Genome-wide association study of depression phenotypes in UK Biobank (n = 322,580) identifies the enrichment of variants in excitatory synaptic pathways

David M. Howard*[1], Mark J. Adams[1], Masoud Shirali[1], Toni-Kim Clarke[1], Riccardo E. Marioni[2], Gail Davies[2,3], Jonathan R. I. Coleman[4,5], Clara Alloza[1], Xueyi Shen[1], Miruna C. Barbu[1], Eleanor M. Wigmore[1], Saskia P. Hagenaars[4,5], Cathryn M. Lewis[4,5], Daniel J. Smith[6], Patrick F. Sullivan[7,8,9], Chris S. Haley[10], Gerome Breen[4,5], Ian J. Deary[2,3], and Andrew M. McIntosh[1,3]

Affiliations:

[1]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK

[2]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

[3]Department of Psychology, University of Edinburgh, Edinburgh, UK

[4]MRC Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK

[5]NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Trust, London, UK

[6]Institute of Health and Wellbeing, University of Glasgow, UK

[7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[8]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA

[9]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA

[10]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

*Corresponding author: David M. Howard

Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK

+44 131 537 6268

(e-mail: D.Howard@ed.ac.uk)

Abstract

Depression is a polygenic trait that causes extensive periods of disability and increases the risk of suicide, a leading cause of death in young people. Previous genetic studies have identified a number of common risk variants which have increased in number in line with increasing sample sizes. We conducted a genome-wide association study (GWAS) in the largest single population-based cohort to date, UK Biobank. This allowed us to estimate the effects of ≈ 20 million genetic variants in 320,000 people for three depression phenotypes: broad depression, probable major depressive disorder (MDD), and International Classification of Diseases (ICD, version 9 or 10) coded MDD. Each phenotype was found to be significantly genetically correlated with the results from a previous independent study of clinically defined MDD. We identified 12 independent loci that were significantly associated ($P < 5 \times 10^{-8}$) with broad depression, two independent variants for probable MDD, and one independent variant for ICD-coded MDD. Gene-based analysis of our GWAS results with MAGMA revealed 52 regions significantly ($P < 2.72 \times 10^{-6}$) associated with broad depression, six regions in probable MDD and three regions in ICD-coded MDD. Gene region-based analysis of our GWAS results with MAGMA revealed 57 regions significantly ($P < 6.01 \times 10^{-6}$) associated with broad depression, of which 35 were also detected by gene-based analysis. Variants for broad depression were enriched in pathways for excitatory neurotransmission and neuron spines. This study provides a number of novel genetic risk variants that can be leveraged to elucidate the mechanisms of MDD and low mood.

Introduction

Depression is ranked as the largest contributor to global disability affecting 322 million people[1]. The heritability (h^2) of major depressive disorder (MDD) is estimated at 37% from twin studies[2] and common single nucleotide polymorphisms (SNPs) contribute approximately 9% to variation in liability[3], providing strong evidence of a genetic contribution to its causation. Previous genetic association studies have used a number of depression phenotypes, including self-declared depression[4],

clinician diagnosed MDD[5] and depression ascertained via electronic health records[6], with some evidence of overlapping genetic architecture between a subset of these definitions. Different definitions of depression are rarely included in large sample studies, although UK Biobank is an exception. The favouring of greater sample size over clinical precision has yielded a steady increase over time in the number of variants for ever more diverse MDD phenotypes[3-5,7]. In the current paper, we extend this approach to the study of three depression-related phenotypes within the large UK Biobank cohort and identify new disease biology based upon our findings.

The UK Biobank cohort provides data on over 500,000 individuals and represents an opportunity to conduct the largest association analysis of depression to date within a single cohort. This cohort has been extensively phenotyped allowing us to derive three depression traits: self-reported past help-seeking for problems with 'nerves, anxiety, tension or depression' (hereby termed 'broad depression'); self-reported depressive symptoms with associated impairment (termed 'probable MDD'); and MDD identified from International Classification of Diseases (ICD)-9 or ICD-10 hospital admission records (termed ICD-coded MDD). We also conducted a gene-based analyses with the MAGMA software package[8] to identify genes and regions associated with each phenotype and used GTEx[9] to identify if the significant variants identified were expression quantitative trait loci (eQTL).

Materials and Methods

The UK Biobank cohort is a population-based cohort consisting of 501,726 individuals, recruited at 23 centres across the United Kingdom. Genotypic data was available for 488,380 individuals and was imputed with IMPUTE4 and used the HRC, UK10K and 1,000 Genomes Phase 3 reference panels to identify ≈ 93M variants for 487,409 individuals[10]. We excluded 131,790 related individuals based on a shared relatedness of up to the third degree using kinship coefficients (> 0.044) calculated using the KING toolset[11], and excluded a further 79,990 individuals that were either not recorded as "white British", outliers based on heterozygosity, or had a variant call rate < 98%. We subsequently added

back in one member of each group of related individuals by creating a genomic relationship matrix and selected individuals with a genetic relatedness less than 0.025 with any other participant (n = 55,745). We removed variants with a call rate < 98%, a minor allele frequency < 0.01, those that deviated from Hardy-Weinberg equilibrium ($P < 10^{-6}$), had an imputation accuracy score < 0.1, or had a minor allele frequency ± 0.2 of that reported in the 1,000G (http://www.internationalgenome.org/data), UK10K (http://www.uk10k.org/data.html), and HRC (http://www.haplotype-reference-consortium.org/site) site lists leaving a total of 19,632,042 variants for 331,374 individuals.

Extensive phenotypic data were collected for UK Biobank participants using health records, biological sampling, physical measures, and touchscreen tests and questionnaires. We used three definitions of depression in the UK Biobank sample, which are explained in greater depth in the Supplementary Information and are summarised below.

