Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank

Howard, David M.; Hall, Lynsey S.; Hafferty, Jonathan D.; Yanni, Zeng; Adams, Mark J.; Clarke, Toni-Kim; Porteous, David J.; Nagy, Reka; Hayward, Caroline; Smith, Blair; Murray, Alison D.; Ryan, Niamh M.; Evans, Kathryn L.; Haley, Chris S.; Deary, Ian J.; Thomson, Pippa A.; McIntosh, Andrew M.

Published in:
Translational Psychiatry

DOI:
10.1038/s41398-017-0010-9

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Howard *et al. Translational Psychiatry* (2017)7:1263
DOI 10.1038/s41398-017-0010-9

Translational Psychiatry

ARTICLE Open Access

Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank

David M. Howard[1], Lynsey S. Hall[1], Jonathan D. Hafferty[1], Yanni Zeng[1,2], Mark J. Adams[1], Toni-Kim Clarke[1], David J. Porteous[3], Reka Nagy[2], Caroline Hayward[2,4], Blair H. Smith[4,5], Alison D. Murray[4,6], Niamh M. Ryan[3], Kathryn L. Evans[3,7], Chris S. Haley[2], Ian J. Deary[4,7,8], Pippa A. Thomson[3,7] and Andrew M. McIntosh[1,4,7]

Abstract

Genome-wide association studies using 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 ($n = 18,773$), as a discovery cohort with UK Biobank used as a population-based replication cohort ($n = 25,035$). Fine mapping of haplotype boundaries was used to account for overlapping haplotypes potentially tagging the same causal variant. Within the discovery cohort, two haplotypes exceeded genome-wide significance ($P < 5 \times 10^{-8}$) for an association with MDD. One of these haplotypes was nominally significant in the replication cohort ($P < 0.05$) and was located in 6q21, a region which has been previously associated with bipolar disorder, a psychiatric disorder that is phenotypically and genetically correlated with MDD. Several haplotypes with $P < 10^{-7}$ in the discovery cohort were located within gene coding regions associated with diseases that are comorbid with MDD. Using such haplotypes to highlight regions for sequencing may lead to the identification of the underlying causal variants.

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 at least 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[4–7], despite heritability estimates ranging from 28 to 37%[8,9]. Single-nucleotide polymorphism (SNP)-based analyses are unlikely to fully capture the variation in regions surrounding the genotyped markers, including 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 the strand of sequence variants and combine information from multiple SNPs to identify rarer causal variants. A number of studies[10–12] 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[13]. A haplotype-based association analysis was conducted using MDD as a

Correspondence: David M Howard (D.Howard@ed.ac.uk)
[1]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
[2]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
Full list of author information is available at the end of the article

phenotype, followed by additional fine mapping of haplotype boundaries with a replication and meta-analysis performed using the UK Biobank cohort[14].

Materials and methods

Discovery cohort

The discovery phase of the study used the family and population-based Generation Scotland: Scottish Family Health Study (GS:SFHS) cohort[13], consisting of 23,960 individuals of whom 20,195 were genotyped with the Illumina OmniExpress BeadChip (706,786 SNPs). Individuals with a genotype call rate <98% were removed, as well as those SNPs with a call rate <98%, a minor allele frequency (MAF) < 0.01 or those deviating from Hardy–Weinberg equilibrium ($P < 10^{-6}$). Individuals who were identified as population outliers through principal component analyses of their genotypic information were also removed[15].

Following quality control there were 19,904 GS:SFHS individuals (11,731 females and 8173 males) that had genotypic information 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. There were 4933 families that had at least two related individuals, this included 1799 families with two members, 1216 families with three members and 829 families with four members. The largest family group consisted of 31 related individuals and there were 1789 individuals that had no other family members within GS:SFHS.

Replication cohort

The population-based UK Biobank[16] (provided as part of project #4844) was used as a replication cohort to assess those haplotypes within GS:SFHS with $P < 10^{-6}$. The UK Biobank data consisted of 152,249 individuals with genomic data for 72,355,667 imputed variants[17]. The SNPs genotyped in GS:SFHS were extracted from the UK Biobank data and those variants with an imputation accuracy <0.8 were removed, leaving 555,782 variants in common between the two cohorts. Those genotyped individuals listed as non-white British and those that had also participated in GS:SFHS were removed from within UK Biobank, leaving a total of 119,955 individuals.

Genotype phasing and haplotype formation

The genotype data for GS:SFHS and UK Biobank was phased using SHAPEIT v2.r837[18]. Genome-wide phasing was conducted on the GS:SFHS cohort, while the phasing of UK Biobank was conducted on a 50 Mb window centred on those haplotypes identified within GS:SFHS with $P < 10^{-6}$. The relatedness within GS:SFHS made it suitable for the application of the duoHMM method, which improves phasing accuracy by also incorporating family information[19]. The default window size of 2 Mb was used for UK Biobank and a 5 Mb window was used for GS:SFHS as larger window sizes have been demonstrated to be beneficial when there is increased identity by descent (IBD) in the population[18]. 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. To calculate the recombination rates between SNPs during phasing the HapMap phase II b37[20] was used. This build was also used to partition the phased data into haplotypes.

