.ncbi.nlm.nih.gov/26432245/)
33. Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling the hidden ancestry of American admixed populations. *Nat Commun*. 2015; 6: 6596. <https://doi.org/10.1038/ncomms7596> PMID: [25803618](https://pubmed.ncbi.nlm.nih.gov/25803618/)
34. Frank M T, Meijer E, Van Der Leeden R. Delete-m Jackknife for Unequal m. *Stat Comput*. Kluwer Academic Publishers; 1999; 9: 3–8.
35. CoreTeam R. R: A Language and Environment for Statistical Computing. Vienna,Austria: R Foundation for Statistical Computing; 2015. 2015.

# Dutch population structure across space, time and GWAS design

Ross P. Byrne <sup>1✉</sup>, Wouter van Rheenen <sup>2</sup>, Project MinE ALS GWAS Consortium\*, Leonard H. van den Berg<sup>2</sup>, Jan H. Veldink <sup>2</sup> & Russell L. McLaughlin <sup>1✉</sup>

Previous genetic studies have identified local population structure within the Netherlands; however their resolution is limited by use of unlinked markers and absence of external reference data. Here we apply advanced haplotype sharing methods (ChromoPainter/fineSTRUCTURE) to study fine-grained population genetic structure and demographic change across the Netherlands using genome-wide single nucleotide polymorphism data (1,626 individuals) with associated geography (1,422 individuals). We identify 40 haplotypic clusters exhibiting strong north/south variation and fine-scale differentiation within provinces. Clustering is tied to country-wide ancestry gradients from neighbouring lands and to locally restricted gene flow across major Dutch rivers. North-south structure is temporally stable, with west-east differentiation more transient, potentially influenced by migrations during the middle ages. Despite superexponential population growth, regional demographic estimates reveal population crashes contemporaneous with the Black Death. Within Dutch and international data, GWAS incorporating fine-grained haplotypic covariates are less confounded than standard methods.

<sup>1</sup>Smurfit Institute of Genetics, Trinity College Dublin, Dublin D02 DK07, Republic of Ireland. <sup>2</sup>Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands. \*A full list of members and their affiliations appears in the Supplementary Information. ✉email: [rbyrne5@tcd.ie](mailto:rbyrne5@tcd.ie); [mclaugr@tcd.ie](mailto:mclaugr@tcd.ie)

The Netherlands is a densely populated country on the northwestern edge of the European continent, bounded by Germany, Belgium and the North Sea. The country is divided into twelve provinces and has a complex demographic history, with occupation by several Germanic peoples since the collapse of the Roman Empire, including the Frisians, the Low Saxons and the Franks. Over 17 million individuals now inhabit this relatively small region (41,500 km<sup>2</sup>), making it one of the most densely populated countries in Europe. Despite its small geographical size, previous genetic studies of the people of the Netherlands have demonstrated coarse population structure that correlates with its geography, as well as apparent heterogeneity in effective population sizes across provinces<sup>1,2</sup>. These observations suggest that the demographic past of the Dutch population has left residual signatures in its present regional genetic structure; however, this has not been fully explained in the context of neighbouring populations and thus far the use of unlinked genetic markers have limited the resolution at which this structure can be described. This resolution limit also confines the extent to which the confounding effects of population structure can be controlled in genomic studies of health and disease such as genome-wide association studies (GWAS). As these studies continue to seek ever-rarer genetic variation with ever-increasing cohort sizes, intricate understanding and fine control of population structure is becoming increasingly relevant, but increasingly challenging<sup>3</sup>.

Recent studies have showcased the power of leveraging shared haplotypes to uncover and characterise previously unrecognised fine-grained genetic structure within populations, yielding novel insights into the demographic composition and history of Britain and Ireland<sup>4–7</sup>, Finland<sup>8</sup>, Japan<sup>9</sup>, Italy<sup>10</sup>, France<sup>11</sup> and Spain<sup>12</sup>. Haplotype sharing has also revealed genetic affinities between populations<sup>13</sup>, enabling inference of historical admixture events using modern populations as proxies for ancestral admixing sources<sup>14</sup>. Furthermore, geographic information can be integrated to model genetic similarity as a function of spatial distance<sup>15</sup> to infer demographic mobility within or between populations; one approach uses the Wishart distribution to estimate and map a surface of effective migration rates based on deviations from a pure isolation by distance model<sup>16</sup>, allowing migrational cold spots to be inferred which may derive from geographical boundaries such as rivers and mountains. Almost half of the area of the Netherlands is reclaimed from the sea and its contemporary land surface is densely subdivided by human-made waterways and naturally-occurring rivers, including the Rhine (Dutch: *Rijn*), Meuse (*Maas*), Waal and IJssel. These rivers have been speculatively linked to genetic differentiation between northern and southern Dutch subpopulations in previous work<sup>1</sup>; however the explicit relationship between Dutch genetic diversity and movement of people within the Netherlands has not been directly modelled.

The Dutch have previously received special interest as a model population<sup>1,2</sup> and form a major component of substantial ongoing efforts to better understand human health, disease, demography and evolution. For example, at the time of writing, over 10% of all studies listed in the NHGRI-EBI genome-wide association study (GWAS) catalogue<sup>17</sup> include the Netherlands in their “Country of recruitment” metadata. As well as offering insights into demography and human history, refined population genetic studies are important to identify and adequately control confounding effects in genomic studies of health and disease, especially if spatially structured environmental factors contribute substantially to variance in phenotype, which in particular impacts rare variants<sup>18</sup>.

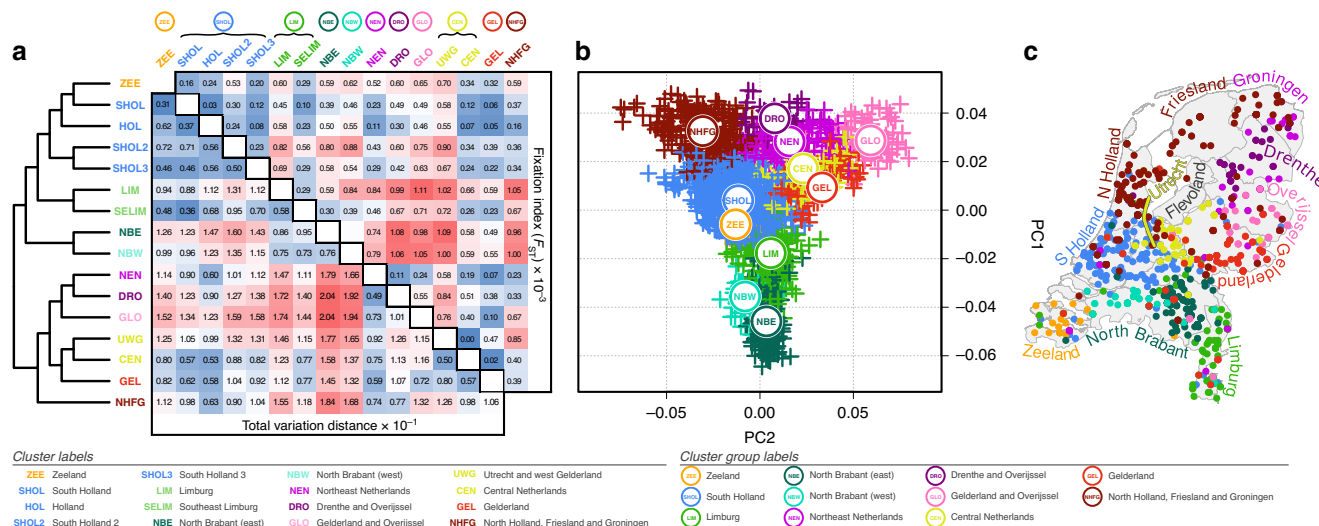
In this study, we harness shared haplotypes to examine the fine-grained genetic structure and demography of the Netherlands. We show that Dutch population structure is more granular

than previously recognised, and is ancient and persistent over time. The strength and stability of the observed structure appears to be tied to the relationship of the Netherlands to neighbouring lands and to its own internal geography, and has likely been shaped over history by migration, but preserved in recent generations by enduring sedentism of genetically similar individuals within regions. We observe genetic evidence of regional population crashes during the Black Death and a countrywide population surge in the 17th century. Finally, we show that the complex genetic structure observed demonstrably confounds GWAS; however, through analysis of the Netherlands and more extensive international data<sup>19</sup>, we demonstrate that using shared haplotypes as GWAS covariates significantly reduces this confounding over standard single-marker methods.