Broad depression *phenotype*

The broadest phenotype (broad depression) was defined using self-reported help-seeking behaviour for mental health difficulties. Case and control status was determined by the touchscreen response to the single question 'Have you ever seen a general practitioner (GP) for nerves, anxiety, tension or depression?'. Caseness for broad depression was determined by answering 'Yes' to this question at either the initial assessment visit or at any repeat assessment visit or if there was a primary or secondary diagnosis of a depressive mood disorder from linked hospital admission records. The remaining respondents were classed as controls if they provided a 'No' response during either the initial assessment visit or the first repeat assessment visit.

Probable MDD phenotype

The second depression phenotype (probable MDD) was derived from touchscreen responses to questions about the presence and duration of low mood and anhedonia, following the definitions from Smith, et al. [12]. Whereby the participant had indicated that they were 'Depressed/down for a whole week; plus at least two weeks duration; plus ever seen a GP or psychiatrist for 'nerves, anxiety, or

depression' OR ever anhedonia for a whole week; plus at least two weeks duration; plus ever seen a GP or psychiatrist for 'nerves, anxiety, or depression'. Cases for the probable MDD definition were supplemented by diagnoses of depressive mood disorder from linked hospital admission records.

ICD-coded phenotype

The ICD-coded MDD phenotype was derived from linked hospital admission records. Participants were classified as cases if they had either an ICD-10 primary or secondary diagnosis for a mood disorder. ICD-coded MDD controls were participants who had linked hospital records but who did not have any diagnosis of a mood disorder and were not probable MDD cases.

For the three UK Biobank depression phenotypes we excluded: participants who were identified with bipolar disorder, schizophrenia, or personality disorder using self-declared data, touchscreen responses (per Smith, et al. [12]), or ICD codes from hospital admission records; and participants who reported having a prescription for an antipsychotic medication during a verbal interview. Further exclusions were applied to control individuals if they had a diagnosis of a depressive mood disorder from hospital admission records, had reported having a prescription for antidepressants, or self-reported depression (see Supplementary Information for full phenotype criteria and UK Biobank field codes). This provided a total of 113,769 cases and 208,811 controls (n_{total} = 322,580, prevalence = 35.27%) for the broad depression phenotype, 30,603 cases and 143,916 controls (n_{total} = 174,519, prevalence = 17.54%) for the probable MDD phenotype, and 8,276 cases and 209,308 controls (n_{total} = 217,584, prevalence = 3.80%) for the ICD-coded MDD phenotype.

To validate the three phenotypes we derived for the UK Biobank cohort, genetic correlations were calculated using Linkage Disequilibrium Score regression (LDSR)[13] using summary statistics from the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. [5] study that predominately made use of a clinically derived phenotype for MDD. We also calculated the genetic correlation with a neuroticism phenotype[14].

Association analysis

We performed a linear association test to assess the effect of each variant using BGENIE v1.1[10]:

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

$$\hat{\mathbf{y}} = \mathbf{X\beta}$$

$$(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{G}b + \boldsymbol{\varepsilon}_2$$

where $\mathbf{y}$ was the vector of binary observations for each phenotype (controls coded as 0 and cases coded as 1). $\boldsymbol{\beta}$ was the matrix of fixed effects, including sex, age, genotyping array, and 8 principal components and $\mathbf{X}$ was the corresponding incidence matrices. $(\mathbf{y} - \hat{\mathbf{y}})$ was a vector of phenotypes residualized on the fixed effect predictors, $\mathbf{G}$ was a vector of expected genotype counts of the effect allele (dosages), b was the effect of the genotype on residualized phenotypes, and $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ were vectors of normally distributed errors.

Genome-wide statistical significance was determined by the conventional threshold of a *P*-value of association $< 5 \times 10^{-8}$. To determine significant variants that were independent the clump command in Plink 1.90b4[15] was applied using --clump-p1 1e-4 --clump-p2 1e-4 --clump-r2 0.1 --clump-kb 3000, mirroring the approach of Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium., et al. [3]. Therefore variants which were within 3Mb of each other and shared a linkage disequilibrium greater than 0.1 were clumped together and only the most significant variant reported. Due to the complexity of major histocompatibility complex (MHC) region an approach similar to that of The Schizophrenia Psychiatric Genome-Wide Association Study [16] was taken and only the most significant variant across that region is reported.

LDSR[13] was used to provide a SNP-based estimate of the heritability of the phenotypes using the whole-genome summary statistics obtained by the association analyses and was also used to examine the data for evidence of inflation of the test statistics based on the intercept due to population stratification.

Gene- and region-based analyses

Two downstream analyses of the results were conducted using MAGMA[8] (Multi-marker Analysis of GenoMic Annotation) by applying a principal component regression model to the results of our association analyses. In the first downstream analysis, a gene-based analysis was performed for each phenotype using the results from the genome-wide association analyses. Genetic variants were assigned to genes based on their position according to the NCBI 37.3 build, resulting in a total of 18,380 genes being analysed. The European panel of the 1,000 Genomes data (phase 1, release 3)[17] was used as a reference panel to account for linkage disequilibrium. A genome-wide significance threshold for gene-based associations was calculated using the Bonferroni method ($\alpha = 0.05 / 18{,}385$; $P < 2.72 \times 10^{-6}$).

In the second downstream analysis, a region-based analysis was performed for each phenotype. To determine the regions, haplotype blocks identified by recombination hotspots were used as described by Shirali, et al. [18] and implemented in an analysis of MDD by Zeng, et al. [19] for detecting causal regions. Block boundaries were defined by hotspots of at least 30 cM per Mb based on a European subset of the 1,000 genome project recombination rates. This resulted in a total of 8,350 regions being analysed using the European panel of the 1,000 Genomes data (phase 1, release 3)[17] as a reference panel to account for linkage disequilibrium. A genome-wide significance threshold for region-based associations was calculated using the Bonferroni correction method ($\alpha = 0.05 / 8{,}310$; $P < 6.01 \times 10^{-6}$).