Three window sizes (1cM, 0.5cM and 0.25cM) were used to establish the SNPs that formed each haplotype[21]. Each window was then moved along the genome by a quarter of the respective window size. There were a total of 97,333 windows with a mean number of SNPs per window of 157, 79 and 34 for the 1, 0.5 and 0.25cM windows, respectively. Windows that were less five SNPs in length were removed. The frequency (p) of each observed haplotype (A) was calculated as:

$$p = \frac{2\,X\,obs(AA) + obs(Aa)}{2\,X(obs(AA) + obs(Aa) + obs(aa))}$$

where a represents all other haplotypes in that window. A chi-squared test for Hardy–Weinberg equilibrium (X^2) for each haplotype was calculated as:

$$\mathrm{X}^2 = \frac{obs(AA) - p^2n}{p^2n} + \frac{obs(Aa) - 2\,pqn}{2\,pqn} + \frac{obs(aa) - q^2n}{q^2n}$$

where n is the number of individuals and $q = 1 - p$. Haplotypes with $0.995 < p < 0.005$ or with $X^2 > 24$ ($P < 10^{-6}$) were not tested for association, however, they were included within the alternative haplotype. Following this quality control there were a total of 2,618,094 haplotypes remaining for analysis. The reported haplotype positions relate to the outermost SNPs within each haplotype are in base pair (bp) position according to GRCh37.

To approximate the number of independently segregating haplotypes the clump command within Plink v1.90[22] was applied. This provides an estimation of the Bonferroni correction required for multiple testing. When applying an LD r^2 threshold of <0.4 there were 1,070,216 independently segregating haplotypes within GS:SFHS, equating to a P-value $< 5 \times 10^{-8}$ for genome-wide significance. This threshold is also frequently applied to SNP-based and sequence-based association studies to account for multiple testing[23].

Phenotype ascertainment

Discovery cohort

Within GS:SFHS 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)[24]. 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 have been described previously[25]. In this study, MDD was defined by at least one instance of a major depressive episode which initially identified 2659 cases, 17,237 controls and 98 missing (phenotype unknown) individuals.

In addition, the psychiatric history of cases and controls was examined using the Scottish Morbidity Record[26]. Within the control group, 1072 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 to 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 access to routine administrative clinical data.

In total there were 2605 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 cohort.

Replication cohort

Within the UK Biobank cohort, 25,035 participants (12,528 males and 12,507 females) completed a touchscreen assessment of depressive symptoms and previous treatment. These participants ranged from 40 to 79 years of age with a mean age of 57.8 years and a standard deviation of 8.0 years. On the basis of their responses to items from the Patient Health Questionnaire, diagnostic status was 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' using the definitions provided by Smith et al.[14]. MDD Cases were defined by reporting that they had ever been depressed/down for a whole week (UK Biobank field number 4598); plus this was for at least a two week period (UK Biobank field number 4609); plus this was for at least one episode (UK Biobank field number 4620); plus ever seen a GP (UK Biobank field number 2090) or psychiatrist (UK Biobank field number 2100) for nerves, anxiety, tension or depression. Alternatively, MDD cases were also defined by reporting that they had ever been uninterested in things or unable to enjoy the things you used to for at least a whole week (UK Biobank field number 4631); plus this was for at least a two week period (UK Biobank field number 5375); plus this was for at least one episode (UK Biobank field number 5386); plus ever seen a GP (UK Biobank field number 2090) or psychiatrist (UK Biobank field number 2100) for nerves, anxiety, tension or depression. In total there were 8508 cases and 16,527 controls, equating to a trait prevalence of 34.0% in this cohort, after the removal of individuals with insufficient information or ambiguous phenotypes.

Statistical approach

Discovery cohort

A mixed linear model was used to conduct an association analysis using GCTA v1.25.0:[27]

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

where $\mathbf{y}$ was the vector of binary observations for MDD. $\boldsymbol{\beta}$ was the matrix of fixed effects, including haplotype, sex, age and age[2]. Each unique haplotype was represented as a distinct allele and was either coded as 0, 1 or 2 depending on the number of haplotypes carried by that individual. $\mathbf{u}$ was fitted as a random effect taking into account the genomic relationships (MVN $(0,\mathbf{G}\sigma^2_{\mathbf{u}})$, where $\mathbf{G}$ was a SNP-based genomic relationship matrix[28]). $\boldsymbol{\nu}$ was a random effect fitting a second genomic relationship matrix $\mathbf{G_t}$(MVN $(0,\mathbf{G_t}\boldsymbol{\sigma}^2_{\mathbf{v}})$ which modelled only the more closely related individuals[29]. $\mathbf{G_t}$ was equal to $\mathbf{G}$ except that off-diagonal elements <0.05 were set to 0. $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ were the corresponding incidence matrices. $\boldsymbol{\varepsilon}$ was the vector of residual effects and was assumed to be normally distributed, MVN $(0,\mathbf{I}\boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}})$.