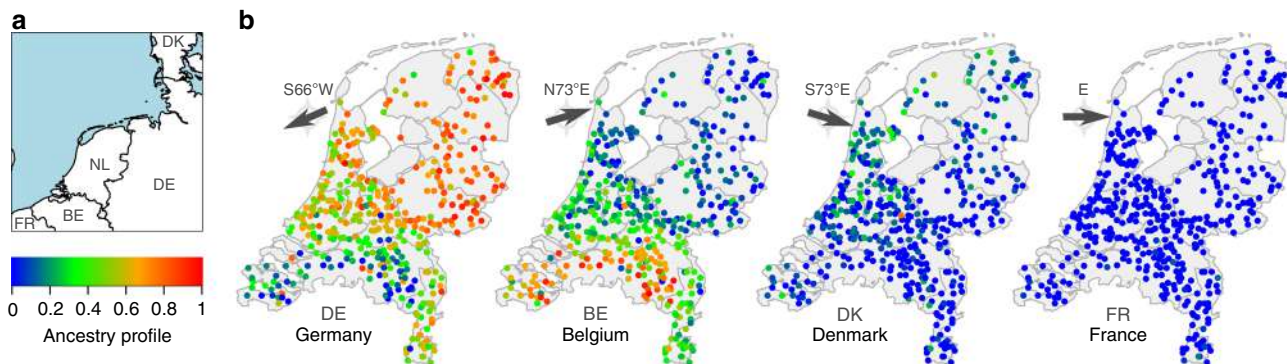
## Results

**The genetic structure of the Dutch population.** We mapped the haplotypic coancestry profiles of 1626 Dutch individuals using ChromoPainter<sup>20</sup> and clustered the resulting matrix using fineSTRUCTURE<sup>20</sup>, identifying 40 genetic clusters at the highest level of the hierarchical tree which segregated with geographical provenance. We explored the clustering from the finest ( $k = 40$ ) to the coarsest level ( $k = 2$ ), settling on  $k = 16$  as it captured the major regional splits sufficiently with little redundancy. Clusters at this level were robustly defined by total variation distance (TVD;  $p < 0.001$ ) and fixation index ( $F_{ST}$ ; Fig. 1a); remarkably, some  $F_{ST}$  values between particularly differentiated Dutch clusters were comparable in magnitude to estimates between European countries (calculated using data from ref. 21; Supplementary Table 1). Some clusters had expansive geographical ranges (for example NHFG, representing individuals from North Holland, Friesland and Groningen), while others neatly distinguished populations on a sub-provincial level (for example, NBE and NBW, representing east and west regions of North Brabant). For visualisation we projected the ChromoPainter coancestry matrix in lower dimensional space using principal component analysis (PCA; Fig. 1b) and assigned cluster labels based on majority sampling location (available for 1422 individuals), arranging neighbouring and genetically similar clusters into cluster groups, as with previous work<sup>6</sup>. The first principal component (PC) of coancestry followed a strong north-south trend (latitude vs mean PC1 per town  $r^2 = 0.52$ ;  $p = 6.8 \times 10^{-72}$ ) with PC2 generally explained by a west-east gradient (longitude vs mean PC2 per town  $r^2 = 0.29$ ;  $p = 3.4 \times 10^{-33}$ ). Further PCs demonstrated more complex relationships with geography (Supplementary Fig. 1).

As previously observed in different populations<sup>6</sup>, the distribution of individuals in this genetic projection generally resembled their geographic distribution (Fig. 1c), with some exceptions. For example, North Brabant is geographically further north than Limburg, but is further separated by PC1 from northern clusters. We explored the possibility that this could instead be explained by relative ancestral affinities to neighbouring lands by modelling the genome of each Dutch individual as a linear mixture of European sources (obtained from ref. 21) using ChromoPainter, retaining source groups that best matched Dutch individuals for at least 5% of the genome<sup>4</sup> (Fig. 2). The resulting profiles of German, Belgian and Danish ancestries were significantly autocorrelated ( $p_{DE}, p_{BE} < 0.0001$ ;  $p_{DK} < 0.001$ ; Moran's I and Mantel's test) and spatially arranged along geographical directions S66°W, N73°E and S73°E respectively, approximately corresponding to declining ancestry gradients directed away from the German and Belgian borders and the North Sea boundary (Fig. 2;  $r_{DE}^2 = 0.31$ ;  $r_{BE}^2 = 0.35$ ;  $r_{DK}^2 = 0.12$ ;  $p_{DE} = 9.4 \times 10^{-119}$ ;  $p_{BE} = 2.7 \times 10^{-133}$ ;  $p_{DK} = 1.1 \times 10^{-39}$ ). The spatial distribution of French ancestry was comparatively



**Fig. 1 The genetic structure of the people of the Netherlands.** **a** fineSTRUCTURE dendrogram of ChromoPainter coancestry matrix showing clustering of 1626 Dutch individuals based on haplotypic similarity. Associated total variation distance (TVD) and fixation index statistics between clusters are shown in the matrix. Permutation testing of TVD yields  $p < 0.001$  for all cluster pairs, indicating that clustering is non-random. Cluster labels derive from Dutch provinces and are arranged into cluster groups for genetically and geographically similar clusters (circled labels). **b** The first two principal components (PCs) of ChromoPainter coancestry matrix for all individuals analysed. Points represent individuals and are coloured and labelled by cluster group. **c** Geographical distribution of 1422 sampled individuals, coloured by cluster groups defined in **a**. Labels represent provinces of the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).



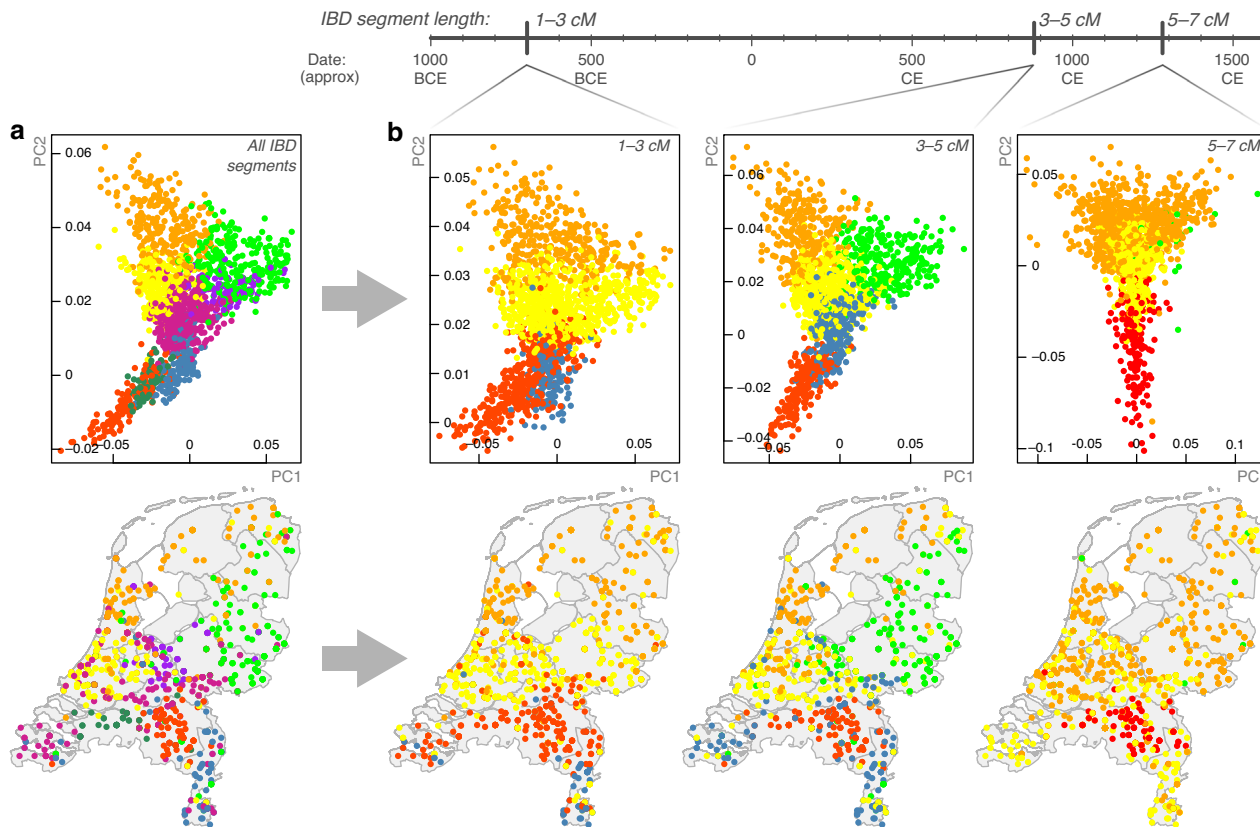
**Fig. 2 The ancestry profile of the Netherlands.** **a** The Netherlands and its geographical relationship to neighbouring lands. **b** German, Belgian, Danish and French haplotypic ancestry profiles for 1422 Dutch individuals. Arrows indicate the predominant directions along which the ancestry gradients are arranged across the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>) and Natural Earth (<https://naturalearthdata.com>).

uniform, with only a modest correlation due east ( $r_{FR}^2 = 0.014$ ;  $p_{FR} = 9.5 \times 10^{-6}$ ). The general trend across the Netherlands was thus of complementary Belgian and German ancestral affinities, decaying with distance from the respective borders. North Brabant, however, showed a greater Belgian profile than Limburg, despite similar, substantial Belgian frontiers in both Dutch provinces. Conversely, the German ancestry profile of Limburg greatly exceeded that of North Brabant, reflecting its 200-kilometre border with Germany and centuries of consequent demographic contact and likely genetic admixture.