Pathway analysis

The pathway analysis was performed on our gene-based analysis results. The analysis was a gene-set enrichment analysis that was conducted utilising gene-annotation files from the Gene Ontology (GO) Consortium (http://geneontology.org/)[20] taken from the Molecular Signatures Database (MSigDB) v5.2. The GO consortium includes gene-sets for three ontologies; molecular function, cellular components and biological function. This annotation file consisted of 5,917 gene-sets which were

corrected for multiple testing correction using the MAGMA default setting correcting for 10,000 permutations.

eQTL identification

The online GTEx portal (https://www.gtexportal.org/home/) was used to determine whether any of the genome-wide significant variants for each phenotype were eQTL[9].

Results

We conducted a genome-wide association analysis testing the effect of 20,019,946 variants on three depression phenotypes using up to 322,580 UK Biobank participants. The study demographics for each UK Biobank phenotype and within the case and control groups are provided in Table 1.

Table 1. Number of individuals, number of each sex, mean age in years, age range in years for each of the assessed UK Biobank phenotypes and within the respective case and control groups

Phenotype	Status	N	Males	Females	Mean Age (st.dev)	Age Range
Broad depression	Cases	113,769	40,477	73,292	56.5 (7.8)	39 -73
	Controls	208,811	109,426	99,385	57.1 (8.1)	39-72
	Total	322,580	149,903	172,677	56.9 (8.0)	39-73
Probable MDD	Cases	30,603	11,346	19,527	56.1 (7.8)	40-70
	Controls	143,916	65,015	78,901	57.1 (7.9)	39-73
	Total	174,519	76,361	98,158	56.9 (7.9)	39-73
ICD-coded MDD	Cases	8,276	3,098	5,178	56.5 (7.9)	40-70
	Controls	209,308	99,961	109,347	57.6 (8.0)	39-73
	Total	217,584	103,059	114,525	57.6 (8.0)	39-73

The estimated SNP-based heritabilities, genetic correlations between each UK Biobank phenotype and genetic correlations with a clinically defined MDD phenotype and obtained from the study conducted by the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. [5] and a neuroticism phenotype[14] for each UK Biobank phenotype are provided in Table 2.

Table 2. The SNP-based heritability (h^2), the genetic correlations (r_g) between each UK Biobank phenotype and the r_g with a depression and a neuroticism phenotype obtained from separate studies† for each of the assessed UK Biobank phenotypes.

Phenotype	**h^2 (s.e)**	**r_g with MDD† (s.e.)**	**r_g with neuroticism† (s.e.)**	**r_g with broad depression (s.e.)**	**r_g with probable MDD (s.e.)**
Broad depression	0.102 (0.004)	0.803 (0.091)	0.671 (0.018)	-	0.871 (0.050)
Probable MDD	0.047 (0.006)	0.691 (0.138)	0.517 (0.050)	0.871 (0.050)	-
ICD-coded MDD	0.067 (0.080)	0.666 (0.121)	0.560 (0.045)	0.864 (0.046)	0.853 (0.053)

†To conduct the genetic correlations with the UK Biobank phenotypes the depression phenotype from Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. [5] and the neuroticism phenotype[14] was used

There were 2,200 variants that were genome-wide significant ($P < 5 \times 10^{-8}$) for an association with broad depression, of which 12 were independent (Table 3). The association analysis of probable MDD identified 24 variants with $P < 5 \times 10^{-8}$ and of these two were independent (Table 4). There was one independent genome-wide significant variants for ICD-coded MDD (Table 5). Manhattan plots of all the variants analysed are provided in Figures 1, 2, and 3 for broad depression, probable MDD, and ICD-coded MDD, respectively. Q-Q plots of the observed *P*-values on those expected are provided in Supplementary Figures 1, 2, and 3 for broad depression, probable MDD, and ICD-coded MDD, respectively. There were 5,614 variants with $P < 1 \times 10^{-6}$ for an association with broad depression (see Supplementary Table 1), 224 variants with $P < 1 \times 10^{-6}$ for an association with probable MDD (see Supplementary Table 2), and 142 variants with $P < 1 \times 10^{-6}$ for an association with ICD-coded MDD (see Supplementary Table 3). None of the phenotypes examined provided evidence of inflation of the test statistics due to population stratification (see Supplementary Table 4).

Table 3. Independent variants with a genome-wide significant ($P < 5 \times 10^{-8}$) association with broad depression

Chr	Marker name	Position	A1	A2	Allele Frequency	Imputation Accuracy	Beta	Standard error	Gene +/- 10kb	*P*-value
1	rs6699744	72825144	T	A	0.617	1.00	0.0061	0.0008	-	1.64×10^{-13}
1	rs7548151	177026983	A	G	0.089	1.00	0.0051	0.0009	*ASTN1*	3.87×10^{-9}
5	rs40465	103981726	G	T	0.325	1.00	0.0052	0.0008	*RP11-6N13.1*	4.45×10^{-10}
6	rs3094054	30333505	T	G	0.136	1.00	-0.0061	0.0008	-	1.79×10^{-13}
7	rs3807865	12250402	A	G	0.421	1.00	0.0057	0.0008	*TMEM106B*	7.28×10^{-12}
7	rs2402273	117600424	C	T	0.405	1.00	0.0050	0.0008	-	1.95×10^{-9}
9	rs263575	17033840	A	G	0.466	1.00	-0.0047	0.0008	-	2.31×10^{-8}
10	rs1021363	106610839	G	A	0.648	1.00	-0.0048	0.0008	*SORCS3*	1.04×10^{-8}
11	rs10501696	88748162	G	A	0.495	0.99	-0.0055	0.0008	*GRM5*	6.73×10^{-11}
11	rs11018449	88797386	T	C	0.327	1.00	0.0045	0.0008	*GRM5*	4.52×10^{-8}
13	rs9530139	31847324	T	C	0.195	1.00	-0.0049	0.0008	*B3GLCT*	2.63×10^{-9}
15	rs28541419	88945878	G	C	0.231	1.00	-0.0046	0.0008	-	2.78×10^{-8}

The allele frequency and reported effect size (beta) is for the A1 allele. The chromosome (Chr) and position in Mb is given with regards to the GRCh37 assembly.