The inclusion of the second genomic relationship matrix, $\mathbf{G_t}$, was deemed desirable as the fitting of the single matrix $\mathbf{G}$ alone resulted in significant population stratification (intercept = 1.029 ± 0.003, λGC = 1.026) following examination with LD score regression[30]. 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).

Replication cohort

A mixed linear model was used to assess the haplotypes in UK Biobank, which were identified in the discovery cohort with $P < 10^{-6}$ using GCTA v1.25.0:[27]

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

where $\mathbf{y}$ was the vector of binary observations for MDD. $\boldsymbol{\beta}$ was the matrix of fixed effects, including haplotype, sex, age, age[2], genotyping batch and recruitment centre. $\mathbf{u}$ was fitted as a random effect taking into account the SNP-based genomic relationships (MVN $(0,\mathbf{G}\sigma^2_{\mathbf{u}})$.$\mathbf{X}$ and $\mathbf{Z}_1$ were the corresponding incidence matrices and $\boldsymbol{\varepsilon}$ was the vector of residual effects and was assumed to be normally distributed, MVN $(0, \mathbf{I}\sigma^2_{\varepsilon})$. Replication success was judged on the statistical significance of each haplotype using an inverse variance-weighted meta-analysis across both cohorts conducted using Metal[31].

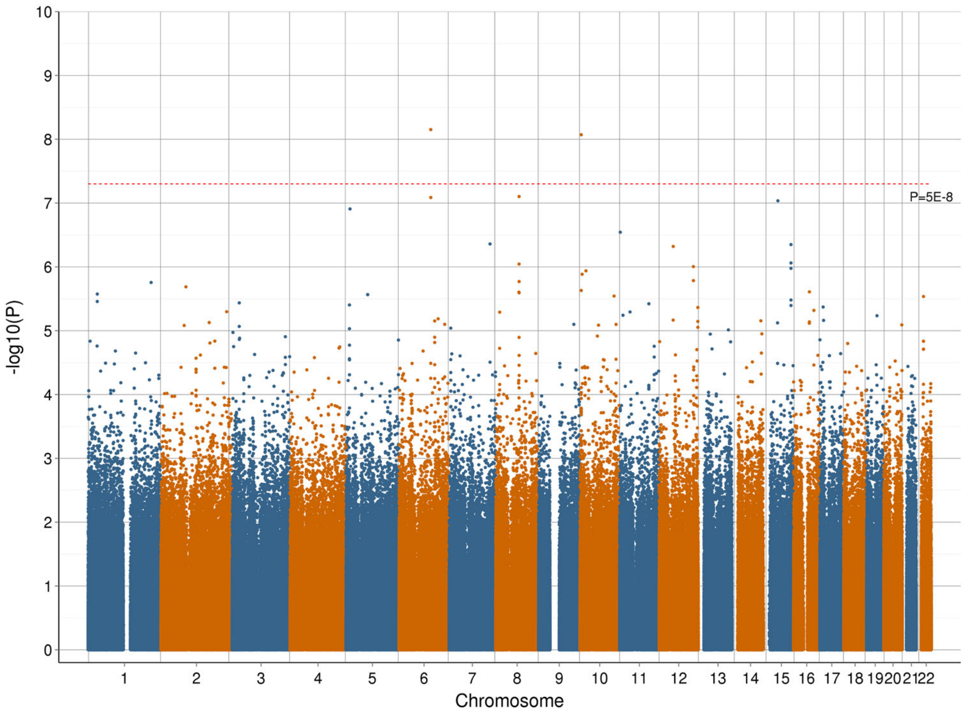

Fig. 1 Manhattan plot representing the $-\log_{10}$ *P*-values for an association between each assessed haplotype in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder

Fine mapping

The method described above examines the effect of each haplotype against all other haplotypes in that window. Therefore, a haplotype could be assessed against similar haplotypes containing the same causal variant, limiting any observed phenotypic association. 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 five SNPs in length. A total of 47 new haplotypes were assessed from within 26 pairs of directly overlapping haplotypes.

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 Fig. 1 with a q–q plot provided in Supplementary Fig. S1. 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. There were 12 haplotypes with $P < 10^{-6}$ in the discovery cohort with replication sought for these haplotypes using UK Biobank. Summary statistics from both cohorts and the meta-analysis for these 12 haplotypes are provided in Table 1. The protein coding genes which overlap these 12 haplotypes along with the observed haplotype frequencies within the two cohorts are provided in Table 2. The SNPs and alleles that constitute these 12 haplotypes are provided in Supplementary Table S1.