**Genome flux and stasis in the Netherlands.** To explore temporal trends in Dutch population structure we called genomic segments of pairwise identity-by-descent (IBD) using RefinedIBD<sup>22</sup>. An IBD haplotype sharing matrix is conceptually similar to a ChromoPainter coancestry matrix<sup>23</sup>, but trades some sensitivity to be more explicitly interpretable. As IBD segment length is inversely related to age<sup>24,25</sup>, different length intervals can inform

on structure at different time depths. Total pairwise IBD between Dutch individuals mirrored the structure observed with ChromoPainter (Fig. 3a), with 8 distinct clusters identified in the IBD sharing matrix that broadly segregated with geography and recapitulated some of the important splits obtained from fineSTRUCTURE, most strikingly the west-east split in North Brabant. Decomposing total IBD by centiMorgan (cM) length into short (1–3 cM), medium (3–5 cM) and long (5–7 cM) bins, we observed stability over time of north-south structure and the emergence of west-east structure embedded in 3–5 cM segments (Fig. 3b), corresponding to an expected time depth around 1120 years ago<sup>25</sup>. As this date and the structure observed is dependent on the (arbitrary) thresholds set for IBD segment length bins, we have also provided an interactive environment in which Dutch population structure can be explored across a range of IBD segment bins (<http://bioinf.gen.tcd.ie/ctg/nlibd>).

Although these observations could potentially be biased by power to detect population structure in longer and shorter bins, the temporally volatile west-east structure contrasts with the



**Fig. 3 The changing genomic structure of the Dutch population over time.** **a** Principal component (PC) analysis of pairwise total identity-by-descent (IBD) for 1626 Dutch individuals (top) and their geographical provenance (bottom). Points represent individuals and are coloured by cluster assignment (mclust on pairwise IBD matrix). **b** PCs (top) and geographical provenance (bottom) for pairwise sharing of 1-3, 3-5 and 5-7 centiMorgan (cM) IBD segments, corresponding to point estimates of expected time depths at ~2700, 1120 and 720 years ago, respectively. Time depths for IBD segment bins have wide distributions<sup>25</sup>; expected values presented here should be interpreted as a guide only and the changing west-east structure over time does not necessarily reflect (for instance) a precisely-timed admixture event. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).

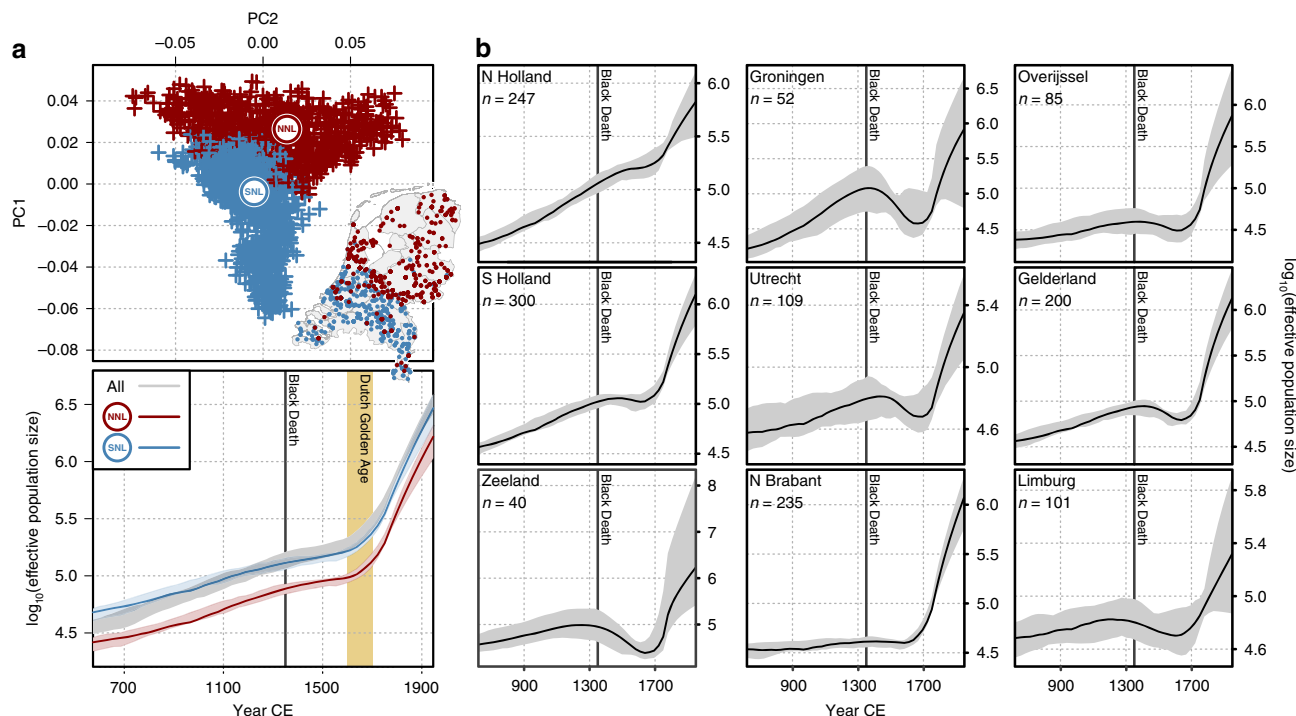
**Table 1 GLOBETROTTER date and source estimates for admixture into the Netherlands.**

Cluster group	Conclusion	Minor	Major	Prop	Date CE	95% CI CE	p
SHOL	One-date multiway	SPA-FRA(2)	GER(5)	0.25	1169	1086-1244	0
ZEE	One-date-multiway	FRA(8)	GER(5)	0.4	1172	771-1773	0
NBE	One-date-multiway	FRA(8)	GER(5)	0.4	1085	939-1262	0
NBW	One-date-multiway	GER(5)	BEL(5)	0.34	1013	668-1383	0
NEN	One-date	SPA-FRA(2)	GER(5)	0.19	1172	925-1364	0
DRO	One-date-multiway	FRA(8)	GER(5)	0.16	1390	1116-1932	0
GLO	One-date	SPA-FRA(2)	GER(5)	0.14	1128	893-1306	0
CEN	One-date	SPA-FRA(2)	GER(5)	0.18	1049	854-1244	0
GEL	One-date	SPA-FRA(2)	GER(5)	0.17	1189	1046-1391	0
NHFG	One-date	GER(9)	DEN(5)	0.36	1060	759-1290	0
LIM	One-date	ITA(8)	GER(5)	0.34	1162	1044-1351	0
ALL	One-date	SPA-FRA(2)	GER(5)	0.25	1088	1004-1111	0

**Minor** and **Major** represent inferred proxy admixing sources. **Prop** represents estimated minor admixture proportion. Admixing sources are derived from ChromoPainter/fineSTRUCTURE clustering of 4514 European reference individuals (Methods); labels represent principal country of origin (SPAin, FRAnce, GERmany, BELgium, DENmark) with cluster numbers arbitrarily assigned within countries. Example coancestry curves are shown in Supplementary Fig. 2.

stability and persistence of old north-south structure and possibly represents a genomic signature of historical demographic flux in the region and its surrounding lands. With this in mind, we investigated possible admixture from outside demographic groups using GLOBETROTTER<sup>14</sup> with 4514 European individuals<sup>21</sup> representing modern proxies for admixing sources. Across the Dutch sample, significant admixture dating to 1088 CE (95% CI 1004-1111 CE) was inferred with the major

contributing source best modelled by modern Germans and the minor source best modelled by southern European groups (France, Spain) (Table 1). This is supported by single-marker ADMIXTURE component estimates showing that the Netherlands has the closest profile to Germanic groups (Supplementary Fig. 3) and is consistent with the ancestry profile gradients detailed in Fig. 2. The timing of the inferred 11th century event was stable across Dutch fineSTRUCTURE clusters (to varying