Table 4. Independent variants with a genome-wide significant ($P < 5 \times 10^{-8}$) association with probable MDD

Chr	Marker name	Position	A1	A2	Allele Frequency	Imputation Accuracy	Beta	Standard error	Gene +/- 10kb	*P*-value
2	rs10929355	15398964	G	T	0.470	1.00	-0.0053	0.0009	*NBAS*	5.84×10^{-9}
7	rs5011432	12268668	C	A	0.418	1.00	0.0051	0.0009	*TMEM106B*	2.23×10^{-8}

The allele frequency and reported effect size (beta) is for the A1 allele. The chromosome (Chr) and position in Mb is given with regards to the GRCh37 assembly.

Table 5. Independent variants with a genome-wide significant ($P < 5 \times 10^{-8}$) association with ICD-coded MDD

Chr	Marker name	Position	A1	A2	Allele Frequency	Imputation Accuracy	Beta	Standard error	Gene +/- 10kb	*P*-value
7	rs1554505	1983929	A	G	0.742	1.00	0.0025	0.0004	*MAD1L1*	2.74×10^{-9}

The allele frequency and reported effect size (beta) is for the A1 allele. The chromosome (Chr) and position in Mb is given with regards to the GRCh37 assembly.

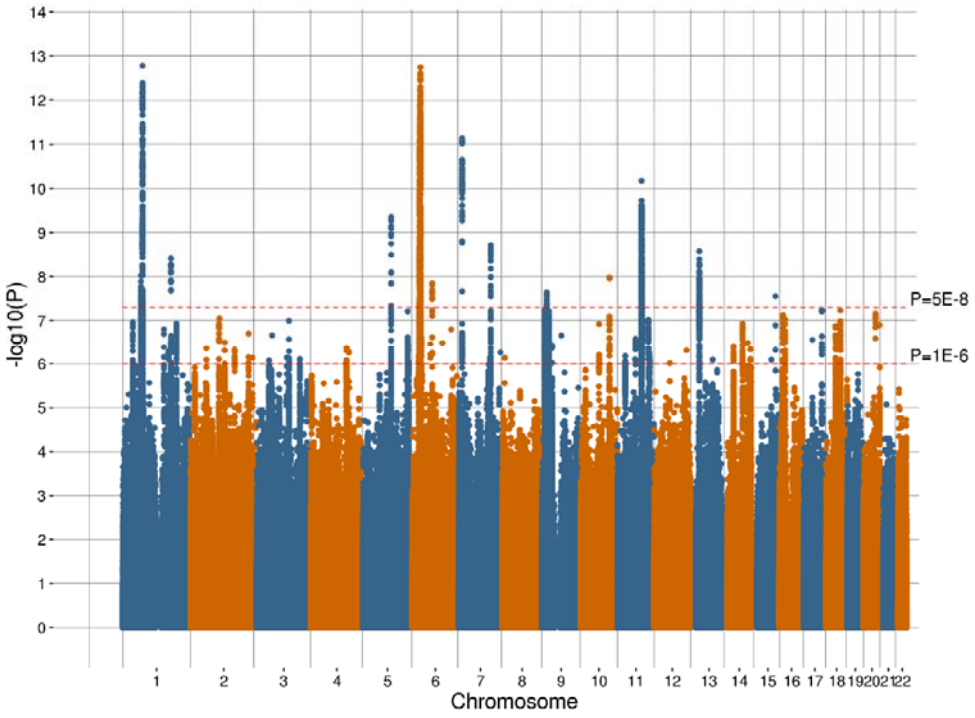

Figure 1. Manhattan plot of the observed $-\log_{10}$ *P*-values of each variant for an association with broad depression in the UK Biobank cohort. Variants are positioned according to the GRCh37 assembly.

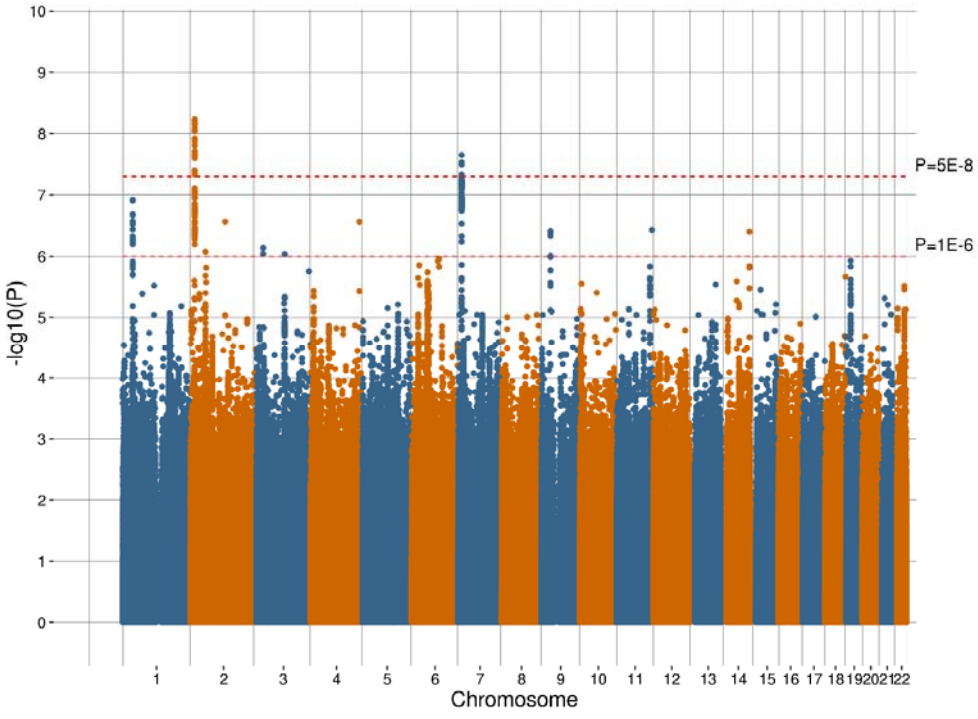

Figure 2. Manhattan plot of the observed $-\log_{10}$ *P*-values of each variant for an association with probable MDD in the UK Biobank cohort. Variants are positioned according to the GRCh37 assembly.

