



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Identifying endophenotypes for depression in Generation Scotland: A Scottish Family Health Study

Lynsey Hall



*Submitted to the University of Edinburgh for the degree of
Doctor of Philosophy*

December 2016
Division of Psychiatry
Kennedy Tower
Royal Edinburgh Hospital
Morningside Park
Edinburgh, EH10 5HF

This work is dedicated to the memory of my late grandpa, Dr. Noel Brian Hopper (1921 - 2006), who always inspired me to do my best through his words of encouragement ("Top of the class!") and stories from his own life and career.

Acknowledgements

I would like to take this opportunity to thank several people for their contribution to my academic development and emotional wellbeing over the past 4 years. Firstly, I would like to thank my supervisors Andrew McIntosh, Chris Haley and Heather Whalley for their intellectual insight, willingness to make time for discussions, and enthusiasm - I have thoroughly enjoyed and appreciated working with you all. I would also like to thank my other colleagues at the Kennedy Tower and IGMM for making the Ph.D. process an enjoyable experience, with particular thanks to Yan-Ni Zeng for her positive outlook and Perl expertise, and Toni-Kim Clarke - I really couldn't have asked for a better or more entertaining office mate!

Outwith my colleagues, I would like to give a massive thanks to my Mum, Dad, brother Stuart and partner Mike for putting up with my elevated stress levels for the past 4 years. I really appreciate all of the phone calls, texts, emails and visitations. And although they cannot read, thank you to my cats Monty and Piggy for upping their purr and cuddle quota as required.

Declaration

I declare that this thesis has been composed by me, that the work described in this thesis is my own (except where otherwise stated), and that the work described in this thesis has not been submitted for any other degree or personal qualification.

Lynsey Hall

Publications

Hall, L.S., Adams, M.J., Clarke, T-K., Howards, D., Arnau-Soler, A., Fernandez-Pujals, A-M., Zeng, Y-N., Thomson, P.A., Hayward, C., MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Haley, C.S., Porteous, D.J., Deary, I.J., McIntosh, A.M. Characterizing The Genetic Basis of Major Depressive Disorder in Generation Scotland: The Scottish Family Health Study. (In preparation)

Hall, L.S., Clarke, T-K., Fernandez-Pujals, A.M., Zeng, Y-N., Thomson, P.A., Hayward, C., MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Haley, C.S., Porteous, D.J., Deary, I.J., McIntosh, A.M. Identifying endophenotypes for depression - a composite trait analysis. (In preparation)

Hall, L.S., Zeng, Y-N., Clarke, T-K., Adams, M.J., Howards, D., Fernandez-Pujals, A-M., Thomson, P.A., Hayward, C., MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Haley, C.S., Porteous, D.J., Deary, I.J., McIntosh, A.M. Dissecting the genetic and environmental basis of smoking and its association with depression and schizophrenia in the Scottish population. (In preparation)

Whalley, H.C., Adams, M.J., **Hall, L.S.**, Clarke, T-K., Fernandez-Pujals, A.M., Gibson, J., Wigmore, E., Hagenaars, S.P., Davies, G., Campbell, A., Hayward, C., Lawrie, S.M., Porteous, D.J., Deary, I.J., McIntosh, A.M. (2016). Dissection of Major Depressive Disorder using polygenic risk scores for Schizophrenia in two independent cohorts. *Translational Psychiatry* 6(11):e938. doi:10.1038/tp.2016.207

McIntosh, A.M., **Hall, L.S.**, Zeng, Y-N., Adams, M.J., Gibson, J., Wigmore, E., Fernandez-Pujals, A.M., Campbell, A.C., Clarke, T-K., Hayward, C., Haley, C.S., Porteous, D.J., Deary, I.J., Smith, D.J., Nicholl, B.I., Hinds, D.A., Jones, A.V., Scollen, S., Meng, W., Smith, B.H., Hocking, L.J. (2016). Genetic and environmental risk for chronic pain and the contribution of risk variants for psychiatric disorders. Results from Generation Scotland: Scottish Family Health Study and UK Biobank. *PLoS Medicine* (accepted)