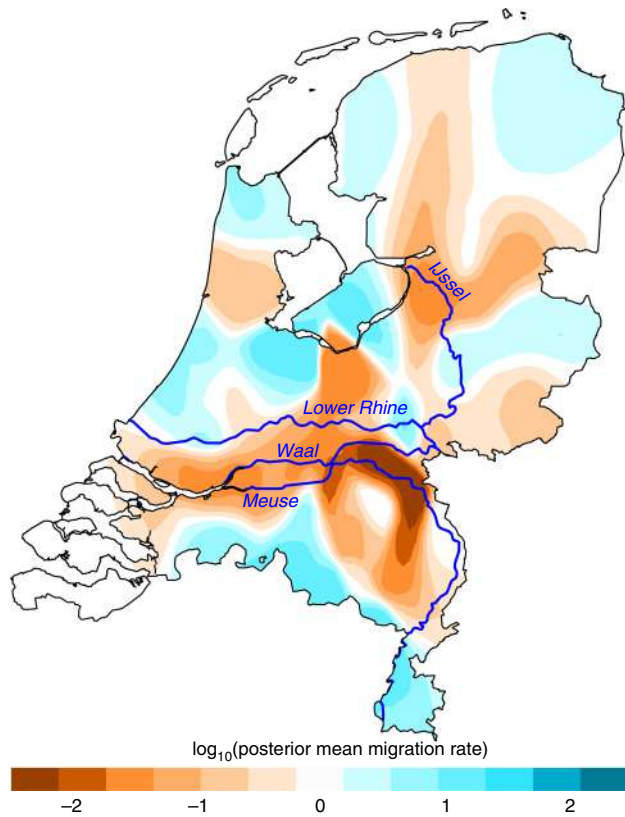
**Fig. 4 Dutch effective population size over time.** **a** Historical change in effective population size ( $N_e$ ) over the past 50 generations for all Dutch individuals and subsets of northerners and southerners. The top plot shows the principal components of ChromoPainter coancestry coloured by the first ( $k = 2$ ) fineSTRUCTURE split, which separates the Dutch population into northern (NNL) and southern (SNL) genetic clusters; inset shows geographical distribution of these individuals. The bottom plot shows growth in effective population size countrywide or per fineSTRUCTURE cluster over the past 50 generations. **b** Historical  $N_e$  trajectories for individual Dutch provinces with more than 40 individuals sampled. For both (**a**, **b**), curves show point estimates for  $N_e$  bounded by a 95% CI estimated from 80 bootstraps of the data (note this is not symmetrically distributed around the point estimates) and assume 28 years per generation and mean year of birth at 1946 CE. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>).

degrees of confidence), suggesting that the signal represents an important period in the establishment of the modern Dutch genome (Table 1); however, given the state of demographic flux in Europe at the time, its exact historical correlate is open to interpretation. Notably, a significant admixture event with a major Danish source was inferred between 759 and 1290 CE in the NHFG cluster group (representing Dutch northern seaboard provinces); this period spans a historical period of recorded Danish Viking contact and rule in northern Dutch territories.

In addition to influence from outside populations, the population structure detailed in Figs. 1 and 3 has likely been shaped by independent demographic histories within the Netherlands. In support of this, we noted that short (1–2 cM) IBD segments shared between northern clusters and provinces outnumbered those shared between southern clusters and provinces (Supplementary Fig. 4), and, as observed previously<sup>2</sup>, northern provinces shared more short segments with southern provinces than southern provinces shared amongst themselves. Together, these results suggest that the north had a smaller ancestral effective population size ( $N_e$ ) than the south and is probably derived from an ancient or historical founder event forming the northern population from a subset of southerners. We formally characterised ancestral trajectories in  $N_e$  between the north and the south of the Netherlands using the nonparametric method IBDNe<sup>26</sup> for the entire Dutch sample and two subsamples representing the principal fineSTRUCTURE north/south split (Fig. 4a), retaining a random sample of 641 individuals from each group. We also characterised historical  $N_e$  within individual Dutch provinces for which genotypes for more than 40 individuals were available. Countrywide,  $N_e$  has grown super-exponentially over the past 50 generations in the Netherlands

(Fig. 4a) and has been consistently lower in the north than the south. Despite this, the pattern of growth in northern and southern groups was identical, with a steady exponential growth up to around 1650 CE, when a major uptick in growth rate was observed. This corresponds to a period of substantial economic development in the Netherlands over the 17th century known to historians as the Dutch Golden Age. Preceding this period, historical  $N_e$  estimates for the entire country and for northern/southern groups showed only a modest response to the Black Death (*Yersinia pestis* plague pandemic) of the 14th century which claimed up to 60% of Europe’s population<sup>27</sup>. Conversely,  $N_e$  estimation within individual Dutch provinces revealed a much more detectable impact of the Black Death (Fig. 4b).

**Genomic signatures of Dutch mobility.** We noted that long (>7 cM) IBD segments, which capture recent shared ancestry, were almost always shared within genetic clusters (and provinces), and rarely between (Supplementary Fig. 4). This indicates a propensity for genetically similar individuals (relatives) to remain mutually geographically proximal, suggesting a degree of sedentism that has likely influenced Dutch population structure over time. It has also previously been argued that genetic structure in the Netherlands may be partially rooted in geographic obstacles imposed by the country’s major waterways<sup>1</sup> so we explicitly modelled genetic similarity as a function of geographic distance using EEMS<sup>16</sup> to infer migrational hot and cold spots (Fig. 5). The resulting effective migration surface showed several apparent barriers to gene flow, the strongest and most contiguous of which runs in an east-west direction across the Netherlands overlapping the courses of the Rhine, Meuse and Waal rivers. This inferred migrational boundary also approximately



**Fig. 5 The effective migration surface of the Netherlands.** Contour map shows the mean of 10 independent EEMS posterior migration rate estimates between 800 demes modelled over the land surface of the Netherlands. A value of 1 (blue) indicates a tenfold greater migration rate over the average;  $-1$  (orange) indicates tenfold lower migration than average. The courses of major rivers are included to highlight their correlation with migrational cold spots. Map boundary data from the Database of Global Administrative Areas (GADM; <https://gadm.org>); river course data from Natural Earth (<https://www.naturalearthdata.com>).

corresponds to the geographical division determining the principal fineSTRUCTURE split between northern and southern Dutch populations (Fig. 4a) as well as the geographical boundaries between clusters inferred from ancient IBD segments (Fig. 3b), suggesting that these rivers have been a historically persistent determinant of Dutch population structure.

**GWAS confounding by fine-grained structure.** As population structure confounds GWAS (for example due to stratification of cases and controls between subpopulations), we investigated the extent to which haplotype sharing captures confounding structure in a Dutch sample of 1963 cases of amyotrophic lateral sclerosis (ALS) and 2774 controls from a recent multi-population GWAS for ALS<sup>19</sup>. PCs of the haplotypic ChromoPainter coancestry matrix for these 4737 individuals explained substantially more variance in ALS phenotype than PCs calculated from single nucleotide polymorphism (SNP) genotypes alone, indicating latent structure captured by ChromoPainter that is stratified between cases and controls (Fig. 6a). To estimate the extent to which this stratified structure confounds GWAS we calculated case-control association statistics using a logistic model covarying for either 20 ChromoPainter PCs or 20 SNP PCs and estimated the linkage disequilibrium (LD) score regression intercepts for both sets of resulting summary statistics. An intercept higher than 1 indicates confounding in the GWAS; Fig. 6a shows that GWAS statistics calculated with ChromoPainter PCs as covariates are less

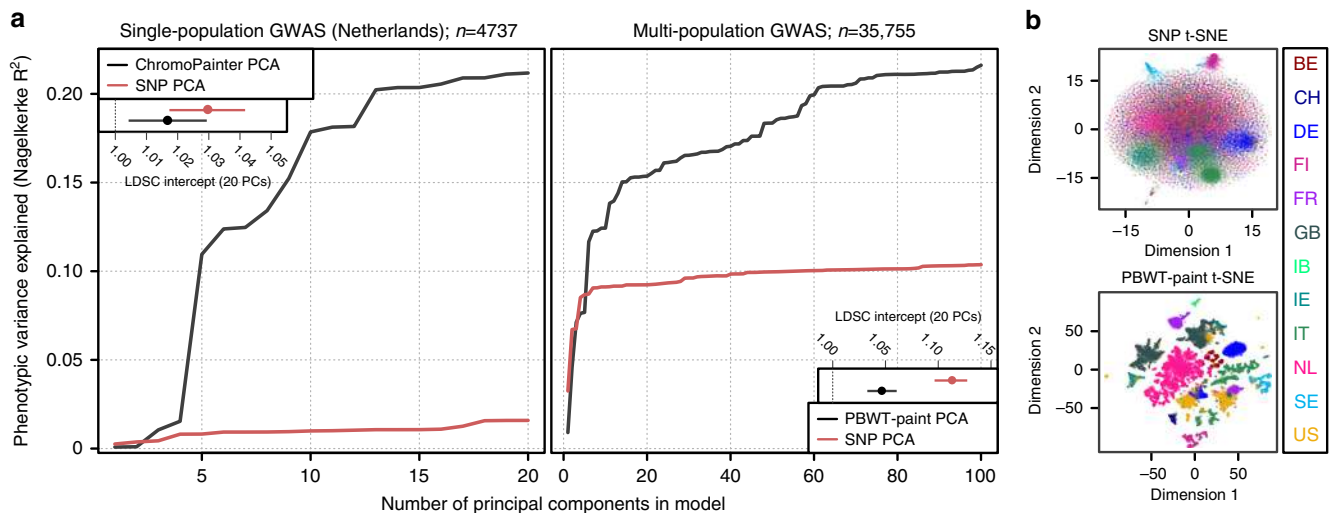
confounded than statistics using SNP PCs, albeit with overlapping confidence intervals for the relatively small Dutch sample. To more adequately represent the large-scale multi-population data typically used in modern GWAS, we extended our analysis to the full ALS case-control dataset from which the Dutch data derive<sup>19</sup>, including 35,755 individuals from twelve European countries and the USA. For computational tractability, instead of ChromoPainter we used PBWT-paint (<https://github.com/richarddurbin/pbwt>), a scalable approximate haplotype painting method based on the positional Burrows-Wheeler transform<sup>28</sup>. When run on our original Dutch dataset of 1626 individuals, the structure rendered by PBWT-paint was almost identical to ChromoPainter ( $r_{PC1}^2 = 0.99$ ;  $r_{PC2}^2 = 0.98$ ; Supplementary Fig. 5), indicating its suitability for this analysis. PBWT-paint captured pervasive global and local structure in the multi-population GWAS data that both separated and subdivided countries (Fig. 6b). Top PCs of PBWT-paint coancestry explained substantially more variance in phenotype than SNP PCs and GWAS statistics including PBWT-paint PCs as covariates were significantly less confounded than statistics corrected by SNP PCA (Fig. 6a, LD score regression intercepts).

## Discussion

The genomes of modern humans contain a detailed record of the intricate histories that shaped them. Genomic signatures of these histories are often reflected in present-day population structure and have the potential to confound genomic studies of health and disease through stratification across phenotypic categories. Here, we have studied the Netherlands as a model population, harnessing information from shared haplotypes and recent developments in spatial modelling to gain intricate insights into the geospatial distribution and likely origin of Dutch population genetic structure. The structure identified through shared haplotypes is surprisingly strong; some Dutch genetic clusters identified this way are more mutually distinct (by  $F_{ST}$ ) than whole European countries. We have also introduced a novel use of length-binned IBD sharing combined with PCA and Gaussian mixture model-based clustering to characterise changing population structure over time, revealing transient genetic structure layered over strong and stable north-south differentiation in the Netherlands. This is contextualised by somewhat distinct demographic histories between genetic groups in the Netherlands, with consistently lower  $N_e$  in the north than the south. A potential source of the north-south differentiation is impaired migration across the east-west courses of the Rhine, Meuse and Waal, which effectively separate southern Dutch populations from the north. The population structure observed in the Netherlands is especially remarkable when considered in terms of the country's size and extensive infrastructure; notably Denmark, which is roughly equal in geographical area, is genetically homogeneous, forming only a single cluster when interrogated using fineSTRUCTURE<sup>29</sup>, despite its island-rich geography. Both the United Kingdom and Ireland also exhibit at least one large indivisible cluster constituting a large fraction of the population<sup>4-6</sup>, however no extraordinarily large clusters dominate the Dutch sample. Mean  $F_{ST}$  between Dutch clusters also greatly outmeasures that observed between Irish clusters, suggesting that the extent of population differentiation is higher in the Netherlands, despite Dutch land area being less than half that of the island of Ireland.

While coarse geographical trends in Dutch genetic structure have previously been described using single-marker PCA<sup>1</sup>, our use of shared haplotypes reveals structure at a much higher resolution, differentiating subpopulations between, and sometimes within, provinces (Fig. 1). As a striking example, individuals from the east and west of North Brabant (NBE and NBW in





**Fig. 6 Fine-grained population structure and genome-wide association study (GWAS) confounding.** **a** Variance in phenotype (amyotrophic lateral sclerosis) explained by principal components (PCs) for a single-population Dutch GWAS (left) and a multi-population GWAS (right). Insets show linkage disequilibrium score regression (LDSC) intercept terms (a summary estimate of GWAS confounding) when the first 20 single nucleotide polymorphism (SNP)-based PCs (SNP PCA) or the first 20 haplotype-based PCs (ChromoPainter/PBWT-paint PCA) are included as GWAS covariates. **b** Summary visualisations (t-distributed stochastic neighbour embedding, t-SNE) of local and global structure in the multi-population GWAS based on SNP genotypes (top) or haplotype sharing inferred using the scalable PBWT-paint chromosome painting algorithm (bottom). Individuals are coloured by country of origin; labels (right) follow ISO 3166-1 country codes, except IB, which was labelled Iberia (containing Spanish and Portuguese data) in the original GWAS dataset. PCA, principal component analysis; PBWT, positional Burrows-Wheeler transform.

Fig. 1) are mutually genetically distinguishable and are more distinct from clusters to their north than Limburg, despite being geographically closer. This deviation from haplotype sharing mirroring geography appears to be driven by strong genetic affinity to Belgium (Fig. 2), reflecting a long history of demographic and sovereign overlap across a 100 km frontier spanning the modern Dutch-Belgian border. In contrast, the majority of ancestral influence in Limburg, which also shares a substantial border with Belgium, is equally split between Belgium to the west and Germany to the east. Notably, the Belgian border with the south of Dutch Limburg is almost entirely described by the course of the Meuse, which may have acted as a historical impediment to migration, thus distinguishing individuals in this region genetically. This is reflected in IBD clustering, in particular the distinction of southern Limburgish individuals from the rest of the Netherlands in short (1–3 cM) segments, which otherwise only describe coarse north-central-south structure (Fig. 3). Future work explicitly modelling Dutch-Belgian and Dutch-German frontiers using additional Belgian and German genetic data with associated geography will resolve the historical and present-day role of the Meuse in distinguishing distinct population clusters in the south of the Netherlands.

Similarly to North Brabant, groups of individuals in North and South Holland show significant genetic separation despite mutual geographic proximity. While we have chosen to group the four South Holland clusters for visual brevity in Fig. 1, they are robustly distinct by TVD permutation analysis ( $p < 0.001$ ), indicating that significant population differentiation exists even within South Holland. Migration and admixture in the highly urbanised *Randstad* has been proposed as a driver of genetic diversity and loss of geographic structure in this region<sup>1</sup>; the overlaid geographical distribution of regional ancestry profiles (Fig. 2) for this area lends support to this hypothesis. However, the geographical ranges of the four South Holland clusters are somewhat independent (Supplementary Fig. 6), indicating that some degree of genetic structure has survived this urbanisation. Previous studies have highlighted the correlation between

decreasing autozygosity and increased urbanisation<sup>30</sup>; future work leveraging the ChromoPainter/fineSTRUCTURE framework coupled with length-binned IBD and Gaussian mixture model-based clustering will more explicitly delineate the interplay between urbanisation and population structure over time. To this end, highly urbanised areas such as the *Randstad* will be particularly informative.

The principal fineSTRUCTURE split in the Netherlands describes north-south genetic differentiation (Fig. 1) that is strong and persistent over time (Fig. 3). We hypothesised that this reflects partially independent demographic histories so we estimated ancestral  $N_e$  for northern (NNL) and southern (SNL) Dutch fineSTRUCTURE populations, revealing superexponential growth in both populations with a sudden increase in rate during the 17th century (Fig. 4a). Historical  $N_e$  follows the same approximate trajectory for both populations but is consistently lower for the northern cluster, corroborating previous observations of increased homozygosity in northern Dutch populations<sup>1</sup> and consistent with a model of northerners representing a founder isolate from southerners (although a more complex demographic model may better explain these observations)<sup>1,2</sup>. The apparent absence of  $N_e$  decline in 14th-century Netherlands initially hints at the possibility that the Black Death had a weaker impact in the region than elsewhere in Europe; although this agrees with the views of some historians, it is hotly debated by others<sup>31</sup>. Per province, however, most  $N_e$  estimates display a prominent dip at this time (Fig. 4b), suggesting that merging non-randomly mating subpopulations into a countrywide group (Fig. 4a) artificially inflates diversity, thus smoothing over any population crash following the Black Death. Population structure is thus important when estimating  $N_e$  and trends countrywide and in NNL and SNL clusters (Fig. 4a) should be interpreted carefully: it is possible that a substantial population crash brought about by the Black Death might have had only a marginal impact on the overall effective size of the breeding population in these merged groups. Indeed, the rate of exponential growth in countrywide  $N_e$  (Fig. 4a) is marginally shallower in the 10 generations

following the Black Death (0.024; 95% CI 0.0235–0.0251) compared to the 10 generations prior (0.017; 95% CI 0.016–0.018), indicating enduring strain on the overall Dutch population prior to its recovery in the 17th century.

Previous works have hinted that north-south genetic differentiation in the Netherlands may have been facilitated by cultural division between the predominantly Catholic south and the Protestant north<sup>1</sup>. Given that the north-south structure observed in 1–3 cM IBD bins (expected time depth ~700 BCE) greatly precedes different forms of Christianity (Fig. 3), our data support a model in which the Protestant Reformation of the 16th and 17th centuries exploited pre-existing demographic subdivisions, leading to correlation between distinct cultural affinities and clusters of genetic similarity which has potentially been further strengthened by assortative mating among religious groups<sup>32</sup>. Geographical modelling supports the role of migrational boundaries in establishing and maintaining this population substructure, especially rivers (Fig. 5). A substantial belt of low inferred migration runs across the Netherlands, corresponding closely to the roughly parallel east-west courses of the Lower Rhine, Waal and Meuse rivers and correlating with the geographical boundary of the principal north-south fineSTRUCTURE split. Absolute assignment of causality to these geographical correlates is, however, not possible and, given the dense network of waterways in the Netherlands, could be misleading. For example, a strong migrational cold spot in the east of the Netherlands runs parallel to the IJssel (Fig. 5), but could potentially be better explained by the course of the Apeldoorn Canal, a politically fraught waterway constructed in the early 19th Century. Similarly, a cold spot in the northwest directly overlays the North Sea Canal (completed in 1876). As both of these are human-made waterways, it is not certain whether their courses are consequences or determinants of low movement of people across their paths.