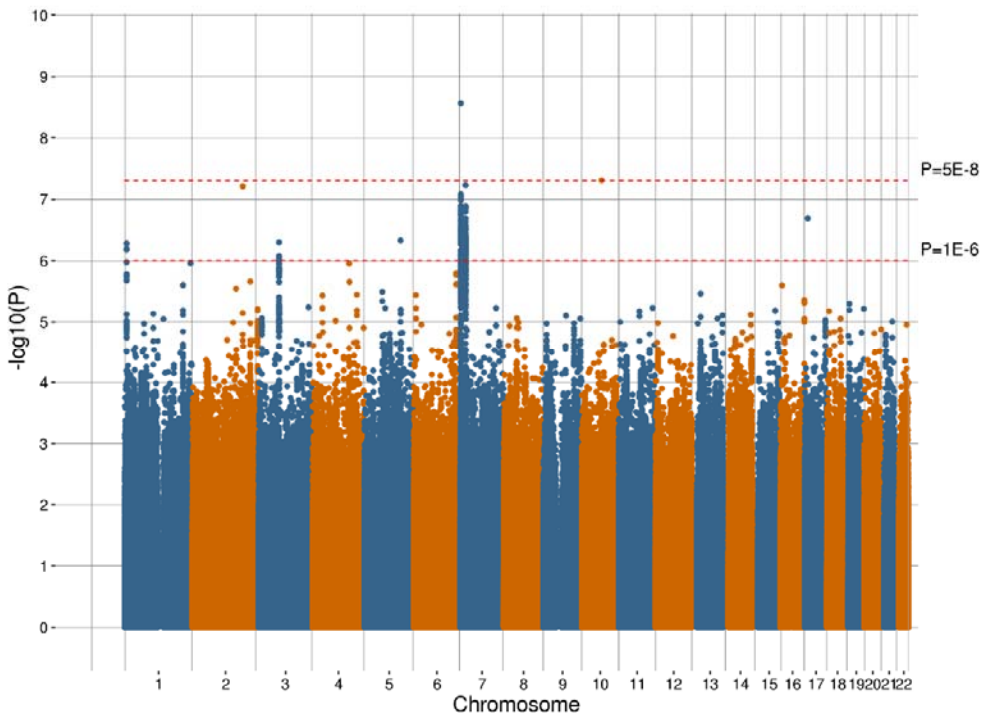

Figure 3. Manhattan plot of the observed $-\log_{10}$ *P*-values of each variant for an association with ICD-coded MDD in the UK Biobank cohort. Variants are positioned according to the GRCh37 assembly.

Gene- and region-based analyses

The MAGMA package was used to identify gene-based regions with a significant effect ($P < 2.72 \times 10^{-6}$) on each trait. This gene-based approach detected 52 significant regions containing 90 genes that were significantly associated with broad depression (Supplementary Table 5), six significant regions containing six genes that were significantly associated with probable MDD (Supplementary Table 6), and three significant regions containing three genes that were significantly associated with ICD-coded MDD (Supplementary Table 7).

MAGMA was also used to identify genomic regions, defined by recombination hotspots, with a statistically significant effect ($P < 6.01 \times 10^{-6}$) on each phenotype. There were 57 significant regions identified for broad depression and a further six regions identified for probable MDD and five regions for ICD-coded MDD. Further details regarding these regions are provided in Supplementary Tables 8, 9, and 10 for broad depression, probable MDD, and ICD-coded MDD, respectively.

Manhattan plots of all the gene/regions analysed are provided in Supplementary Figures 4, 5, and 6 for broad depression, probable MDD, and ICD-coded MDD, respectively.

Pathway analysis

Gene-set enrichment analysis identified five significant pathways for broad depression after applying correction for multiple testing; excitatory synapse ($P_{corrected}$ = 0.006, beta = 0.434, s.e. = 0.090) and neuron spine ($P_{corrected}$ = 0.026, beta = 0.324, s.e. = 0.072) (Table 6). No significant pathways (P > 0.05) were associated with probable MDD or ICD-coded MDD after multiple testing correction.

Table 6. Pathways with a significant effect ($P_{corrected}$ < 0.05) on broad depression following multiple testing correction identified through gene-set enrichment analysis

Phenotype	Pathway	Number of genes	Beta	Standard error	*P*-value	$P_{Corrected}$
Broad depression	Excitatory synapse	114	0.434	0.090	7.35×10^{-7}	0.0059
	Neuron spine	182	0.324	0.072	3.76×10^{-6}	0.0256

eQTL identification

Across the three phenotypes examined seven variants were identified as potential eQTLs (Supplementary Table 11), of these included 4 variants found to be eQTL for brain expressed genes. rs6699744 is associated with the expression of *RPL31P12* in the cerebellum and rs9530139 is associated with expression of *B3GALTL* in the cortex. rs40465 and rs68141011 are broad eQTLs affecting the expression of a number of zinc finger protein encoding genes across various brain tissues including *ZNF391, ZNF204P, ZNF192P1, ZSCAN31* and *ZSCAN23*.

Discussion

This study describes the largest analysis of depression using a single population-based cohort to date. Up to 322,580 individuals from the UK Biobank cohort were used to test the effect of approximately 20 million genetic variants on three depression phenotypes. A total of 15 independent genome-wide significant ($P < 5 \times 10^{-8}$) variants were identified across the three phenotypes. The broadest definition of the phenotype, broad depression, providing the greatest number of individuals for analysis and also the largest number of significant hits (12 independent variants). The probable MDD phenotype was

obtained using the approach of Smith, et al. [12] within a smaller interim release of 150,000 UK Biobank participants; although this phenotype previously yielded no significantly associated variants. However, with the current data release of over 500,000 participants analysed in this study there were two independent genome-wide significant variants. The strictest phenotype was ICD-based MDD, which was dependent on linked hospital admission records for a primary or secondary diagnosis of a mood disorder, had one independent significant variant.