The two haplotypes on chromosome 6 (LD $r^2 = 0.74$) with $P < 10^{-6}$ in the discovery cohort both achieved nominal significance ($P < 0.05$) in the replication cohort (although these would not survive multiple testing correction for the 12 SNPs tested in the replication data set), with one reaching genome-wide significance ($P < 5 \times 10^{-8}$) in the meta-analysis. A regional association plot of the region surrounding these haplotypes within GS:SFHS is provided in Fig. 2. Fine mapping was used to form the most significant haplotype within the discovery cohort. Two directly overlapping 0.5 cM 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},

Table 1 The genetic association between major depressive disorder and 12 haplotypes in the generation Scotland: Scottish Family Health Study (GS:SFHS) discovery cohort (where $P < 10^{-6}$), the replication cohort (UK Biobank) and a meta-analysis

Haplotype			GS:SFHS		UK biobank		Meta-analysis	
Chr.	Position (bp)	Window size (cM)	Odds ratio (95% CI)	*P*-value	Odds ratio (95% CI)	*P*-value	Odds ratio (95% CI)	*P*-value
{ 6[a]	108,338,267 – 108,454,437	0.34	**1.83 (1.53–2.16)**	**7.06×10^{-9}**	**1.11 (1.01–1.22)**	**3.62×10^{-2}**	1.26 (1.16–1.37)	3.14×10^{-7}
6	108,407,662–108,454,437	0.25	1.68 (1.42–1.96)	8.17×10^{-8}	**1.14 (1.04–1.24)**	**4.47×10^{-3}**	**1.25 (1.16–1.35)**	**4.38×10^{-8}**
7	139,682,412–139,708,901	0.25	2.17 (1.67–2.73)	4.37×10^{-7}	0.87 (0.68–1.08)	2.20×10^{-1}	1.28 (1.08–1.49)	4.67×10^{-3}
{ 8	79,700,362–80,387,861	0.5	1.98 (1.56–2.46)	9.02×10^{-7}	1.06 (0.86–1.28)	5.93×10^{-1}	1.36 (1.18–1.56)	6.29×10^{-5}
8	79,759,499–80,156,474	0.25	1.77 (1.47–2.10)	7.90×10^{-8}	1.05 (0.91–1.21)	5.06×10^{-1}	1.28 (1.15–1.42)	1.14×10^{-5}
10	4,588,261–4,822,210	0.5	**2.33 (1.83–2.91)**	**8.50×10^{-9}**	1.15 (0.80–1.59)	4.39×10^{-1}	1.67 (1.40–1.98)	7.92×10^{-8}
11[a]	2,260,854–2,437,425	0.41	1.64 (1.38–1.91)	2.86×10^{-7}	1.00 (0.87–1.34)	9.91×10^{-1}	1.26 (1.10–1.34)	1.32×10^{-4}
12	48,159,721–48,263,828	0.25	2.00 (1.58–2.47)	4.78×10^{-7}	0.97 (0.79–1.17)	7.36×10^{-1}	1.29 (1.12–1.48)	6.51×10^{-4}
12	116,904,503–117,062,860	0.25	2.13 (1.64–2.69)	9.90×10^{-7}	1.04 (0.79–1.34)	7.79×10^{-1}	1.45 (1.22–1.71)	5.37×10^{-5}
15	49,206,902–49,260,601	0.25	2.03 (1.62–2.48)	9.21×10^{-8}	1.09 (0.88–1.32)	4.04×10^{-1}	1.41 (1.22–1.61)	4.39×10^{-6}
{ 15	93,806,447–93,851,224	0.5	1.58 (1.34–1.83)	4.47×10^{-7}	0.93 (0.81–1.05)	2.38×10^{-1}	1.16 (1.05–1.27)	2.50×10^{-3}
15	93,821,340–93,845,622	0.25	1.52 (1.31–1.75)	8.67×10^{-7}	0.91 (0.81–1.03)	1.37×10^{-1}	1.13 (1.03–1.23)	6.97×10^{-3}

Bold values indicate genome-wide statistical significance ($P < 5 \times 10^{-8}$) was achieved in the GS:SFHS cohort or the meta-analysis, or that nominal statistical significance ($P < 0.05$) was achieved in the UK Biobank. Base pair (bp) positions are based on build GRCh37. [a]indicates haplotype boundaries defined by the fine mapping approach. { indicates linkage disequilibrium (r^2) > 0.5 between haplotypes in the GS:SFHS cohort

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 ($P = 7.06 \times 10^{-9}$). Calculating the effect size at the population level[32], the estimates of the contribution of the two haplotypes to the total genetic variance was 2.09×10^{-4} and 2.38×10^{-4}, respectively, within GS:SFHS. None of the individual SNPs located within either haplotype were 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 within GS:SFHS using a 0.5 cM window. A regional association plot of the region surrounding this haplotype is provided in Fig. 3. This haplotype had an odds ratio (OR) of 2.33 (95% confidence interval (CI): 1.83 – 2.91) in the discovery cohort and an OR of 1.15 (95% CI: 0.80–1.59) in the replication cohort. These were the highest ORs observed in the respective cohorts. The estimate of the contribution of this haplotype to the total genetic variance was 2.29×10^{-4} in the discovery cohort. 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*-values < 0.001.

All 12 of the haplotypes with a *P*-value for association $<10^{-6}$ in the GS:SFHS discovery cohort were risk factors for MDD (OR > 1). Within the replication cohort, 7 out of these 12 haplotypes had OR > 1, however, only of two of these had the lower bound of the 95% confidence interval > 1. None of the 95% confidence intervals for the replication ORs overlapped the 95% confidence intervals of the discovery GS:SFHS cohort.

Discussion

Twelve haplotypes were identified in the discovery cohort with $P < 10^{-6}$ of which two were significant at the genome-wide level ($P < 5 \times 10^{-8}$) in the discovery cohort and one which was genome-wide significant ($P < 5 \times 10^{-8}$) in the meta-analysis. A power analysis[33] was conducted using the genotype relative risks observed in the discovery cohort, the sample sizes and haplotype frequencies in the replication cohort and the prevalence of MDD reported for a structured clinical diagnosis of MDD in other high income counties (14.6%)[34]. There was sufficient power (>0.99) to detect the twelve haplotypes with $P < 10^{-6}$ identified in the discovery cohort within the replication cohort at a significance threshold of 0.05.

There are several reasons why the effect sizes observed in the replication cohort were lower than those observed in the discovery cohort. The causal loci may have been in lower LD with the assessed haplotypes in the replication cohort than in the discovery cohort lessening the observed effect. The phenotypes across the two cohorts were potentially heterogeneous (certainly with regards to the prevalence in each population) so the assessed haplotypes may have had differing effects on each cohort's phenotype. A complementary approach to replication is to identify the gene coding regions within haplotypes that potentially provide a biologically informative explanation for an association with MDD. Those haplotypes with

Table 2 Protein coding genes located overlapping with the 12 haplotypes with $P < 10^{-6}$ in the generation Scotland: Scottish family health study (GS:SFHS) discovery cohort and the frequencies of those haplotypes in GS:SFHS and UK Biobank

			Haplotype frequency	
Chr.	Position (bp)	Protein coding genes	GS:SFHS	UK Biobank
{ 6	108,338,267–108,454,437	OSTM1	0.0152	0.0197
6	108,407,662–108,454,437	OSTM1	0.0193	0.0241
7	139,682,412–139,708,901	TBXAS1	0.0066	0.0069
{ 8	79,700,362–80,387,861	IL7	0.0076	0.0081
8	79,759,499–80,156,474	IL7	0.0147	0.0157
10	4,588,261–4,822,210		0.0064	0.0027
11	2,260,854–2,437,425	ASCL2, CLorf21, TSPAN32, CD81, TSSC4, TRPM5	0.0196	0.0187
12	48,159,721–48,263,828	SLC48A1, RAPGEF3, HDAC7, VDR	0.0078	0.0090
12	116,904,503–117,062,860	MAP1LC3B2	0.0057	0.0045
15	49,206,902–49,260,601	SHC4	0.0082	0.0080
{ 15	93,806,447–93,851,224		0.0224	0.0206
15	93,821,340–93,845,622		0.0265	0.0243

Base pair (bp) positions are based on build GRCh37 with protein coding regions obtained from Ensembl, GRCh37.p13. Haplotype frequencies were calculated using unrelated individuals and excluding UK Biobank participants recruited in Glasgow or Edinburgh. { indicates a linkage disequilibrium (r^2) > 0.5 between haplotypes in the GS:SFHS cohort

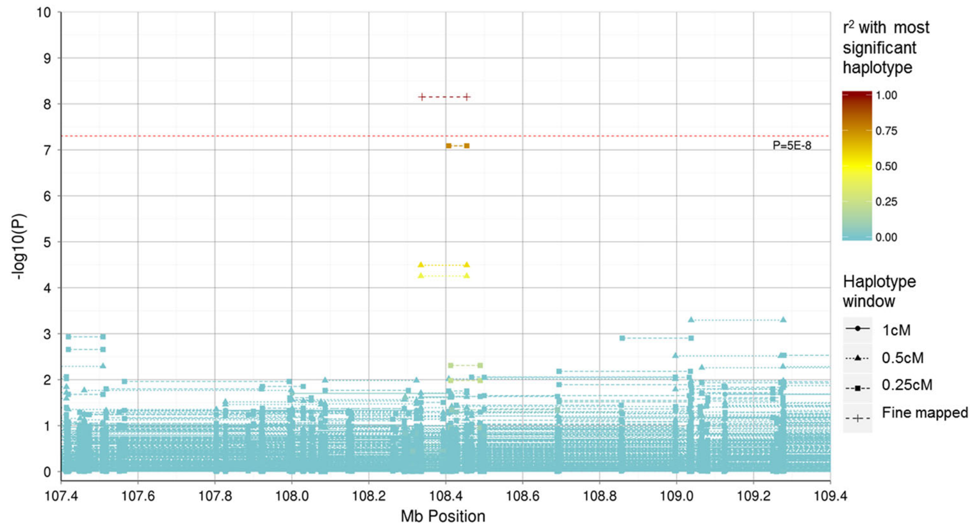

Fig. 2 Regional association plot representing the $-\log_{10}$ *P*-values for an association between haplotypes in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder within the 107.4–107.6 Mb region on chromosome 6. The start and end position (using build GRCh37) 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

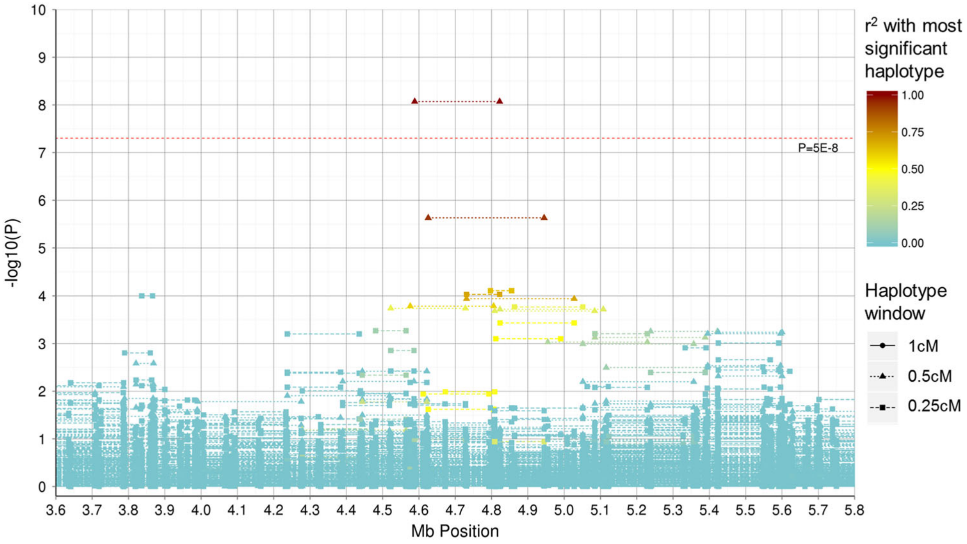

Fig. 3 Regional association plot representing the $-\log_{10}$ *P*-values for an association between haplotypes in the Generation Scotland: Scottish Family Health Study cohort and Major Depressive Disorder within the 3.6–5.8 Mb region on chromosome 10. The start and end position (using build GRCh37) 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

$P < 10^{-7}$ in the discovery cohort and the gene coding regions that they overlap are discussed below.

The two haplotypes on chromosome 6 overlapped with the Osteopetrosis Associated Transmembrane Protein 1 (*OSTM1*) coding gene. *OSTM1* is associated with neurodegeneration[35,36] and melanocyte function[37], and alpha-melanocyte-stimulating hormone has been shown to have an effect on depression-like symptoms[38–40]. This haplotype lies within the 6q21 region that has been associated with bipolar disorder[41–45], a disease that shares symptoms with MDD and has a correlated phenotypic liability of 0.64[46]. This may indicate 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 8 overlapped with the Interleukin 7 (*IL7*) protein coding region. *IL7* is involved in maintaining T-cell homoeostasis[47] and proliferation[48], 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[49], with depressed subjects found to have elevated[50] or depressed levels[51] of *IL7* serum. There is conjecture as to whether MDD causes inflammation or represents a reaction to an increased inflammatory response[52,53], but it is most likely to be a bidirectional relationship[51].

The haplotype on chromosome 10 overlapped with two RNA genes: long intergenic non-protein coding RNA 704 (*LINC00704*) and long intergenic non-protein coding RNA 705 (*LINC00705*). The function of these non-protein coding genes is unreported. However, a study of cardiac neonatal lupus, which is a rare autoimmune disease demonstrated an association for a SNP (rs1391511) which is 15kb from *LINC00705*.

Two Dutch studies[54,55] have identified a variant (rs8023445) on chromosome 15 located within the SRC (Src homology 2 domain containing) family, member 4 (*SHC4*) gene coding region that has a moderate degree of association with MDD ($P = 1.64 \times 10^{-5}$ and $P = 9 \times 10^{-6}$, respectively). A variant (rs10519201) within the *SHC4* coding region was also found to have an association ($P = 6.16 \times 10^{-6}$) with Obsessive-Compulsive Personality Disorder in a UK-based study[56]. *SHC4* is expressed in neurons[57] and regulates BDNF-induced MAPK activation[58], which has been shown to be a key factor in MDD pathophysiology[59]. The *SHC4* region overlaps with the haplotype on chromosome 15 identified in the discovery cohort (located at 49,206,902–49,260,601 bp) and, therefore, further research to examine the association between the *SHC4* region and psychiatric disorders could be warranted.