Power, R.A., Tansey, K.E., Buttenschøn, H., Cohen-Woods, S., Bigdeli, T., **Hall, L.S.**, Kutalik, Z., Lee, S.H., Ripke, S., Steinberg, S., Teumer, A., Viktorin, A., Wray, N.R., Arolt, V., Baune, B.T., Boomsma, D.I., Børglum, A.D., Byrne, E.M., Castelao, E., Craddock, N., Craig, I., Dannlowski, U., Deary, I.J., Degenhardt, F., Forstner, A.J., Gordon, S.D., Grabe, H.J., Grove, J., Hamilton, S.P., Hayward, C., Heath, A.C., Hocking, L.J., Homuth, G., Hottenga, J.J., Kloiber, S., Krogh, J., Landén, M., Lang, M., Levinson, D.F., Lichtenstein, P., Lucae, S., MacIntyre, D.J., Madden, P., Magnusson, P.K.E., Martin, N.G., McIntosh, A.M., Middeldorp, C.M., Milaneschi, Y., Montgomery, G.W., Mors, O., MüllerMyhsok, B., Nyholt, D.R., Oskarsson, H., Owen, M.J., Padmanabhan, S., Penninx, B.W.J.H., Pergadia, M.L., Porteous, D.J., Potash, J.B., Preisig, M., Rivera-Sanchez, M., Shi, J., Shyn, S., Sigurdsson, E., Smit, J.H., Smith, B.H., Stefansson, K., Stefansson, H., Strohmaier, J., Sullivan, P.F., Thomson, P.A., Thorgeirsson, T.E., Van der Auwera, S., Weissman, M.M., CONVERGE consortium, CARDIoGRAM Consortium, GERAD1 Consortium, Breen, G., and Lewis, C.M. (2016). Genome-wide association for major depression through age at onset stratification. The Psychiatric Genomics Consortium Major Depressive Disorder Working Group. *Biol Psychiatry*. (accepted)

Clarke, T-K., Obsteter, J., **Hall, L.S.**, Hayward, C., Thomson, P.A., Smith, B.H., Padmanabhan, S., Hocking, L.J., Deary, I.J., Porteous, D.J., McIntosh, A.M. (2016). Investigating shared aetiology between Type 2 Diabetes and Major Depressive Disorder in a population based cohort. *Am J Med Genet B Neuropsychiatr Genet*. (accepted)

Lawrie, S., Thomson, P.A., Duff, B., Blackwood, D., Romaniuk, L., Watson, A., Whalley, H., Xang, L., Dauvermann, M., Moorhead, T.W., Bois, C., Ryan, N., Morris, S., Redpath, H., **Hall, L.S.**, van Beek, E., Roberts, N., Porteous, D.J., St. Clair, D., Whitcher, B., Dunlop, J., Brandon, N., Hughes, Z., Hall, J., McIntosh, A.M. (2016). Balanced translocation linked to psychiatric disorder, glutamate and cortical structure/function. *npj Schizophrenia*. (accepted)

Zeng, Y-N, Navarro, P., Fernandez-Pujals, A.M., **Hall, L.S.**, Clarke, T-K., Thomson, P.A., Smith, B.H., Hocking, L.J., Padmanabhan, S., Hayward, C., MacIntyre, D.J., Wray, N.R., Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Deary, I.J., Porteous, D.J., Haley, C.S., McIntosh, A.M. (2016). A combined pathway and regional heritability analysis indicates NETRIN1 pathway is associated with Major Depressive Disorder. *Biol Psychiatry* (accepted).

Hagenaars, S.P., Harris, S.E., Clarke, T-K., **Hall, L.S.**, Luciano, M., Fernandez-Pujals, A.M., Davies, G., Hayward, C., Generation Scotland, Starr, J.M., Porteous, D.J., McIntosh, A.M., Deary, I.J. (2016). Polygenic Risk for Coronary Artery Disease is Associated with Cognitive Ability in Older Adults. *Int J Epidemiol* (accepted)

Aas, M., Blokland, G., Chawner, S., Choi, S-W., Estrada, J., Forsingdal, A., Friedrich, M., Ganesham, S., **Hall, L.S.**, Haslinger, D., Huckins, L., Loken, E., Malan-Müller, S., Martin, J., Misiewicz, Z., Pagliaroli, L., Pardiñas, A., