A genome-wide haplotype association analysis of major depressive disorder identifies two genome-wide significant haplotypes

David M. Howard*[1], Lynsey S. Hall[1], Jonathan D. Hafferty[1], Yanni Zeng[1], Mark J. Adams[1], Toni-Kim Clarke[1], David J. Porteous[2], Caroline Hayward[3], Blair H. Smith[4], Alison D. Murray[5], Niamh M. Ryan[2], Kathryn L. Evans[2], Chris S. Haley[3], Ian J. Deary[6, 7], Pippa A. Thomson[2] and Andrew M. McIntosh[1, 7]

Affiliations:
[1]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
[2]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[3]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[4]Division of Population Health Sciences, University of Dundee, Dundee, UK
[5]Aberdeen Biomedical Imaging Centre, University of Aberdeen, Aberdeen, UK
[6]Department of Psychology, The University of Edinburgh, Edinburgh, UK
[7]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK

*Corresponding author: David M. Howard (e-mail: D.Howard@ed.ac.uk)

ABSTRACT

Genome-wide association studies using SNP genotype data have had limited success in the identification of variants associated with major depressive disorder (MDD). Haplotype data provide an alternative method for detecting associations between variants in weak linkage disequilibrium with genotyped variants and a given trait of interest. A genome-wide haplotype association study for MDD was undertaken utilising a family-based population cohort, Generation Scotland: Scottish Family Health Study, as a discovery sample with a population-based cohort, UK Biobank, used as a replication sample. Fine mapping of haplotype boundaries was used to account for overlapping haplotypes tagging causal variants. Within the discovery cohort, two haplotypes exceeded genome-wide significance ($P<5 \times 10^{-8}$) for an association for MDD. One of these haplotypes was located in 6q21, in a region which has been previously associated with bipolar disorder, a psychiatric disorder that is phenotypically and genetically correlated with MDD. The detection of associated haplotypes potentially allows the causal stratification of MDD into biologically informative aetiological subtypes.

INTRODUCTION

Major depressive disorder (MDD) is a complex and clinically heterogeneous condition with core symptoms of low mood and/or anhedonia over a period of two weeks. MDD is frequently comorbid with other clinical conditions, such as cardiovascular disease,[1] cancer[2] and inflammatory diseases.[3] This complexity and comorbidity suggests heterogeneity of aetiology and may explain why there has been limited success in identifying causal genetic variants,[4-6] despite heritability estimates ranging from 28% to 37%.[7, 8] Single nucleotide polymorphism (SNP) based analyses are unlikely to fully capture the variation in surrounding regions, especially untyped lower-frequency variants and those that are in weak linkage disequilibrium (LD) with the common SNPs on many genotyping arrays.

Haplotype-based analysis may help improve the detection of causal genetic variants as, unlike single SNP-based analysis, it is possible to assign both the strand and parent of origin of sequence variants and combine information from multiple SNPs to identify rarer causal variants. A number of studies[9-11] have identified haplotypes associated with MDD, albeit by focussing on particular regions of interest. In the current study, a family and population-based cohort Generation Scotland: Scottish Family Health Study (GS:SFHS) was utilised to ascertain genome-wide haplotypes in closely and distantly related individuals.[12] A haplotype-based association analysis was conducted using MDD as a phenotype, followed by additional fine-mapping of haplotype boundaries and a replication study performed using the UK Biobank cohort.[13]

MATERIALS AND METHOD

Discovery cohort

The discovery phase of the study used the family and population-based Generation Scotland: Scottish Family Health Study (GS:SFHS) cohort,[12] consisting of 23 960 individuals of which 20 195 were genotyped using the Illumina OmniExpress BeadChip (706 786 SNPs). Quality control procedures were applied, removing individuals with a genotype call rate < 98% and those SNPs with a call rate < 98%, a minor allele frequency (MAF) < 0.01 or those deviating from Hardy-Weinberg equilibrium ($P<10^{-6}$).

Following quality control there were 19 994 GS:SFHS individuals (11 774 females and 8 220 males) genotyped for 561 125 autosomal SNPs. These individuals ranged from 18-99 years of age with an average age of 47.4 years and a standard deviation of 15.0 years. Within the cohort were 4 933 families containing at least two related individuals, including 1 799 families with two members, 1 216 families with three members and 829 families with four members with the largest family containing 31 individuals. There were 1 789 individuals with no other family members in this dataset.

Genotype phasing and haplotype formation

Phasing of the genotype data was conducting using SHAPEIT v2.r837.[14] The relatedness within GS:SFHS made it suitable for the application of the duoHMM method, which combined the results of a MCMC algorithm with pedigree information to improve phasing accuracy.[15] A 5Mb window size was used (rather than the default of 2Mb) as this has been shown to be advantageous when larger amounts of identity by descent (IBD) sharing are present.[14] The number of conditioning states per SNP was increased from the default of 100 states to 200 states to improve phasing accuracy, with the default effective population size of 15 000 used. HapMap phase II b37[16] was used to calculate the recombination rates between SNPs during phasing and for the subsequent partitioning of phased data into haplotypes.

Window sizes of 1cM, 0.5cM and 0.25cM were used to determine the SNPs included within each haplotype.[17] A sliding window was used, sliding the window along a quarter of the respective window size. This produced a total of 97 333 windows with a mean number of SNPs per window of 157, 79 and 34 for the 1cM, 0.5cM and 0.25cM windows, respectively. The haplotype positions reported subsequently are given in base pair (bp) position (using GRCh37) and correspond to the outermost SNPs located within each haplotype. Quality control procedures were applied to the haplotypes in each window, with those haplotypes containing less than 5 SNPs, with a haplotype frequency < 0.005 or deviating from Hardy-Weinberg equilibrium ($P<10^{-6}$) removed. Following quality control there were 2 618 094 haplotypes for further analysis.

To estimate the correction required for multiple testing, the clump command within Plink v1.90[18] was used to determine the number of independently segregating haplotypes. An LD threshold of 0.4 was used to classify a haplotype as independent and at this threshold there were 1 070 216 independently segregating haplotypes in the discovery cohort. Therefore, when assessing all 2.6 million haplotypes a Bonferroni correction required $P<5 \times 10^{-8}$ for genome-wide significance which is in alignment with the conventional level for significance used for sequence and SNP-based genome-wide association studies.[19] Therefore in this study, and for future genome-wide haplotype-based analyses using cohorts similar to GS:SFHS, then the conventional *P*-value for significance can be applied.

Phenotype ascertainment and patient linkage

In the discovery cohort, a diagnosis of MDD was made using initial screening questions and the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders (SCID).[20] The SCID is an internationally validated approach to identifying episodes of depression and was conducted by clinical nurses trained in its administration. Further details regarding this diagnostic assessment are described previously.[21] In this study, MDD was defined by at least one instance of a major depressive episode which initially identified 2 659 cases, 17 237 controls and 98 missing (phenotype unknown) individuals.

In addition, the psychiatric history of cases and controls was examined using record linkage to the Scottish Morbidity Record.[22] Within the control group, 1 072 participants were found to have attended at least one psychiatry outpatient clinic and were excluded from the study. In addition, 47 of the MDD cases were found have additional diagnoses of either bipolar disorder or schizophrenia in psychiatric inpatient records and were also excluded from the study. These participants had given prior consent for anonymised record linkage to routine administrative clinical data. Individuals who were identified as population outliers through principal component analyses of their genotypes were also removed.[23]

In total there were 2 605 MDD cases and 16 168 controls following the removal of individuals based on patient records and population stratification, equating to a prevalence of 13.9% for MDD in this sample.

Statistical approach

A mixed linear model was used to conduct an association analysis using GCTA v1.25.0:[24]

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u} + \boldsymbol{Z}_2\boldsymbol{v} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ was the vector of binary observations for MDD. $\boldsymbol{\beta}$ was the matrix of fixed effects, including haplotype, sex, age and age^2. $\boldsymbol{u}$ was fitted as a random effect taking into account the genomic relationships (MVN $(0, \boldsymbol{G}\sigma_u^2)$, where $\boldsymbol{G}$ was a genomic relationship matrix[25]). $\boldsymbol{v}$ was a random effect fitting a second genomic relationship matrix $\boldsymbol{G}_t$ (MVN $(0, \boldsymbol{G}_t\sigma_v^2)$ which modelled only the more closely related individuals.[26] $\boldsymbol{G}_t$ was equal to $\boldsymbol{G}$ except that off-diagonal elements < 0.05 were set to 0. $\boldsymbol{X}$, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ were the corresponding incidence matrices. $\boldsymbol{\varepsilon}$ was the vector of residual effects and was assumed to be normally distributed, MVN $(0, \mathbf{I}\sigma_\varepsilon^2)$.

The inclusion of the second genomic relationship matrix, $\boldsymbol{G}_t$, was deemed desirable as the fitting of the single matrix $\boldsymbol{G}$ alone resulted in significant population stratification (intercept = 1.029 ± 0.003, λGC = 1.026) following examination with LD score regression.[27] The fitting of both genomic relationship matrices simultaneously produced no evidence of bias due to population stratification (intercept = 1.002 ± 0.003, λGC = 1.005).

Fine mapping

The method described above examines the effect of each haplotype against all other haplotypes in that window. To investigate whether there were causal variants located within directly overlapping haplotypes of the same window size, fine mapping of haplotype boundaries was used. Where there were directly overlapping haplotypes, each with $P<10^{-3}$ and with an effect in the same direction, i.e. both causal or both preventative, then any shared consecutive regions formed a new haplotype that was assessed using the mixed model described previously. This new haplotype was assessed using all

individuals and was required to be at least 5 SNPs in length. A total of 47 new haplotypes were assessed from within 26 pairs of directly overlapping haplotypes.

Replication cohort

UK Biobank[28] (provided as part of project #4844) was used as a replication cohort for the haplotypes in GS:SFHS with $P<10^{-6}$. The UK Biobank data consisted of 152 249 individuals with genomic data for 72 355 667 imputed variants.[29] The SNPs genotyped in GS:SFHS were extracted from the UK Biobank data, identifying 557 813 variants in common between the two datasets. Using the variants in common, phasing was conducted on a 50Mb window centred on the region containing the haplotype of interest using all individuals in UK Biobank. The same parameters for SHAPEIT v2.r837 were used for phasing as that described previously for GS:SFHS, except that duoHMM was not used and the default window size of 2Mb was applied, due to the expectation of lower IBD sharing in the UK Biobank cohort.

To produce an unrelated dataset suitable for replication, firstly those individuals listed as non-white British were removed. Secondly, participants in GS:SFHS and their relatives (up to and including 3rd degree relatives) were removed from within UK Biobank. Finally, up to and including 3rd degree relatives of the remaining UK Biobank participants were removed to create an unrelated dataset. The relatives to be removed were identified by a kinship coefficient ≥ 0.0442 with another participant in UK Biobank using the KING toolset,[30] leaving a total of 112 024 individuals. The phenotype in the replication dataset was based on a less restrictive self-diagnosed touchscreen assessment with case status defined as either 'probable single lifetime episode of major depression' or 'probable recurrent major depression (moderate and severe)' and with control status defined as 'no mood disorder'.[13] In total there were 8 003 cases, 15 540 controls and 88 481 with missing records, equating to a trait prevalence of 34.0% in this sample.

A logistic association analysis was implemented in Plink v1.90[18] assessing the haplotypes in UK Biobank which were identified in the discovery cohort with $P<10^{-6}$:

$$y = X\beta + \varepsilon$$

where y was the vector of binary observations for MDD. β was the matrix of fixed effects, including haplotype, sex, age, age^2, the first 8 genetic principal components, genotyping array, genotyping batch and recruitment centre. X was the corresponding incidence matrices and ε was the vector of residual effects and was assumed to be normally distributed, MVN (0, $\mathbf{I}\sigma_\varepsilon^2$). Replication success was judged on the statistical significance of each haplotype and a meta-analysis conducted using the R package meta[31] with the default methods used for the cohort pooling.

RESULTS

An association analysis for MDD was conducted using 2 618 094 haplotypes and 47 fine mapped haplotypes within the discovery cohort, GS:SFHS. A genome-wide Manhattan plot of $-\log_{10}$ *P*-values for these haplotypes is provided in Figure 1 with a q-q plot provided in Supplementary Figure 1. Within the discovery cohort, 13 haplotypes had $P<10^{-6}$ with replication sought for these haplotypes using UK Biobank. Summary statistics from both cohorts regarding these 13 haplotypes are provided in Table 1. There was no evidence for an association with MDD for any of these haplotypes in the replication cohort. Table 2 shows the protein coding genes which overlap these 13 haplotypes with the SNPs and alleles that constitute these haplotypes provided in Supplementary Table 1.

Within the discovery cohort, two haplotypes exceeded genome-wide significance ($P<5 \times 10^{-8}$) for an association with MDD, one located on chromosome 6 and the other located on chromosome 10.

On chromosome 6, two directly overlapping 0.5cM haplotypes consisting of 28 SNPs were identified between 108 335 345 and 108 454 437 bp (rs7749081 - rs212829). These two haplotypes had *P*-values of 3.24×10^{-5} and 5.57×10^{-5}, respectively and differed at a single SNP (rs7749081). Exclusion of this single SNP defined a new 27 SNP haplotype that had a genome-wide significant association with MDD ($P=7.06 \times 10^{-9}$). A Manhattan plot of the region surrounding the genome-wide significant haplotype on chromosome 6 is provided in Figure 2. This haplotype had an odds ratio (OR) of 1.87 (95% confidence interval (CI): 1.53 – 2.29) in the discovery dataset and an OR of 1.09 (95% CI: 0.95

– 1.25) in the replication dataset. The disease prevalence in the discovery dataset was 22.7% amongst carriers compared to 13.6% prevalence amongst non-carriers. Calculating the effect size at the population level,[32] the estimate of the contribution of this haplotype to the total genetic variance was 2.32×10^{-4}. None of the SNPs within this haplotype were individually associated with MDD in either cohort ($P \geq 0.05$).

A genome-wide significant haplotype ($P=8.50 \times 10^{-9}$) was identified on chromosome 10 using a 0.5cM window. A Manhattan plot of the region surrounding this haplotype is provided in Figure 3. This haplotype had an OR of 2.35 (95% CI: 1.76 - 3.14) in the discovery cohort, which was the highest OR observed in this study and a disease prevalence of 27.2% amongst carriers of this haplotype, compared to a prevalence of 13.7% amongst non-carriers. The replication cohort had an OR of 1.13 (95% CI: 0.79 - 1.63) for the haplotype on chromosome 10. The estimate of the contribution of this haplotype to the total genetic variance was 2.29×10^{-4}. Association analysis of the 92 SNPs on this haplotype revealed that one SNP in GS:SFHS (rs17133585) and two SNPs in UK Biobank (rs12413638 and rs10904290) were nominally significant ($P<0.05$), although none had $P<0.001$.

The SuRFR R package was used to integrate functional annotation and prior biological knowledge to prioritise potentially functional variants from within the associated regions.[33] The variant, rs313455 located within the haplotype on chromosome 10, was highlighted by SuRFR as a well conserved and interesting regulatory candidate, being located in a DNase HS region, and overlapping a binding site for the brain-expressed transcription factor FOS, although the variant itself was not significantly associated with MDD ($P \geq 0.05$).

All 13 of the haplotypes with a *P*-value for association $< 10^{-6}$ in the GS:SFHS discovery cohort were risk factors for MDD (OR > 1), see Supplementary Table 2. Within the replication cohort, 9 out of these 13 haplotypes were shown to be risk factors, however none of the 95% confidence intervals for the replication ORs overlapped the 95% confidence intervals of the discovery GS:SFHS cohort. All 13 haplotypes providing evidence of significant heterogeneity ($P<0.05$, I-squared[34] $\geq 88.2\%$) between

the two datasets. Meta-analysis of UK Biobank and GS:SFHS, gave no evidence for association of these haplotypes with MDD.

DISCUSSION

Haplotype-based analyses are capable of tagging variants due to the LD between the untyped variants and the multiple flanking genotyped variants which make up the inherited haplotype. This approach should provide greater power when there is comparatively higher IBD sharing. The cohort used in the discovery phase of the analysis was recruited as a population and family-based dataset and therefore should have greater power for detecting haplotypes associated with MDD compared to an unrelated dataset, which may partly explain why replication was unsuccessful. Thirteen haplotypes were identified in the discovery cohort with $P<10^{-6}$ of which two reached genome-wide significance ($P<5 \times 10^{-8}$), although none of these were significant in the replication cohort.

The prevalence of MDD observed in the discovery cohort (13.7%) was comparable to that reported for structured clinical diagnosis of MDD in other high income counties (14.6%).[35] However in the replication cohort, the trait prevalence was notably higher (34.0%), most likely due to the differing methods of phenotypic ascertainment. This difference potentially limits the power of the replication cohort to validate the findings from the discovery cohort. A complementary approach to replication is to identify gene coding regions within haplotypes that provided a biologically informative explanation for an association with MDD.

The genome-wide significant haplotype on chromosome 6 overlaps with the Osteopetrosis Associated Transmembrane Protein 1 (OSTM1) coding gene. OSTM1 is associated with neurodegeneration[36, 37] and melanocyte function,[38] with alpha-melanocyte stimulating hormone known to be associated with depression.[39-41] This haplotype lies within the 6q21 region which has been associated with bipolar disorder,[42-45] which shares similar symptoms with MDD and which has a correlated liability of 0.64 with MDD.[46] This potentially suggests either a pleiotropic effect or clinical heterogeneity, whereby patients may be misdiagnosed, i.e. patients may have MDD and transition to bipolar disorder in the future or are sub-threshold for bipolar disorder and instead given a diagnosis of MDD.

The haplotype identified on chromosome 7 was within 10kb of the variant rs10260531 which had moderate association (P=3.21 x 10^{-5}) with MDD in a large mega-analysis.[4] Both the variant and the haplotype are located within the Thromboxane A synthase 1 (TBXAS1) gene which is involved in platelet aggregation and maintaining haemostasis[47, 48] and is a mediator in cardiovascular disease.[49] Multiple studies have demonstrated comorbidity between cardiovascular disease and MDD[50, 51] and this region certainly warrants further investigation.

The two haplotypes on chromosome 8 with $P<10^{-6}$ both overlapped with the Interleukin 7 (IL7) protein coding region. IL7 is involved in maintaining T cell homeostasis[52] and proliferation,[53] which in turn contributes to the immune response to pathogens. It has been proposed that impaired T cell function may be a factor in the development of MDD,[54] with depressed subjects found to have elevated[55] or depressed levels[56] of IL7 serum. There is conjecture as to whether MDD causes inflammation or represents a reaction to an increased inflammatory response,[57, 58] but it is most likely to be a bidirectional relationship.[56] In the present study, an association between a region containing the IL7 gene and MDD suggests that, at least for a subset of cases, it was immunocompromised individuals which developed MDD, rather than MDD causing an altered immune response.

The haplotype on chromosome 12 located between 48 159 721 – 48 263 828 bp overlapped with four gene coding regions. One of these encodes the vitamin D receptor (VDR) for which there is a suggestive association with MDD,[59-61] however this association appears to be population dependant.[62] The discovery cohort was based in Scotland which is at a latitude of approximately 55° to 60°N and is a country known for its rather cloudy and inclement climate. Studies have demonstrated that vitamin D deficiencies are highly prevalent in Scotland[63, 64] and therefore this cohort may be suitable for further studies examining the relationships between the VDR coding region, vitamin D levels and MDD.

Two Dutch studies[65, 66] have identified a variant (rs8023445) on chromosome 15 located within the SRC (Src homology 2 domain containing) family, member 4 (SHC4) gene coding region with a moderate degree of association with MDD (P=1.64 x 10^{-5} and P=9 x 10^{-6}, respectively). A variant

(rs10519201) within the SHC4 coding region was also found to have an association with Obsessive-Compulsive Personality Disorder in a UK-based study (P=6.16 x 10^{-6}).[67] SHC4 is expressed in neurons[68] and regulates BDNF-induced MAPK activation[69] which has been shown to be a key factor in MDD pathophysiology.[70] The SHC4 region overlaps with the haplotype on chromosome 15 identified in the discovery cohort (49 206 902 – 49 260 601 bp position with P=9.21 x 10^{-8}) and therefore further research should be undertaken examining the association between the SHC4 region and psychiatric disorders.

As haplotype windows increase in size the individual haplotypes become rarer within the population and consequently have lower power to detect an association as they are carried by fewer individuals.[71] This was apparent here from comparatively fewer 1cM haplotypes remaining after restricting the haplotype frequency to > 0.005 and with no 1cM haplotypes having $P<10^{-6}$. As different sizes of haplotypes across a region are likely to tag different untyped variants, it was potentially beneficial to analyse multiple window sizes. Two-point non-parametric linkage analysis was conducted on the best 13 haplotypes, coding the haplotypes as a single marker, using Merlin v1.1.2.[72] No evidence was found significant linkage with MDD in GS:SFHS, suggesting that the associated haplotypes were not segregating within multiple large families. This was unsurprising due to the window sizes used to define the haplotype blocks in this study.

Conclusions and future research

The study identified two haplotypes within the discovery that exceeded genome-wide significance for an association with a clinically diagnosed MDD phenotype. The haplotype on chromosome 6 was identified by applying a fine mapping technique that amended the haplotype boundaries and strengthened the signal detected. This haplotype was located on 6q21 which has been shown in previous studies to be related to other psychiatric disorders. Additional work could seek to replicate the findings in additional cohorts, as well as a wider analysis of all haplotypes in the regions identified within replication cohorts. An additive model was used to analyse the haplotypes and alternative approaches could implement a dominant model or an analysis of diplotypes (haplotype pairs) for an

association with MDD. There were a number of haplotypes approaching genome-wide significance located within gene coding regions associated with diseases that share a comorbidity with MDD and therefore these regions warrant further investigation. The total genetic variance explained by the haplotypes identified was small, however these haplotypes potentially represent biologically informative aetiological subtypes for MDD and merit additional analysis using alternative biological approaches or next generation sequencing.

ACKNOWLEDGMENTS

Generation Scotland received core funding from the Chief Scientist Office of the Scottish Government Health Directorate CZD/16/6 and the Scottish Funding Council HR03006. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the UK's Medical Research Council and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) (Reference 104036/Z/14/Z). YZ acknowledges support from China Scholarship Council. IJD is supported by the Centre for Cognitive Ageing and Cognitive Epidemiology which is funded by the Medical Research Council and the Biotechnology and Biological Sciences Research Council (MR/K026992/1). AMMcI acknowledges support from the Dr Mortimer and Theresa Sackler Foundation.

We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. Ethics approval for the study was given by the NHS Tayside committee on research ethics (reference 05/S1401/8)

CONFLICT OF INTEREST

The authors declare that no conflict of interest exists

REFERENCES

1. Huffman JC, Celano CM, Beach SR, Motiwala SR, Januzzi JL. Depression and cardiac disease: epidemiology, mechanisms, and diagnosis. *Cardiovascular Psychiatry and Neurology* 2013; **2013:** 14.

2. Kang H-J, Kim S-Y, Bae K-Y, Kim S-W, Shin I-S, Yoon J-S *et al.* Comorbidity of depression with physical disorders: research and clinical implications. *Chonnam Medical Journal* 2015; **51**(1)**:** 8-18.

3. Raison CL, Capuron L, Miller AH. Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends in Immunology* 2006; **27**(1)**:** 24-31.

4. Major Depressive Disorder Working Group of the Psychiatric Gwas Consortium. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* 2013; **18**(4)**:** 497-511.

5. Converge Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 2015; **523**(7562)**:** 588-591.

6. Levinson DF, Mostafavi S, Milaneschi Y, Rivera M, Ripke S, Wray NR *et al.* Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it? *Biological Psychiatry* 2014; **76**(7)**:** 510-512.

7. Lubke GH, Hottenga JJ, Walters R, Laurin C, de Geus EJC, Willemsen G *et al.* Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biological Psychiatry* 2012; **72**(8)**:** 707-709.

8. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry* 2000; **157**(10)**:** 1552-1562.

9. Zhang Z, Ni J, Zhang J, Tang W, Li X, Wu Z *et al.* A haplotype in the 5'-upstream region of the NDUFV2 gene is associated with major depressive disorder in Han Chinese. *Journal of Affective Disorders* 2016; **190:** 329-332.

10. Kim J-J, Mandelli L, Pae C-U, De Ronchi D, Jun T-Y, Lee C *et al.* Is there protective haplotype of dysbindin gene (DTNBP1) 3 polymorphisms for major depressive disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2008; **32**(2)**:** 375-379.

11. Klok MD, Giltay EJ, Van der Does AJW, Geleijnse JM, Antypa N, Penninx BWJH *et al.* A common and functional mineralocorticoid receptor haplotype enhances optimism and protects against depression in females. *Translational Psychiatry* 2011; **1:** e62.

12. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM *et al.* Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants

and their potential for genetic research on health and illness. *International Journal of Epidemiology* 2013; **42**(3)**:** 689-700.

13. Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans J *et al.* Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: cross-sectional study of 172,751 participants. *PLoS ONE* 2013; **8**(11)**:** e75362.

14. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 2013; **10**(1)**:** 5-6.

15. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* 2014; **10**(4)**:** e1004234.

16. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**(7164)**:** 851-861.

17. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 2013; **194**(2)**:** 459-471.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira Manuel A, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 2007; **81**(3)**:** 559-575.

19. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014; **15**(5)**:** 335-346.

20. First MB, Spitzer RL, Gibbon Miriam., Williams JBW. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P) 2002.

21. Fernandez-Pujals AM, Adams MJ, Thomson P, McKechanie AG, Blackwood DHR, Smith BH *et al.* Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation scotland: scottish family health study (GS:SFHS). *PLoS ONE* 2015; **10**(11)**:** e0142197.

22. Information Services Division. SMR Data Manual. http://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets, 2016.

23. Amador C, Huffman J, Trochet H, Campbell A, Porteous D, Wilson JF *et al.* Recent genomic heritage in Scotland. *BMC Genomics* 2015; **16**(1)**:** 1-17.

24. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* 2014; **46**(2)**:** 100-106.

25. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 2010; **42**(7)**:** 565-569.

26. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics* 2013; **9**(5)**:** e1003520.

27. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 2015; **47**(3)**:** 291-295.

28. Allen NE, Sudlow C, Peakman T, Collins R. UK biobank data: come and get it. *Science Translational Medicine* 2014; **6**(224)**:** 224ed4.

29. Marchini J. UK Biobank phasing and imputation documentation. Version 1.2. http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf, 2015.

30. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**(22)**:** 2867-2873.

31. Schwarzer G. meta: general package for meta-analysis. R package version 4.3-2. 2015.

32. Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* 2010; **42**(7)**:** 570-575.

33. Ryan NM, Morris SW, Porteous DJ, Taylor MS, Evans KL. SuRFing the genomics wave: an R package for prioritising SNPs by functionality. *Genome Medicine* 2014; **6**(10)**:** 1-13.

34. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11)**:** 1539-1558.

35. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G *et al.* Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine* 2011; **9**(1)**:** 1-16.

36. Kasper D, Planells-Cases R, Fuhrmann JC, Scheel O, Zeitz O, Ruether K *et al.* Loss of the chloride channel ClC-7 leads to lysosomal storage disease and neurodegeneration. *The EMBO Journal* 2005; **24**(5)**:** 1079-1091.

37. Pandruvada SNM, Beauregard J, Benjannet S, Pata M, Lazure C, Seidah NG *et al.* Role of ostm1 cytosolic complex with kinesin 5B in intracellular dispersion and trafficking. *Molecular and Cellular Biology* 2016; **36**(3)**:** 507-521.

38. Hoek KS, Schlegel NC, Eichhoff OM, Widmer DS, Praetorius C, Einarsson SO *et al.* Novel MITF targets identified using a two-step DNA microarray strategy. *Pigment Cell & Melanoma Research* 2008; **21**(6)**:** 665-676.

39. Maes M, DeJonckheere C, Vandervorst C, Schotte C, Cosyns P, Raus J *et al.* Abnormal pituitary function during melancholia: Reduced α-melanocyte-stimulating hormone secretion and increased intact ACTH non-suppression. *Journal of Affective Disorders* 1991; **22**(3)**:** 149-157.

40. Goyal SN, Kokare DM, Chopde CT, Subhedar NK. Alpha-melanocyte stimulating hormone antagonizes antidepressant-like effect of neuropeptide Y in Porsolt's test in rats. *Pharmacology Biochemistry and Behavior* 2006; **85**(2)**:** 369-377.

41. Kokare DM, Singru PS, Dandekar MP, Chopde CT, Subhedar NK. Involvement of alpha-melanocyte stimulating hormone (α-MSH) in differential ethanol exposure and withdrawal related depression in rat: Neuroanatomical–behavioral correlates. *Brain Research* 2008; **1216:** 53-67.

42. Knight J, Rochberg NS, Saccone SF, Nurnberger JI, Rice JP. An investigation of candidate regions for association with bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2010; **153B**(7)**:** 1292-1297.

43. Dick DM, Foroud T, Flury L, Bowman ES, Miller MJ, Rau NL *et al.* Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the national institute of mental health genetics initiative. *American Journal of Human Genetics* 2003; **73**(1)**:** 107-114.

44. Park N, Juo SH, Cheng R, Liu J, Loth JE, Lilliston B *et al.* Linkage analysis of psychosis in bipolar pedigrees suggests novel putative loci for bipolar disorder and shared susceptibility with schizophrenia. *Molecular Psychiatry* 2004; **9**(12)**:** 1091-1099.

45. Pato CN, Pato MT, Kirby A, Petryshen TL, Medeiros H, Carvalho C *et al.* Genome-wide scan in Portuguese Island families implicates multiple loci in bipolar disorder: Fine mapping adds support on chromosomes 6 and 11. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 2004; **127B**(1)**:** 30-34.

46. McGuffin P, Rijsdijk F, Andrew M, Sham P, Katz R, Cardno A. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Archives of General Psychiatry* 2003; **60**(5)**:** 497-502.

47. Needleman P, Truk J, Jakschik BA, A R Morrison a, Lefkowith JB. Arachidonic acid metabolism. *Annual Review of Biochemistry* 1986; **55**(1)**:** 69-102.

48. Gresele P, Arnout J, Deckmyn H, Huybrechts E, Pieters G, Vermylen J. Role of proaggregatory and antiaggregatory prostaglandins in hemostasis. Studies with combined thromboxane synthase inhibition and thromboxane receptor antagonism. *Journal of Clinical Investigation* 1987; **80**(5)**:** 1435-1445.

49. Smyth EM. Thromboxane and the thromboxane receptor in cardiovascular disease. *Clinical lipidology* 2010; **5**(2)**:** 209-219.

50. Nicholson A, Kuper H, Hemingway H. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6 362 events among 146 538 participants in 54 observational studies. *European Heart Journal* 2006; **27**(23)**:** 2763-2774.

51. Hare DL, Toukhsati SR, Johansson P, Jaarsma T. Depression and cardiovascular disease: a clinical review. *European Heart Journal* 2013**:** 1365-1372.

52. Surh CD, Sprent J. Homeostasis of Naive and Memory T Cells. *Immunity* 2008; **29**(6)**:** 848-862.

53. Kittipatarin C, Khaled AR. Interlinking interleukin-7. *Cytokine* 2007; **39**(1)**:** 75-83.

54. Miller AH. Depression and immunity: A role for T cells? *Brain, Behavior, and Immunity* 2010; **24**(1)**:** 1-8.

55. Simon NM, McNamara K, Chow CW, Maser RS, Papakostas GI, Pollack MH *et al.* A detailed examination of cytokine abnormalities in major depressive disorder. *European Neuropsychopharmacology* 2008; **18**(3)**:** 230-233.

56. Lehto SM, Huotari A, Niskanen L, Herzig K-H, Tolmunen T, Viinamäki H *et al.* Serum IL-7 and G-CSF in major depressive disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2010; **34**(6)**:** 846-851.

57. Stewart JC, Rand KL, Muldoon MF, Kamarck TW. A prospective evaluation of the directionality of the depression-inflammation relationship. *Brain, Behavior, and Immunity* 2009; **23**(7)**:** 936-944.

58. Irwin MR, Miller AH. Depressive disorders and immunity: 20 years of progress and discovery. *Brain, Behavior, and Immunity* 2007; **21**(4)**:** 374-383.

59. Kuningas M, Mooijaart SP, Jolles J, Slagboom PE, Westendorp RGJ, van Heemst D. VDR gene variants associate with cognitive function and depressive symptoms in old age. *Neurobiology of Aging* 2009; **30**(3)**:** 466-473.

60. Ganji V, Milone C, Cody MM, McCarty F, Wang YT. Serum vitamin D concentrations are related to depression in young adult US population: the third national health and nutrition examination survey. *International Archives of Medicine* 2010; **3:** 29.

61. Glocke M, Lang F, Schaeffeler E, Lang T, Schwab M, Lang UE. Impact of vitamin D receptor VDR rs2228570 polymorphism in oldest old. *Kidney and Blood Pressure Research* 2013; **37**(4-5)**:** 311-322.

62. Bertone-Johnson ER. Vitamin D and the occurrence of depression: causal association or circumstantial evidence? *Nutrition Reviews* 2009; **67**(8)**:** 481-492.

63. Zgaga L, Theodoratou E, Farrington SM, Agakov F, Tenesa A, Walker M *et al.* Diet, environmental factors, and lifestyle underlie the high prevalence of vitamin D deficiency in healthy adults in Scotland, and supplementation reduces the proportion that are severely deficient. *The Journal of Nutrition* 2011; **141**(8)**:** 1535-1542.

64. Hyppönen E, Power C. Hypovitaminosis D in British adults at age 45 y: nationwide cohort study of dietary and lifestyle predictors. *American Journal of Clinical Nutrition* 2007; **85**(3)**:** 860-868.

65. Aragam N, Wang K-S, Pan Y. Genome-wide association analysis of gender differences in major depressive disorder in the Netherlands NESDA and NTR population-based samples. *Journal of Affective Disorders* 2011; **133**(3)**:** 516-521.

66. Sullivan PF, de Geus EJC, Willemsen G, James MR, Smit JH, Zandbelt T *et al.* Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Molecular Psychiatry* 2008; **14**(4)**:** 359-375.

67. Boraska V, Davis OSP, Cherkas LF, Helder SG, Harris J, Krug I *et al.* Genome-wide association analysis of eating disorder-related symptoms, behaviors, and personality traits. *American Journal of Medical Genetics* 2012; **159B**(7)**:** 803-811.

68. Hawley SP, Wills MKB, Rabalski AJ, Bendall AJ, Jones N. Expression patterns of ShcD and Shc family adaptor proteins during mouse embryonic development. *Developmental Dynamics* 2011; **240**(1)**:** 221-231.

69. You Y, Li W, Gong Y, Yin B, Qiang B, Yuan J *et al.* ShcD interacts with TrkB via its PTB and SH2 domains and regulates BDNF-induced MAPK activation. *BMB Rep* 2010; **43**(7)**:** 485-490.

70. Duric V, Banasr M, Licznerski P, Schmidt HD, Stockmeier CA, Simen AA *et al.* A negative regulator of MAP kinase causes depressive behavior. *Nature Medicine* 2010; **16**(11)**:** 1328-1332.

71. Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics* 2014; **95**(1)**:** 5-23.

72. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 2002; **30**(1)**:** 97-101.

Figure 1. Manhattan plot representing the $-\log_{10}$ *P*-values for an association between each assessed haplotype in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder

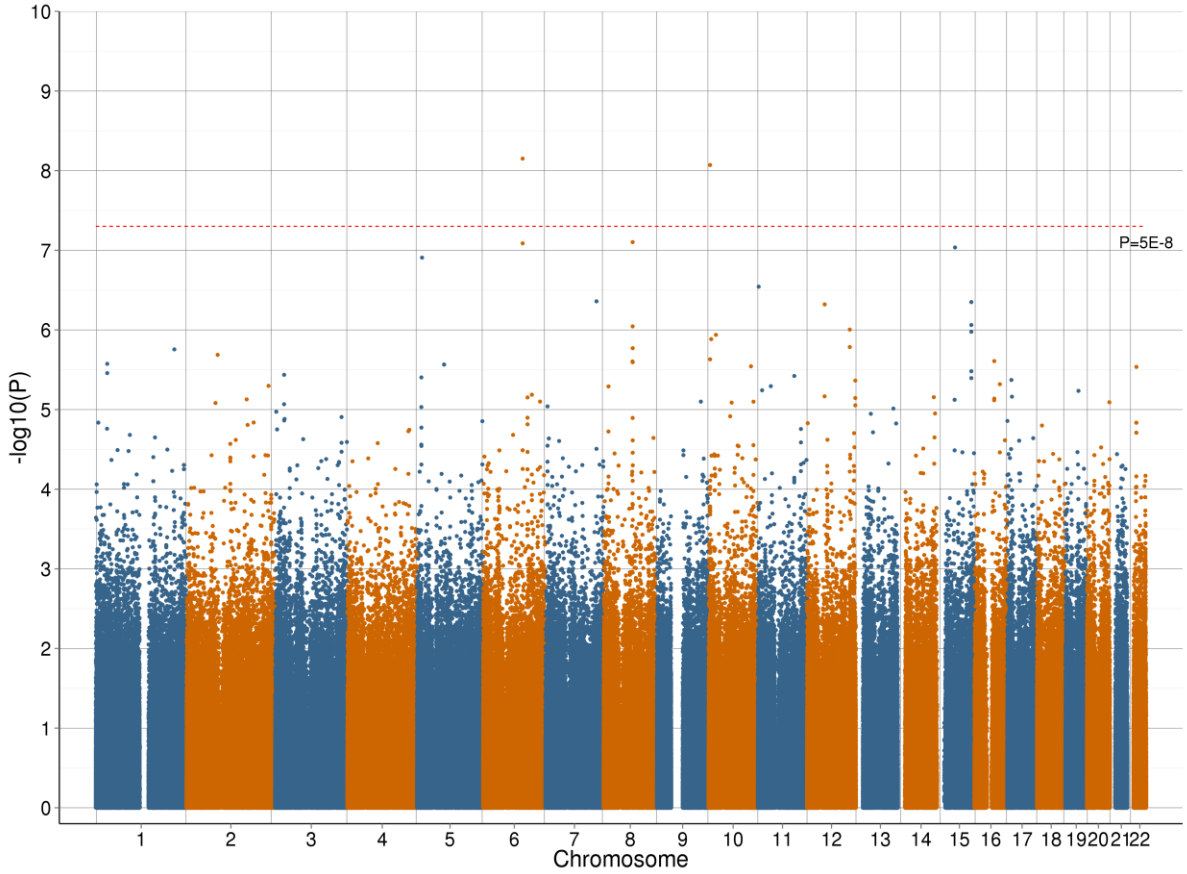

Figure 2. Manhattan plot representing the $-\log_{10}$ *P*-values for an association between haplotypes in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder within the 107.4 – 107.6 Mb region on chromosome 6. The start and end position (using build GRCh37) of haplotypes represent the outermost SNP positions within the windows examined. The warmth of colour represents the r^2 with the genome-wide significant haplotype located between 108 338 267 and 108 454 437 bp.

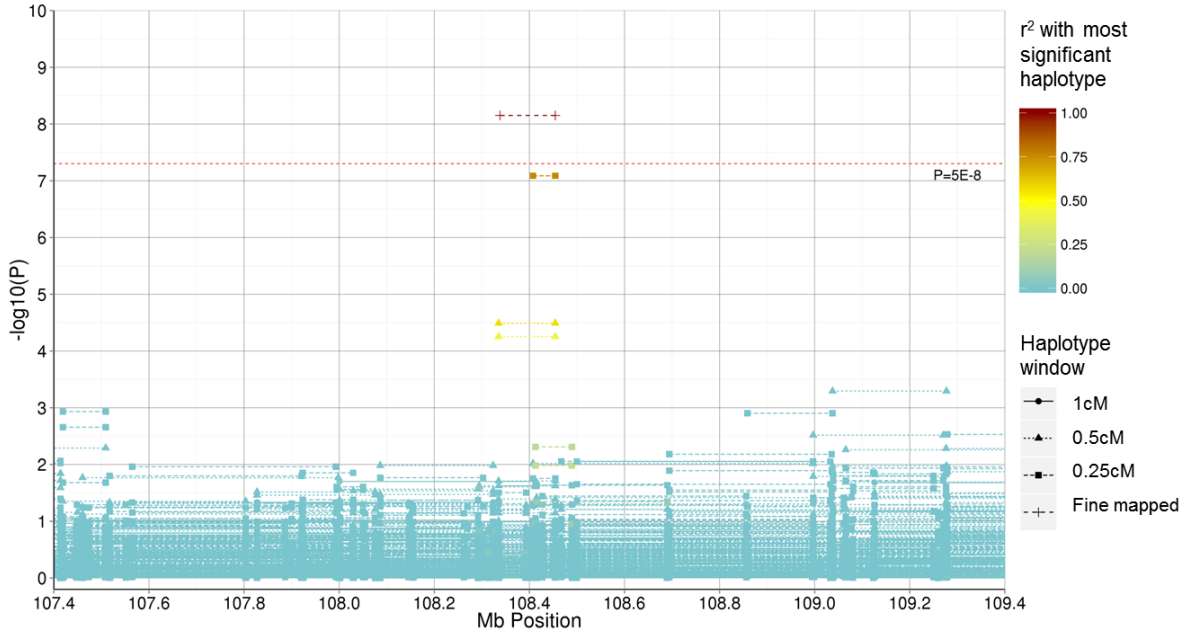

Figure 3. Manhattan plot representing the $-\log_{10}$ *P*-values for an association between haplotypes in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder within the 3.6 – 5.8 Mb region on chromosome 10. The start and end position (using build GRCh37) of haplotypes represent the outermost SNP positions within the windows examined. The warmth of colour represents the r^2 with the genome-wide significant haplotype located between 4 588 261 and 4 822 210 bp.

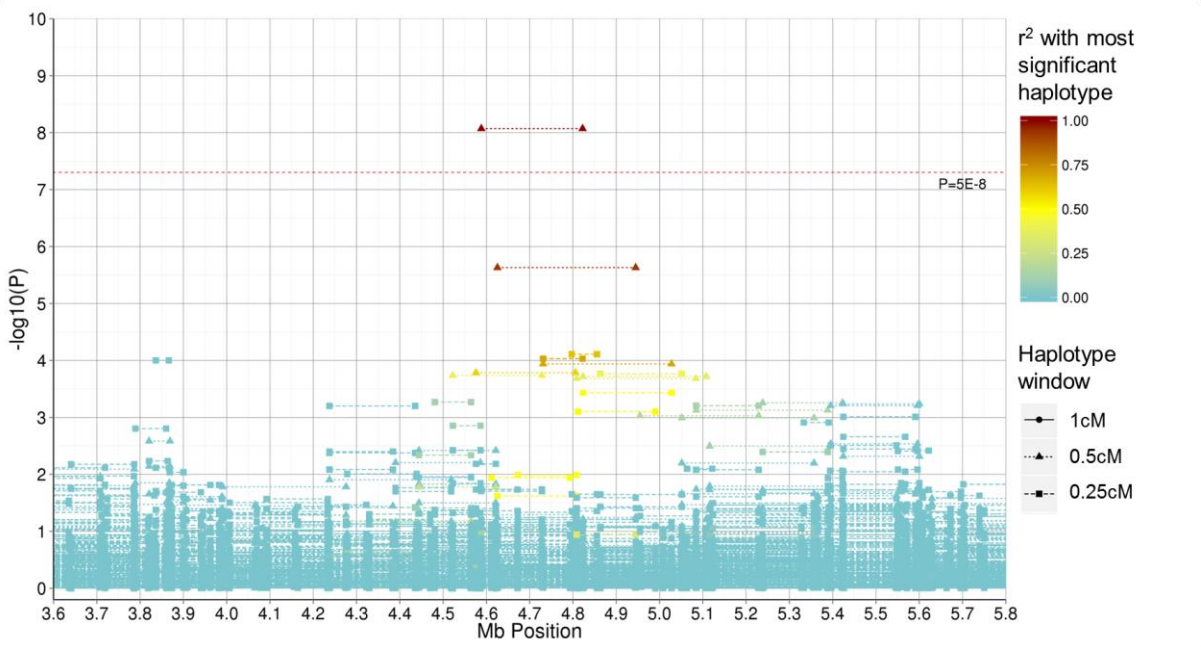

Table 1. The genetic association between Major Depressive Disorder and 13 haplotypes in the Generation Scotland: Scottish Family Health Study (GS:SFHS) discovery cohort (where $P<10^{-6}$) and a replication dataset (UK Biobank). Base pair (bp) positions are based on build GRCh37. Bold values indicate genome-wide statistical significance ($P<5 \times 10^{-8}$) was achieved in the GS:SFHS cohort. * indicates haplotype boundaries defined by the fine mapping approach. { indicates linkage disequilibrium (r^2) > 0.5 between haplotypes in the GS:SFHS cohort. Haplotype frequencies were calculated using unrelated individuals and excluding UK Biobank participants recruited in Glasgow or Edinburgh

	Haplotype					GS:SFHS		UK Biobank	
	Chr.	Position (bp)	RSID range	Window Size (cM)	SNPs in window	Haplotype Frequency	*P*	Haplotype Frequency	*P*
	5	13611631 - 13737444	rs408619 – rs17263496	0.25	24	0.0045	1.24×10^{-7}	0.0036	6.98×10^{-1}
{	**6 ***	**108338267 - 108454437**	**rs17069173 – rs218289**	**0.34**	**27**	**0.0152**	$\mathbf{7.06 \times 10^{-9}}$	0.0198	1.88×10^{-1}
{	6	108407662 - 108454437	rs2001144 – rs218289	0.25	12	0.0193	8.17×10^{-8}	0.0241	7.70×10^{-2}
	7	139682412 - 139708901	rs8192846 – rs12703488	0.25	27	0.0066	4.37×10^{-7}	0.0070	3.79×10^{-1}
{	8	79700362 - 80387861	rs11990466 – rs11777412	0.5	125	0.0076	9.02×10^{-7}	0.0080	7.21×10^{-1}
{	8	79759499 - 80156474	rs2010128 – rs1227634	0.25	68	0.0147	7.90×10^{-8}	0.0158	5.69×10^{-1}
	10	**4588261 - 4822210**	**rs11814411 – rs7082636**	**0.5**	**92**	**0.0064**	$\mathbf{8.50 \times 10^{-9}}$	0.0027	3.44×10^{-1}
	11 *	2260854 - 2437425	rs11021859 – rs2301698	0.41	42	0.0196	2.86×10^{-7}	0.0188	8.51×10^{-1}
	12	48159721 - 48263828	rs7135791 – rs2189480	0.25	26	0.0078	4.78×10^{-7}	0.0089	9.26×10^{-1}
	12	116904503 - 117062860	rs16946695 – rs10850685	0.25	45	0.0057	9.90×10^{-7}	0.0046	8.96×10^{-1}
	15	49206902 - 49260601	rs934741 – rs4474633	0.25	28	0.0082	9.21×10^{-8}	0.0080	3.93×10^{-1}
{	15	93806477 - 93851224	rs13313429 – rs13329166	0.5	24	0.0224	4.47×10^{-7}	0.0204	3.18×10^{-1}
{	15	93821340 - 93845622	rs4777809 – rs10083565	0.25	14	0.0265	8.67×10^{-7}	0.0241	1.96×10^{-1}

Table 2. Protein coding genes located overlapping with the 13 haplotypes with $P<10^{-6}$ in the Generation Scotland: Scottish Family Health Study (GS:SFHS) discovery cohort. Base pair (bp) positions are based on build GRCh37 with protein coding regions obtained from Ensembl, GRCh37.p13. { indicates a linkage disequilibrium (r^2) > 0.5 between haplotypes in the GS:SFHS cohort

	Chr.	Position (bp)	Protein coding genes
	5	13611631 - 13737444	DNAH5
{	6	108338267 - 108454437	OSTM1
{	6	108407662 - 108454437	OSTM1
	7	139682412 - 139708901	TBXAS1
{	8	79700362 - 80387861	IL7
{	8	79759499 - 80156474	IL7
	10	4588261 - 4822210	
	11	2260854 - 2437425	ASCL2, CLorf21, TSPAN32, CD81, TSSC4, TRPM5
	12	48159721 - 48263828	SLC48A1, RAPGEF3, HDAC7, VDR
	12	116904503 - 117062860	MAP1LC3B2
	15	49206902 - 49260601	SHC4
{	15	93806477 - 93851224	
{	15	93821340 - 93845622	