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, such as in GS:SFHS which was a family-based cohort, where there is a greater likelihood that a single haplotype is tagging the same causal variant across that population. The UK Biobank was selected as replication cohort as it is a large population-based sample that was expected to be genetically similar to the GS:SFHS discovery cohort. This was confirmed by the similarity of the observed haplotype frequencies (Table 2) between the two cohorts. The prevalence of MDD observed in the discovery cohort (13.7%) was comparable to that reported (14.6%) within similar populations[34]. However, in the replication cohort, the trait prevalence was notably higher (34.0%), most likely due to the differing methods of phenotypic ascertainment. Additional work could seek to replicate the findings in further cohorts, as well as full meta-analysis of all haplotypes within those 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 association with MDD.

Conclusions

This study identified two haplotypes within the discovery cohort that exceeded genome-wide significance for association with a clinically diagnosed MDD phenotype. One of these haplotypes was nominally significant in the replication cohort and was in LD with a haplotype that was genome-wide significant in the meta-analysis. The genome-wide significant haplotype on chromosome 6 was located on 6q21, which has been shown previously to be related to psychiatric disorders. There were a number of haplotypes approaching genome-wide significance located within genic regions associated with diseases that are comorbid 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 further analysis.

Acknowledgements

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 GS:SFHS 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). 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). This research has been conducted using the UK Biobank resource–application number 4844; we are grateful to UK Biobank participants. The UK Biobank study was conducted under generic approval from the NHS National Research Ethics Service (approval letter dated 17th June 2011, Ref 11/NW/0382). Y.Z. acknowledges support from China Scholarship Council. I.J.D. 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). A.M. McI. and T.-K.C. acknowledge support from the Dr. Mortimer and Theresa Sackler Foundation.

Author details

[1]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK. [2]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [4]Generation Scotland, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5]Division of Population Health Sciences, University of Dundee, Dundee, UK. [6]Aberdeen Biomedical Imaging Centre, University of Aberdeen, Aberdeen, UK. [7]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK. [8]Department of Psychology, The University of Edinburgh, Edinburgh, UK

Competing interests

D.J.P. and I.J.P. are participants in UK Biobank. The authors declare that they have no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

The online version of this article (https://doi.org/10.1038/s41398-017-0010-9) contains supplementary material.

Received: 20 March 2017 Revised: 16 August 2017 Accepted: 20 August 2017
Published online: 30 November 2017

References

1. Huffman, J. C., Celano, C. M., Beach, S. R., Motiwala, S. R. & Januzzi, J. L. Depression and cardiac disease: epidemiology, mechanisms, and diagnosis. *Cardiovasc. Psychiatr. Neurol.* **2013**, 14 (2013).
2. Kang, H. -J. et al. Comorbidity of depression with physical disorders: research and clinical implications. *Chonnam Med. J.* **51**, 8–18 (2015).
3. Raison, C. L., Capuron, L. & Miller, A. H. Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends Immunol.* **27**, 24–31 (2006).
4. Major Depressive Disorder Working Group of the Psychiatric Gwas Consortium. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatr.* **18**, 497–511 (2013).
5. Converge Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
6. Levinson, D. F. et al. Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it? *Biol. Psychiatr.* **76**, 510–512 (2014).
7. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
8. Lubke, G. H. et al. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol. Psychiatr.* **72**, 707–709 (2012).
9. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic epidemiology of major depression: review and meta-analysis. *Am. J. Psychiatr.* **157**, 1552–1562 (2000).
10. Zhang, Z. et al. A haplotype in the 5′-upstream region of the NDUFV2 gene is associated with major depressive disorder in Han Chinese. *J. Affect. Disord.* **190**, 329–332 (2016).
11. Kim, J. -J. et al. Is there protective haplotype of dysbindin gene (DTNBP1) 3 polymorphisms for major depressive disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatr.* **32**, 375–379 (2008).
12. Klok, M. D. et al.A common and functional mineralocorticoid receptor haplotype enhances optimism and protects against depression in females. *Transl. Psychiatr.* **1**, e62 (2011).
13. Smith, B. H. 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. *Int. J. Epidemiol.* **42**, 689–700 (2013).