As well as internal geography, outside populations have also played an important and significant role in the establishment of population structure in the Netherlands (Fig. 2; Table 1); however the variety and extent of demographic upheaval and mobility of European populations over history obscure the likely historical provenance of most inferred admixture signals. As an important exception, however, ancestry profiles show a small but significant contribution of Danish haplotypes in the north and west of the Netherlands, a possible vestige of Viking raids in coastal areas in the 9th and 10th centuries. This is corroborated by an inferred GLOBETROTTER single-date admixture event in the NHFG (North Holland, Friesland and Groningen) cluster (Fig. 1) between 759 and 1290 CE with Danish haplotypes as a major admixing source (Table 1). The demographic legacy of more than a century of Danish Viking raids and settlement in the Netherlands has been the subject of some debate; from our data, it appears that the modern Dutch genome has indeed been partially shaped by historical Viking admixture. This Danish Viking contact is contemporaneous with a critical period in the establishment of the modern Dutch genome from other outside sources (1004–1111 CE; Table 1), although the precise historical correlates of the admixture events detected in the remaining Dutch regions are less obvious. Future densely sampled ancient DNA datasets from informative time depths in the Netherlands and northwest Europe will enable direct estimation of ancestral population structure, admixture, demographic affinities and effective population sizes, improving precision over the current study which depends on proxy patterns of haplotype sharing between modern individuals. Similarly, regional ancestry and admixture inference are limited by the use of modern proxy populations in place of true ancestral sources; nevertheless, there are ample advantages to the use of modern data, including large

sample size and relevance to research on modern human health and disease. In particular, as in our previous work in Ireland<sup>6</sup>, samples in the current Dutch dataset were not specifically selected to have pure ancestry in each geographical area (eg all grandparents from the same region<sup>4</sup>) meaning the degree of structure observed is not idealised or exaggerated by sampling, but instead representative of the structure expected in any GWAS that includes Dutch data.

We therefore explored the impact of fine-scale genetic structure described in this study and others<sup>4–12</sup> on GWAS statistics, using the ALS study from which the Dutch data derive as an exemplar trait. Generally, population-based PCs should not predict case/control status (in the absence of any disease-ancestry interaction); if they do, this indicates that (sub)populations are stratified between cases and controls, introducing bias that artificially inflates GWAS statistics. In both Dutch-only and multi-population analyses, fine-scale genetic structure detected by haplotype sharing (ChromoPainter or PBWT-paint) explained substantially more variance in phenotype (ALS case/control status) than standard SNP-only PCA (Fig. 6a). This demonstrates the power of shared haplotypes to simultaneously capture subtle genetic structure within single countries (that is potentially invisible to standard single-marker PCA) along with broader structure between countries and potential cryptic technical artefacts such as platform- or imputation-derived bias. We found that shared haplotypes are effective for controlling GWAS inflation: statistics calculated using haplotype-based PCs as covariates showed lower overall confounding than single marker-based covariates, as measured by LD score regression intercepts (Fig. 6a). In the age of large-scale, single-country and cross-population biobanks, the additional power of haplotype sharing methods to detect fine-scale local population structure will be crucial for ensuring robust GWAS results unconfounded by ancestry. For example, a recent study of latent structure in the UK Biobank demonstrated that a GWAS for birth location returned significant loci even after correction for 40 single-marker PCs<sup>33</sup>, suggesting that residual fine-grained population structure may influence other GWAS from this cohort (although others suggest a role for socioeconomically-driven migration in this phenomenon<sup>34</sup>). Ongoing developments in scalable haplotype sharing algorithms such as PBWT-paint will help to address this problem by facilitating the creation of biobank-scale haplotype sharing resources, simultaneously improving studies of human health and disease and enabling large-scale, fine-grained population genetic studies of human demography. Such resources will likely be particularly useful in studies of rare variation, motivating future work exploring the efficacy of such strategies in correcting confounding where rare variation is a factor.

## Methods

**Data and quality control.** We mapped fine-grained genetic structure in the Netherlands using a population-based Dutch ALS case-control dataset ( $n = 1626$ ; subset of stratum sNL3 from a GWAS for amyotrophic lateral sclerosis<sup>19</sup>) and a European reference dataset subsampled from a GWAS for multiple sclerosis<sup>21</sup> (MS;  $n = 4514$ ; EGA accession ID EGAD00000000120 [<https://www.ebi.ac.uk/ega/datasets/EGAD00000000120>]). 1422 Dutch individuals had associated residential data (hometown at time of sampling) which were used for geographical analyses. For estimating GWAS confounding, we separately analysed the Netherlands on its own using a larger ALS case/control dataset ( $n = 4753$ ; strata sNL1, sNL3 and sNL4 from ref. <sup>19</sup>) and the complete multi-population GWAS dataset<sup>19</sup> ( $n = 36,052$ ) from which this Dutch subset was derived. Data handling for estimating confounding is further described under “Estimating GWAS confounding” below. For population structure analyses, we applied quality control (QC) using PLINK v1.9<sup>35</sup>; briefly we removed samples with high missingness (>10%), high heterozygosity (>3 median absolute deviations from median) and single-marker PCA outliers (>5 standard deviations from mean for PCs 1–20). We also filtered out A/T and G/C SNPs and SNPs with minor allele frequency <0.05, high missingness (>2%) or in Hardy Weinberg disequilibrium ( $p < 1 \times 10^{-6}$ ). Before running ChromoPainter/fineSTRUCTURE we retained only one individual from any pair or group that

exhibited greater than 7.5% genomic relatedness ( $\hat{\pi}$ ) and removed SNPs with any missing genotypes as the algorithm does not tolerate missingness or relatedness well. For European reference data we also removed individuals suggested by the QC of the source study<sup>21</sup> and we extracted individuals only of European descent. As this European dataset included MS patients, we filtered out SNPs in a 15 Mb region surrounding the strongly associated HLA locus (GRCh37 position chr6:22,915,594–37,945,593) to avoid bias generated from this association, following previous works. The final Dutch and European reference datasets contained 374,629 SNPs and 363,396 SNPs respectively at zero missingness. The merge of these datasets contained 147,097 SNPs at zero missingness. Data were phased per chromosome with the 1000 Genomes Project phase 3 reference panel<sup>36</sup> using SHAPEIT v2<sup>37</sup> (for ChromoPainter/fineSTRUCTURE) and Beagle v4.1 (for IBD estimation). For these and all subsequent runs of SHAPEIT and ChromoPainter, we used the 1000 Genomes Project Phase 3 genetic map; IBD analyses with Beagle were carried out using the Hapmap phase 2 genetic map<sup>38</sup> as used in the RefinedIBD and IBDNe source papers<sup>22,26</sup>. Both programmes were run with default settings; allele concordance was checked prior to phasing (SHAPEIT: -check; Beagle: conform-gt utility).

**fineSTRUCTURE analysis.** We used ChromoPainter/fineSTRUCTURE<sup>20</sup> to detect fine-grained population structure using default settings. In brief, each individual was painted using all other individuals (-a 0 0), first estimating  $N_e$  and  $\mu$  (switch rate and mutation rate) with 10 expectation-maximisation (EM) iterations (using all samples and chromosomes), then the model was finally run using these parameter estimates. The fineSTRUCTURE Markov chain Monte Carlo (MCMC) model was then run on the resulting Dutch coancestry matrix with two chains for 3,000,000 burnin and 1,000,000 sampling iterations, sampling every 10,000 iterations. To define European clusters for use in GLOBETROTTER and ancestry profile estimation we instead used 1,000,000 burnin and sampling iterations, sampling every 1000 iterations (due to large sample size). We extracted the state with the maximum posterior probability and performed an additional 10,000 burnin iterations before inferring the final trees using both the climbtree and maximum concordance methods. For all subsequent analyses the maximum concordance tree was used.