The three UK Biobank phenotypes for depression all had significant genetic correlations ($r_g \geq 0.666$ (0.122), $P \leq 5.34 \times 10^{-7}$) with the results from a mega-analysis of MDD[5], based on an anchor set of clinically defined cases. Interestingly, it was the broad depression phenotype that had the highest genetic correlation with that clinically-defined MDD phenotype. Neuroticism and MDD do share similar symptoms and we did identify significant genetic correlations ($P \leq 5.16 \times 10^{-25}$) between these phenotypes, although as expected the genetic correlations were not as high as those with clinically diagnosed MDD. Each UK Biobank phenotype produced different variants and genes that demonstrated an association. This variability in the variants underlying each phenotype indicates that depression phenotypes may differ markedly in their tractability for genetic studies, and that they may also have somewhat different aetiologies.

A genome-wide significant variant associated with broad depression and identified as an eQTL by GTEx on chromosome 1, rs6699744 (72,825,144 Mb; $P = 1.64 \times 10^{-13}$) was close to another significant variant (rs11209948; $P = 8.38 \times 10^{-11}$) at 72,811,904 Mb associated with MDD within the Hyde, et al. [4] study. Both these variants were close to the Neural Growth Regulator 1 (*NEGR1*) gene which was associated with MDD in the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium., et al. [3] study. Another significant variant (rs7548151) for broad depression on chromosome 1, located at 177,026,983 Mb, was found within 30 Kb of the microRNA 488 (*MIR488*) coding region. *MIR488* is a brain-enriched miRNA that has been implicated in post-transcriptional regulation of gene expression affecting both stability and translation of mRNAs. Transcriptome analysis has demonstrated that altered expression of *MIR488* is nominally associated with stress response and panic disorder[21].

The most proximal gene coding region to the significant variant (rs40465) for broad depression on chromosome 5 was RNU6-334P, which is a pseudogene. Although this variant is located within a region that contains no protein-coding sequences and has no known biological function, the region has been associated with depression and depressive symptoms[7,22]. The GTEx analysis identified this variant as an eQTL for brain expressed genes. A significant variant (rs1021363, 106,610,839 Mb, $P = 1.04 \times 10^{-8}$) on chromosome 10 was associated with broad depression in our study and was within 4 Kb of another variant (rs10786831; 106,614,571 Mb; $P = 8.11 \times 10^{-9}$) found to be associated within the Hyde, et al. [4] MDD study and close to a variant (rs61867293, 106,563,924 Mb, $P = 7.0 \times 10^{-10}$) associated with MDD in the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium., et al. [3] study.

There were up to 16 variants in the gene-rich MHC region that could have been classified as independent however the complexity of the genetic architecture across this region may confounded this. Therefore the authors took the decision to report only the most significant variant (rs3094054). The MHC region has been associated with both schizophrenia and bipolar disorder across multiple studies[23-25], as well as an early-onset and recurrent form of depression[26]. A closer examination of the MHC region is certainly warranted with regarding psychiatric disease based on previous studies and the results obtained in this paper.

There were two genome-wide significant variants located on chromosome 11 that overlap the glutamate metabotropic receptor 5 (*GRM5*) protein coding gene. *GRM5* is expressed in the brain and facilitates glutamatergic neurotransmission. *GRM5* has previously been associated with a range of behavioural and neurological phenotypes such as depression[27], OCD[28], epilepsy[29], smoking[30], Alzheimer's disease[31,32], autism[33] and schizophrenia[34]. A recent study found a role for Metabotropic glutamate receptor 5 (*mGluR5*) in relation to stress-induced depression in mice[35] and *GRM5* antagonists have been shown to have anxiolytic and anti-depressant properties[36,37].

A variant on chromosome 7, rs1554505, was associated with ICD-coded MDD. This variant is located in the MAD1 mitotic arrest deficient-like 1 (*MAD1L1*) gene coding region. *MAD1L1* is a known

susceptibility locus for schizophrenia[38-40] and a recent study showed that differential reward processing during an fMRI task in carriers of a *MAD1L1* bipolar risk allele [41].

Gene-based analyses

The gene-based analysis identified a total of 97 genome-wide significant ($P < 2.72 \times 10^{-6}$) genes across the three phenotypes. Beta 3-glucosyltransferase (*B3GALTL*) and transmembrane protein 106B (*TMEM106B*)) were identified in both the broad depression and probable MDD, and the broad depression and ICD-coded MDD analyses, respectively.

The gene-based analysis of the ICD-coded MDD phenotype highlighted the Ankyrin Repeat and SOCS Box Containing 1 (*ASB1*) protein coding gene. This gene is involved in multiple pathways, including the innate immune system and Class I MHC mediated antigen processing and presentation (Antigen processing- Ubiquitination and Proteasome degradation). Emeny, et al. [42] found an association between increased CpG methylation of the promoter of *ASB1* and anxiety with severe anxiety associated with an increase of almost 50% of CpG methylation in the promoter of ASB1.

Region-based analysis

Our region-based analysis identified 57 genome-wide significant ($P < 6.01 \times 10-6$) regions across the three phenotypes, with four of these regions detected in more than one phenotype. The region-based method detected regions harbouring several known genes that are reported to have effect on depression and other mental diseases that were not detected in our gene-based analysis.

The region-based analysis across all three phenotypes detected a significant region on chromosome 1 that contained the glutamate ionotropic receptor kainate type subunit 3 (*GRIK3*) protein coding region. Glutamate receptors are the main excitatory neurotransmitter receptors in the mammalian brain. Moreover, these receptors are active in several neurophysiologic processes. *GRIK3* has been associated with schizophrenia[43,44], neuroticism[45] and recurrent MDD[46,47]. Higher levels of *GRIK3* have been reported in MDD suicides compared to MDD non-suicides with *GRIK3* expression a strong predictor of suicide[48].

The analysis of broad depression and ICD-coded MDD both detected a significant region containing the Receptor tyrosine-protein kinase erbB-4X (*ERBB4*) protein coding region. *ERBB4* is a member of the Tyr protein kinase family and the epidermal growth factor receptor subfamily. *ERBB4* has been previously linked to schizophrenia[49] and impairments in the link between Neuregulin 1 (*NRG1*) and *ERBB4* signalling are associated with schizophrenia [50] and anxiety behaviours[51].

The broad depression analysis identified a region containing the dihydropyrimidine dehydrogenase (*DPYD*) gene coding region which has been associated with schizophrenia and bipolar disorder[52] and borderline personality disorder[53]. Also identified were regions containing the Neurexin 1 (*NRXN1*) gene coding region which has been associated with Tourette syndrome[54] and non-syndromic autism spectrum disorder[55] and the regulator of G protein signalling 6 (*RGS6*) coding region which has been previously associated with alcoholism[56], depression/anxiety[57] and Parkinson's disease[58].

Pathway analysis

Two gene-sets were significantly enriched in broad depression, both of which were cellular components (where the genes are active) and associated with parts of the nervous system and demonstrates that different gene activity in these components could be attributing to depression.

Our study analysed a large single population-based cohort and replication of the significant genes and variants was not sought. Follow-up analyses should be undertaken in additional cohorts to provide validation of the results obtained in this study. Although useful for studies of this kind, each of the three MDD phenotypes have limitations. None are based on a formal structured diagnostic assessment (such as the Structured Clinical Interview for DSM Axis 1 Disorders interview) and both the broad depression and probable MDD phenotypes are based on self-reported information, which can be subject to recall biases. Broad depression is also likely to be endorsed by a wider range of individuals than traditional depression definitions, including those with internalising disorders other than depression and those with depressive symptoms that would not meet diagnostic criteria for MDD. The ICD-coded MDD phenotype is based on hospital admission records, which can sometimes be incomplete. The size of the full UK Biobank cohort is close to half a million

participants; however, the size of an unrelated sample is approximately a third smaller. One solution to this is to fit a genomic relationship matrix to account for the covariance between individuals. However, it is currently computational challenging to generate and fit within a model a square matrix with the dimensions of 500,000 x 500,000.

Conclusion

In a large genome-wide analysis of a broad depression phenotype in UK Biobank, we identified 15 risk variants implicating perturbations of excitatory neurotransmission in depression and high genetic correlations with more comprehensive interview based methods. These findings suggest that a broad depression phenotype may provide a more tractable target for future genetic studies, allowing the inclusion of many more samples. These findings also provide new genetic instruments for the discovery of disease mechanisms, pharmacological treatments and potentially modifiable factors.

Acknowledgements

This research has been conducted using the UK Biobank Resource – application number 4844. We are grateful to the UK Biobank and all its voluntary 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). All participants gave full informed written consent.

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, IJD and T-KC acknowledge support from the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z and the Dr Mortimer and Theresa Sackler Foundation. This investigation represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. DJS supported by Lister Institute Prize Fellowship 2016-2021.

Competing interests

IJD is a participant in UK Biobank. The authors report that no other conflicts of interest exist.

References

1. World Health Organization. Depression and other common mental disorders. (2017).
2. Sullivan, P.F., Neale, M.C. & Kendler, K.S. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry* **157**, 1552-1562 (2000).
3. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium., Wray, N.R. & Sullivan, P.F. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv* (2017).
4. Hyde, C.L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* **48**, 1031-1036 (2016).
5. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* **18**, 497-511 (2013).
6. Converge consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588-591 (2015).
7. Hek, K. *et al.* A Genome-Wide Association Study of Depressive Symptoms. *Biological psychiatry* **73**, 10.1016/j.biopsych.2012.09.033 (2013).
8. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLOS Computational Biology* **11**, e1004219 (2015).
9. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
10. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).
11. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
12. 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).
13. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
14. Luciano, M. *et al.* 147 independent genetic variants influence the neuroticism personality trait in over 329,000 UK Biobank individuals. (In Prep).
15. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
16. The Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969-976 (2011).
17. The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
18. Shirali, M. *et al.* Using haplotype mapping to uncover the missing heritability: a simulation study. in *10th World Congress in Genetics Applied to Livestock Production* (2014).
19. Zeng, Y. *et al.* Genome-wide regional heritability mapping identifies a locus within the TOX2 gene associated with major depressive disorder. *Biological Psychiatry* (2016).
20. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**, D331-D338 (2017).