14. Smith, D. 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* **8**, e75362 (2013).
15. Amador, C. et al. Recent genomic heritage in Scotland. *BMC Genomics* **16**, 1–17 (2015).
16. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK biobank data: come and get it. *Sci. Transl. Med.* **6**, 224ed224 (2014).
17. Marchini J. UK Biobank phasing and imputation documentation. Version 1.2: http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf (2015).
18. Delaneau, O., Zagury, J. -F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
19. O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
20. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
21. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
22. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
23. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
24. First, M. B., Spitzer, R. L., Miriam, G., Williams, J. B. W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders*, Research Version, Patient Edition. (SCID-I/P) (2002).
25. Fernandez-Pujals, A. M. 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* **10**, e0142197 (2015).
26. Information Services Division. SMR Data Manual: http://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets (2016).
27. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
28. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
29. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
30. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
31. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
32. Park, J. -H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
33. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
34. Bromet, E. et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med.* **9**, 1–16 (2011).
35. Kasper, D. et al. Loss of the chloride channel ClC-7 leads to lysosomal storage disease and neurodegeneration. *EMBO J.* **24**, 1079–1091 (2005).
36. Pandruvada, S. N. M. et al. Role of ostm1 cytosolic complex with kinesin 5B in intracellular dispersion and trafficking. *Mol. Cell. Biol.* **36**, 507–521 (2016).
37. Hoek, K. S. et al. *Novel MITF targets identified usin*g a two-step DNA microarray strategy. *Pigment Cell Melanoma Res.* **21**, 665–676 (2008).
38. Maes, M. et al. Abnormal pituitary function during melancholia: Reduced α-melanocyte-stimulating hormone secretion and increased intact ACTH non-suppression. *J. Affect. Disord.* **22**, 149–157 (1991).
39. Goyal, S. N., Kokare, D. M., Chopde, C. T. & Subhedar, N. K. Alpha-melanocyte stimulating hormone antagonizes antidepressant-like effect of neuropeptide Y in Porsolt's test in rats. *Pharmacol. Biochem. Behav.* **85**, 369–377 (2006).
40. Kokare, D. M., Singru, P. S., Dandekar, M. P., Chopde, C. T. & Subhedar, N. K. Involvement of alpha-melanocyte stimulating hormone (α-MSH) in differential ethanol exposure and withdrawal related depression in rat: Neuroanatomical–behavioral correlates. *Brain Res.* **1216**, 53–67 (2008).
41. Knight, J., Rochberg, N. S., Saccone, S. F., Nurnberger, J. I. & Rice, J. P. An investigation of candidate regions for association with bipolar disorder. *Am. J. Med. Genet. Part B: Neuropsychiatr. Genet.* **153B**, 1292–1297 (2010).
42. Dick, D. M. et al. Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the national institute of mental health genetics initiative. *Am. J. Hum. Genet.* **73**, 107–114 (2003).
43. Park, N. et al. Linkage analysis of psychosis in bipolar pedigrees suggests novel putative loci for bipolar disorder and shared susceptibility with schizophrenia. *Mol. Psychiatr.* **9**, 1091–1099 (2004).
44. Pato, C. N. et al. Genome-wide scan in Portuguese Island families implicates multiple loci in bipolar disorder: Fine mapping adds support on chromosomes 6 and 11. *Am. J. Med. Genet. Part B: Neuropsychiatr. Genet.* **127B**, 30–34 (2004).
45. Fabbri, C. & Serretti, A. Genetics of long-term treatment outcome in bipolar disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatr.* **65**, 17–24 (2016).
46. McGuffin, P. et al. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatr.* **60**, 497–502 (2003).
47. Surh, C. D. & Sprent, J. Homeostasis of naive and memory T Cells. *Immunity* **29**, 848–862 (2008).
48. Kittipatarin, C. & Khaled, A. R. Interlinking interleukin-7. *Cytokine* **39**, 75–83 (2007).
49. Miller, A. H. Depression and immunity: A role for T cells? *Brain Behav. Immun.* **24**, 1–8 (2010).
50. Simon, N. M. et al. A detailed examination of cytokine abnormalities in major depressive disorder. *Eur. Neuropsychopharmacol.* **18**, 230–233 (2008).
51. Lehto, S. M. et al. Serum IL-7 and G-CSF in major depressive disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatr.* **34**, 846–851 (2010).
52. Stewart, J. C., Rand, K. L., Muldoon, M. F. & Kamarck, T. W. A prospective evaluation of the directionality of the depression-inflammation relationship. *Brain Behav. Immun.* **23**, 936–944 (2009).
53. Irwin, M. R. & Miller, A. H. Depressive disorders and immunity: 20 years of progress and discovery. *Brain Behav. Immun.* **21**, 374–383 (2007).
54. 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. *J. Affect. Disord.* **133**, 516–521 (2011).
55. Sullivan, P. F. et al. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol. Psychiatr.* **14**, 359–375 (2008).
56. Boraska, V. et al. Genome-wide association analysis of eating disorder-related symptoms, behaviors, and personality traits. *Am. J. Med. Genet.* **159B**, 803–811 (2012).
57. Hawley, S. P., Wills, M. K. B., Rabalski, A. J., Bendall, A. J. & Jones, N. Expression patterns of ShcD and Shc family adaptor proteins during mouse embryonic development. *Dev. Dynam.* **240**, 221–231 (2011).
58. You, Y. et al. ShcD interacts with TrkB via its PTB and SH2 domains and regulates BDNF-induced MAPK activation. *BMB Rep.* **43**, 485–490 (2010).
59. Duric, V. et al. A negative regulator of MAP kinase causes depressive behavior. *Nat. Med.* **16**, 1328–1332 (2010).