**Cluster robustness and differentiation.** To assess the robustness of clustering in the Dutch data we calculated TVD<sup>4</sup> and  $F_{ST}$ . TVD is a distance metric for assessing the distinctness of pairs of clusters, calculated from the ChromoPainter chunk-length matrix. TVD is calculated as the sum of the absolute differences between copying vectors for all pairs of clusters, where the copying vector for a given cluster  $A$  is a vector of the average lengths of DNA donated to individuals in  $A$  by all clusters. Intuitively, the TVD of two clusters reflects distance between those clusters in terms of haplotype sharing amongst all clusters, and is a meaningful method for assessing the effectiveness of fineSTRUCTURE clustering. To assess whether the observed clustering performed better than chance we permuted individuals between cluster pairs (maintaining cluster size) and calculated the number of permutations that exceeded our original TVD score for that pairing of clusters. We used 1000 permutations where possible, and otherwise used the maximum number of unique permutations.  $P$  values were calculated from the number of permutations greater than or equal to the observed TVD divided by the total permutations; all  $p$ -values were less than 0.001, indicating robust clustering. We generated a TVD tree for clusters from the  $k = 16$  fineSTRUCTURE split by merging pairs of clusters with the lowest TVD successively using methods developed in ref. <sup>8</sup>, with the goal of providing an alternative representation of cluster relationships that is independent of sample size (Supplementary Fig. 7). The tree was built in  $k-1$  steps, with TVD recalculated at each step from the remaining populations. Branch lengths were scaled proportional to the TVD value of the corresponding pair of populations using adapted code from the original paper<sup>8</sup>. Finally, to assess cluster differentiation independently of the ChromoPainter model,  $F_{ST}$  was calculated between Dutch clusters using PLINK 1.9. For this analysis we used the SNP overlap between Dutch and European datasets, pruning for LD (--indep-pairwise 1000 50 0.25) and simultaneously calculating  $F_{ST}$  between European countries present in ref. <sup>21</sup> for comparison.

**Ancestry profiles.** We assessed the ancestral profile of Dutch samples in terms of a European reference made up of 4514 European individuals<sup>21</sup> from Belgium, Denmark, Finland, France, Germany, Italy, Norway, Poland, Spain and Sweden. European samples were first assigned to homogeneous genetic clusters using the fineSTRUCTURE maximum concordance tree<sup>6</sup> to reduce noise in painting profiles. We then modelled each Dutch individual's genome as a linear mixture of the European donor groups using ChromoPainter, and applied ancestry profile estimation method developed in ref. <sup>4</sup> and implemented in GLOBETROTTER<sup>14</sup> (num.mixing.iterations: 0). This method estimates the proportion of DNA which is most closely shared with each individual from each donor group calculated from a normalised ChromoPainter chunklength output matrix, and then implements a multiple linear regression of the form

$$Y_p = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_G X_G \quad (1)$$

to correct for noise caused by similarities between donor populations. Here,  $Y_p$  is a

vector of the proportion of DNA that individual  $p$  copies from each donor group, and  $X_g$  is the vector describing the average proportion of DNA that individuals in donor group  $g$  copy from other donor groups  $G$ , including their own. The coefficients of this equation  $\beta_1 \dots \beta_G$  are thus interpreted as the “cleaned” proportions of the genome that target individual  $p$  copies from each donor group, hence the ancestral contribution of each donor group to that individual. The equation is solved using a non-negative-least squares (NNLS) function such that  $\beta_g \geq 0$  and the sum of proportions across groups equals 1. We discarded European groups that contributed less than 5% total to any individual, and refit to eliminate noise. We then aggregated sharing proportions across donor groups (genetically homogenous clusters) from the same country to estimate total sharing between an individual and a given country to investigate the regional distribution of sharing profiles. Auto-correlation of ancestry profiles was assessed by Moran's  $I$  and Mantel's test (10,000 permutations) in R version 3.2.3. Geographical directions of ancestry gradients were determined by rotating the plane of latitude-longitude between  $0^\circ$  and  $360^\circ$  in  $1^\circ$  steps and finding the axis  $Y$  that maximised the coefficient of determination for the linear regression  $Y \sim A_c$ , where  $A_c$  is the aggregated ancestry proportion for country  $c$ .

Additionally we compared the ancestry profiles estimated by the NNLS method to those estimated using the recently developed Bayesian algorithm SOURCEFIND<sup>13</sup>. We ran SOURCEFIND on the ChromoPainter output described above using 50,000 burnin and 200,000 MCMC iterations, sampling every 5000 iterations. For each Dutch individual we took the weighted average (weighted by posterior probability) of ancestry estimates with the highest posterior probability taken from 50 independent runs of the algorithm. We aggregated sharing portions across donor groups from the same country to estimate total sharing between an individual and a given country to investigate the regional distribution of sharing profiles. Ancestry gradients generated by each method were regressed against one another to estimate correlation. We report both the results of both NNLS (Fig. 2) and SOURCEFIND (Supplementary Fig. 8) for comparison.

**Identity-by-descent analyses.** IBD segments were called in phased data using RefinedIBD<sup>22</sup> (default settings) to generate pairwise matrices of total length of IBD shared between individuals for bins of different segment lengths. To identify population structure captured by IBD sharing patterns we performed PCA on these matrices using the prcomp function in R version 3.2.3<sup>39</sup> and clustered the IBD matrices using a Gaussian mixture model implemented in the R package mclust<sup>40</sup>. Plots of model selection are shown in Supplementary Fig. 9. We note that while previous work<sup>23</sup> has shown that IBD matrices underperform the linked ChromoPainter matrix in identifying population structure, they are arguably more interpretable for visualising temporal change as they can be subdivided into cM bins corresponding to different time periods, a feature leveraged by emerging work on local population structure<sup>25</sup>. Patterns in IBD sharing that identify population subgroups in older (shorter) cM bins which are preserved in more recent (longer) bins are interpreted as persistent population structure that has been influenced by mating patterns in old and recent generations. Structure which emerges in a specific cM bin and is lost is likely to reflect transient changes in panmixia that have not necessarily persisted. We approximated the age of segments in a given cM bin using equation s19 from ref. <sup>25</sup>, under the assumption that the population is sufficiently large:

$$\lim_{N \rightarrow \infty} E[T | \mu \leq l \leq v] = 75 \left( \frac{1}{L_1} + \frac{1}{L_2} \right), \quad (2)$$

where  $T$  is the random coalescence time in generations,  $l$  is the length of a segment (in base pairs),  $\mu$  and  $v$  are the upper and lower segment length bounds of the interval (in base pairs) and  $L_2$  and  $L_1$  are the upper and lower bounds of the interval rescaled to centiMorgan (i.e. multiplied by 100, where  $r$  is the recombination rate). For the age estimates given in Fig. 3, we multiplied the expected coalescence time in generations by the approximate human generation time (28 years).

**Inferring admixture events.** To infer and date admixture events from European sources we ran GLOBETROTTER<sup>14</sup> with the Netherlands dataset as a whole and in individual cluster groups defined from the Dutch fineSTRUCTURE maximum concordance tree (Fig. 1). To define European donor groups we used the European fineSTRUCTURE maximum concordance tree to ensure genetically homogenous donor populations. We used ChromoPainter v2 to paint Dutch and European individuals using European clusters as donor groups (estimating  $N_e$  and  $\mu$  using the weighted average of 10 EM iterations on chromosomes 1, 8, 15 and 20, using all samples). This generated a copying matrix (chunklengths file) and 10 painting samples for each Dutch individual. GLOBETROTTER was run for 5 mixing iterations twice: once using the null.ind:1 setting to test for evidence of admixture accounting for unusual linkage disequilibrium (LD) patterns and once using null.ind:0 to finally infer dates and sources. We further ran 100 bootstraps for the admixture date and calculated the probability of no admixture as the proportion of nonsensical inferred dates (<1 or >400 generations). Confidence intervals were calculated from the bootstraps from the standard model (null.ind:0) using the empirical bootstrap method, and a generation time of 28 years.

**ADMIXTURE analysis.** We performed ADMIXTURE analysis<sup>41</sup> on the combined Dutch and European samples to explore single marker-based population structure in a set of 41,675 SNPs (LD-pruned using PLINK 1.9;  $r^2 > 0.1$ ; sliding window 50 SNPs advancing 10 SNPs at a time). ADMIXTURE was run for  $k = 1-10$  populations, using 5 EM iterations at each  $k$  value. The  $k$  value with the lowest cross-validation error was selected for further analysis using 15 fold cross-validation; where two  $k$  values had equal CV-error the lower  $k$  value was taken for parsimony (Supplementary Fig. 10). We analysed the distribution of proportions for each ADMIXTURE cluster across the Dutch dataset, and its relationship with geography.

**Computing mean pairwise shared IBD within and between groups.** We compared IBD sharing within and between both clusters and provinces (Supplementary Fig. 4) using the mean number of segments within a given length range (e.g. 1–2 cM) shared between individuals. To calculate this mean for a single group of size  $N$  with itself the denominator was  $(N^2 - N)/2$ ; when comparing two groups of sizes  $N$  and  $M$  the denominator was  $NM$ .

**Estimating recent changes in population sizes.** We used IBDNe<sup>26</sup> to estimate historical changes in  $N_e$ . IBDNe leverages information from the length distribution of IBD segments to accurately estimate effective population size over recent generations, with a resolution limit of about 50 generations for SNP data. We followed the authors' protocol and detected IBD segments using IBDseq version r1206<sup>42</sup> with default settings and ran IBDNe on the resulting output with default settings, removing IBD segments shorter than 4 cM (minibd = 4, the recommended threshold for genotype data). We compared estimated  $N_e$  with recorded census size (<https://opendata.cbs.nl/staline/#/CBS/nl/dataset/37296ned/table?ts=1520261958200>) for approximately equivalent dates (starting at 1946 CE for generation 0 and assuming 1 generation is 28 years) and found that for generations 0 - 3 our  $N_e$  estimates were approximately 1/3 of the census population (Supplementary Fig. 11), which follows expectation if lifespan is  $\sim 3\times$  the generation time<sup>26,43</sup>. The slope of the ratios for the three generations is near zero suggesting that our model tracks well with the census population; this is consistent with reported expectation<sup>26</sup>.

**Estimating effective migration surfaces.** To model geographic barriers to geneflow in the Netherlands we ran EEMS<sup>16</sup>. This software provides a visualisation of hot and coldspots for geneflow across a habitat using a geocoded genetic dataset. To run EEMS, we generated an average pairwise genetic dissimilarity matrix from our genotype data using the bed2diffs utility provided with the software. We initially ran the EEMS model with 10 randomly initialised MCMC chains for a short run of 100,000 burn-in and 200,000 sampling iterations, thinning every 999 iterations, to find a suitable starting point. For these runs we placed the data in 800 demes and used default settings with the following adjustments to the proposal variances: qEffectProposals2 = 0.000088888888; qSeedsProposals2 = 0.7; mEffectProposals2 = 0.7. The resulting chain with the highest log-likelihood was then used as the starting point for a further ten chains for 1,000,000 burn-in iterations and 2,000,000 sampling iterations, thinning every 9999 iterations. The model was run with the following adjustments to the proposal variances: qEffectProposals2 = 0.000088888888; qSeedsProposals2 = 0.7; mEffectProposals2 = 0.7. We plotted the results of our analysis using the rEEMSplot package in R and modified the resulting vector graphics using Inkscape v0.91 to remove display artefacts caused by non-overlapping polygons. MCMC convergence was assessed by inspecting the log-posterior traces (Supplementary Fig. 12).

**Estimating GWAS confounding.** To examine the contribution of observed fine-grained population structure to GWAS confounding, we estimated how well phenotype could be predicted by principal components of haplotype sharing matrices in a 2016 GWAS for ALS<sup>19</sup>, comparing our results to those obtained using standard single marker PCA. We separately analysed 1,060,224 zero-missingness Hapmap3 SNPs that passed QC in the original GWAS for Dutch data alone (1963 cases, 2774 controls) and for the complete multi-population GWAS (12,480 cases, 23,275 controls). Haplotypes for unrelated individuals ( $\hat{r} < 0.075$ ) were phased using SHAPEIT v2<sup>37</sup> and painted in terms of one another using ChromoPainter v2<sup>20</sup> for the Dutch dataset (estimating  $N_e$  and  $\mu$  using the weighted average of 10 EM iterations on chromosomes 1, 8, 15 and 20 in 10% of samples), and PBWT-paint (<https://github.com/richarddurbin/pbwt>) for the considerably larger multi-population GWAS dataset. PBWT-paint is a fast approximate implementation of ChromoPainter suitable for large datasets. PCs of the resulting coancestry matrices were calculated using the fineSTRUCTURE R tools (<http://www.paintmychromosomes.com>), removing extreme haplotype PCA outliers ( $>20$  SD from mean on PC1-10) followed by repainting as an additional QC step. For comparison we also calculated PCs on independent markers from the SNP datasets using Plink v1.9, first removing long range LD regions<sup>44</sup> ([https://genome.sph.umich.edu/wiki/Regions\\_of\\_high\\_linkage\\_disequilibrium\\_\(LD\)](https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD))) and pruning for LD ( $--indep-pairwise$  500 50 0.8). Variance in ALS phenotype explained by ChromoPainter/PBWT-paint PCs and SNP PCs (Nagelkerke  $R^2$ ) was estimated using the glm() function and fmsb package<sup>45</sup> in R version 3.2.3. To estimate confounding in GWAS inflation, we implemented a logistic regression model GWAS ( $--logistic$ ) in PLINK v1.9 for each dataset using a range of ChromoPainter/

PBWT-paint PCs or SNP PCs (10, 20, 30 and 40 PCs) as covariates and ran LD score regression<sup>46</sup> on the resulting summary statistics using recommended settings (Fig. 6 and Supplementary Fig. 13). Structure evident in the PBWT-paint matrix was visualised and contrasted with corresponding SNP data in 2 dimensions using t-distributed stochastic neighbour embedding (t-SNE)<sup>47</sup> implemented in the Rtsne package in R version 3.2.3 (5000 iterations; perplexity 30; top 100 PCs provided as initial dimensions).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data used in this study are available for academic use through the Project MinE Consortium at <https://www.projectmine.com/research/data-sharing/>. MS GWAS data used for European reference populations were downloaded from the European Genome-phenome Archive under accession EGAD00000000120. Data availability subject to any conditions outlined by source studies.

Received: 15 January 2020; Accepted: 21 August 2020;

Published online: 11 September 2020

## References

- Abdellaoui, A. et al. Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).
- Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
- Lawson, D. J. et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* <https://doi.org/10.1007/s00439-019-02014-8> (2019).
- Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
- Gilbert, E. et al. The Irish DNA Atlas: revealing fine-scale population structure and history within Ireland. *Sci. Rep.* **7**, 17199 (2017).
- Byrne, R. P. et al. Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* **14**, e1007152 (2018).
- Gilbert, E. et al. The genetic landscape of Scotland and the Isles. *Proc. Natl Acad. Sci. USA* **116**, 19064–19070 (2019).
- Kerminen, S. et al. Fine-Scale Genetic Structure in Finland. *G3* **7**, 3459–3468 (2017).
- Takeuchi, F. et al. The fine-scale genetic structure and evolution of the Japanese population. *PLoS One* **12**, e0185487 (2017).
- Raveane, A. et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci. Adv.* **5**, eaaw3492 (2019).
- Saint Pierre, A. et al. The genetic history of France. *Eur. J. Hum. Genet.* <https://doi.org/10.1038/s41431-020-0584-1> (2020).
- Bycroft, C. et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **10**, 551 (2019).
- Chacón-Duque, J.-C. et al. Latin Americans show wide-spread *Converso* ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).
- Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
- Novembre, J. & Peter, B. M. Recent advances in the study of fine-scale population structure in humans. *Curr. Opin. Genet. Dev.* **41**, 98–105 (2016).
- Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
- van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
- Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
- Sawcer, S. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
- Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

23. Lawson, D. J. & Falush, D. Population identification using genetic data. *Ann. Rev. Genom. Hum. Genet.* **13**, 337–361 (2012).
24. Palamara, P. F. Population genetics of identity by descent. Preprint at <https://arxiv.org/abs/1403.4987> (2014).
25. Al-Asadi, H., Petkova, D., Stephens, M. & Novembre, J. Estimating recent migration and population-size surfaces. *PLoS Genet.* **15**, e1007908 (2019).
26. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
27. Herlihy, D. *The Black Death and the Transformation of the West*. (Harvard University Press, 1997).
28. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinforma. Oxf. Engl.* **30**, 1266–1272 (2014).
29. Athanasiadis, G. et al. Nationwide genomic study in Denmark reveals remarkable population homogeneity. *Genetics* **204**, 711–722 (2016).
30. Nalls, M. A. et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet.* **5**, e1000415 (2009).
31. Roosen, J. & Curtis, D. R. The ‘light touch’ of the Black Death in the Southern Netherlands: an urban trick? *Econ. Hist. Rev.* **72**, 32–56 (2019).
32. Abdellaoui, A. et al. Association between autozygosity and major depression: stratification due to religious assortment. *Behav. Genet.* **43**, 455–467 (2013).
33. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
34. Abdellaoui, A. et al. Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* **3**, 1332–1342 (2019).
35. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
36. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Delaneau, O., Marchini, J. & Zagury, J.-F. cois. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
38. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
39. CoreTeam, R. R.: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; (2015).
40. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R. J.* **8**, 289–317 (2016).
41. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
42. Browning, B. L. & Browning, S. R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).
43. Felsenstein, J. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68**, 581–597 (1971).
44. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 135–139 (2008).
45. Nakazawa, M. *fmsb: functions for medical statistics book with some demographic data, 2014* (R Package, 2018).
46. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
47. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

## Acknowledgements

This work has been supported by Science Foundation Ireland (17/CDA/4737), the Motor Neurone Disease Association of England, Wales and Northern Ireland (957-799) and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 772376 – ESORIAL). The collaboration project is co-funded by the PPP Allowance made available by Health-Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. The authors wish to acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

## Author contributions

R.P.B. and R.L.M. conceived the study. R.P.B., W.V.R., J.H.V. and R.L.M. contributed to study design. R.P.B. and R.L.M. conducted the analyses. R.P.B. and R.L.M. drafted the manuscript. W.V.R., L.H.V.D.B. and J.H.V. provided data and critical revision of the manuscript.

## Competing interests

The authors declare no competing interests.

## Ethics

Sample collection and data sharing were approved by country-specific institutional review boards and informed consent was obtained from study participants as detailed in the source studies<sup>19,21</sup>.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-18418-4>.

**Correspondence** and requests for materials should be addressed to R.P.B. or R.L.M.

**Peer review information** *Nature Communications* thanks Abdel Abdellaoui and Javier Mendoza-Revilla for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020