21. Muiños-Gimeno, M. *et al.* Human microRNAs miR-22, miR-138-2, miR-148a, and miR-488 Are Associated with Panic Disorder and Regulate Several Anxiety Candidate Genes and Related Pathways. *Biological Psychiatry* **69**, 526-533 (2011).
22. Dunn, E.C. *et al.* Genetic determinants of depression: Recent findings and future directions. *Harvard review of psychiatry* **23**, 1-18 (2015).
23. Bergen, S.E. *et al.* Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry* **17**, 880-886 (2012).
24. Ruderfer, D.M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* **19**, 1017-1024 (2014).
25. Sleiman, P. *et al.* GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus. **3**, 3075 (2013).
26. Shyn, S.I. *et al.* Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry* **16**, 202-215 (2011).
27. Chandley, M.J. *et al.* Elevated gene expression of glutamate receptors in noradrenergic neurons from the locus coeruleus in major depression. *International Journal of Neuropsychopharmacology* **17**, 1569-1578 (2014).
28. Akkus, F. *et al.* Metabotropic glutamate receptor 5 binding in patients with obsessive-compulsive disorder. *International Journal of Neuropsychopharmacology* **17**, 1915-1922 (2014).
29. Kandratavicius, L. *et al.* Distinct increased metabotropic glutamate receptor type 5 (mGluR5) in temporal lobe epilepsy with and without hippocampal sclerosis. *Hippocampus* **23**, 1212-1230 (2013).
30. Hulka, L.M. *et al.* Smoking but not cocaine use is associated with lower cerebral metabotropic glutamate receptor 5 density in humans. *Mol Psychiatry* **19**, 625-632 (2014).
31. Tsamis, K.I., Mytilinaios, D.G., Njau, S.N. & Baloyannis, S.J. Glutamate receptors in human caudate nucleus in normal aging and Alzheimer's disease. *Current Alzheimer Research* **10**, 469-475 (2013).
32. Haas, L.T. *et al.* Metabotropic glutamate receptor 5 couples cellular prion protein to intracellular signalling in Alzheimer's disease. *Brain* **139**, 526-546 (2016).
33. Fatemi, S.H., Folsom, T.D., Kneeland, R.E. & Liesch, S.B. Metabotropic glutamate receptor 5 upregulation in children with autism is associated with underexpression of both Fragile X mental retardation protein and GABA(A) receptor beta 3 in adults with autism. *Anatomical record (Hoboken, N.J. : 2007)* **294**, 10.1002/ar.21299 (2011).
34. Matosin, N. *et al.* Alterations of mGluR5 and its endogenous regulators Norbin, Tamalin and Preso1 in schizophrenia: towards a model of mGluR5 dysregulation. *Acta Neuropathologica* **130**, 119-129 (2015).
35. Shin, S. *et al.* mGluR5 in the nucleus accumbens is critical for promoting resilience to chronic stress. *Nat Neurosci* **18**, 1017-1024 (2015).
36. Tatarczyńska, E. *et al.* Potential anxiolytic- and antidepressant-like effects of MPEP, a potent, selective and systemically active mGlu5 receptor antagonist. *British Journal of Pharmacology* **132**, 1423-1430 (2001).
37. Spooren, W., Lesage, A., Lavreysen, H., Gasparini, F. & Steckler, T. Metabotropic glutamate receptors: their therapeutic potential in anxiety. in *Behavioral Neurobiology of Anxiety and Its Treatment* (eds. Stein, M.B. & Steckler, T.) 391-413 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).
38. O'Dushlaine, C. *et al.* Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatry* **16**, 286-292 (2011).
39. Su, L. *et al.* Genetic association of GWAS-supported MAD1L1 gene polymorphism rs12666575 with schizophrenia susceptibility in a Chinese population. *Neuroscience Letters* **610**, 98-103 (2016).
40. Ripke, S. *et al.* Genome-wide association analysis identifies 14 new risk loci for schizophrenia. *Nature genetics* **45**, 1150-1159 (2013).

41. Trost, S. *et al.* Investigating the impact of a genome-wide supported bipolar risk variant of MAD1L1 on the human reward system. *Neuropsychopharmacology* **41**, 2679-2687 (2016).
42. Emeny, R.T. *et al.* Anxiety associated increased CpG methylation in the promoter of asb1: a translational approach evidenced by epidemiological and clinical studies and a murine model. *Neuropsychopharmacology* (2017).
43. Djurovic, S. *et al.* A possible association between schizophrenia and GRIK3 polymorphisms in a multicenter sample of Scandinavian origin (SCOPE). *Schizophrenia Research* **107**, 242-248 (2009).
44. Greenwood, T.A. *et al.* Genetic assessment of additional endophenotypes from the Consortium on the Genetics of Schizophrenia Family Study. *Schizophrenia Research* **170**, 30-40 (2016).
45. Smith, D.J. *et al.* Genome-wide analysis of over 106.000 individuals identifies 9 neuroticism-associated loci. *Mol Psychiatry* **21**, 749-757 (2016).
46. Luciano, M. *et al.* Association of existing and new candidate genes for anxiety, depression and personality traits in older people. *Behavior Genetics* **40**, 518-532 (2010).
47. Schiffer, H.H. & Heinemann, S.F. Association of the human kainate receptor GluR7 gene (GRIK3) with recurrent major depressive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **144B**, 20-26 (2007).
48. Gray, A.L., Hyde, T.M., Deep-Soboslay, A., Kleinman, J.E. & Sodhi, M.S. Sex differences in glutamate receptor gene expression in major depression and suicide. *Mol Psychiatry* **20**, 1057-1068 (2015).
49. Law, A.J., Kleinman, J.E., Weinberger, D.R. & Weickert, C.S. Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. *Human Molecular Genetics* **16**, 129-141 (2007).
50. Li, B., Woo, R.-S., Mei, L. & Malinow, R. The neuregulin-1 receptor ErbB4 controls glutamatergic synapse maturation and plasticity. *Neuron* **54**, 583-597 (2007).
51. Bi, L.-L. *et al.* Amygdala NRG1-ErbB4 is critical for the modulation of anxiety-like behaviors. *Neuropsychopharmacology* **40**, 974-986 (2015).
52. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* **45**, 1150-9 (2013).
53. Witt, S.H. *et al.* Genome-wide association study of borderline personality disorder reveals genetic overlap with bipolar disorder, major depression and schizophrenia. *Transl Psychiatry* **7**, e1155 (2017).
54. Huang, A.Y. *et al.* Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron* **94**, 1101-1111.e7 (2017).
55. Onay, H. *et al.* Mutation analysis of the NRXN1 gene in autism spectrum disorders. *Balkan J Med Genet* **19**, 17-22 (2016).
56. Stewart, A. *et al.* Regulator of G protein signaling 6 is a critical mediator of both reward-related behavioral and pathological responses to alcohol. *Proc Natl Acad Sci U S A* **112**, E786-95 (2015).
57. Stewart, A. *et al.* Regulator of G-protein signaling 6 (RGS6) promotes anxiety and depression by attenuating serotonin-mediated activation of the 5-HT(1A) receptor-adenylyl cyclase axis. *FASEB J* **28**, 1735-44 (2014).
58. Bifsha, P., Yang, J., Fisher, R.A. & Drouin, J. Rgs6 is required for adult maintenance of dopaminergic neurons in the ventral substantia nigra. *PLoS Genet* **10**, e1004863 (2014).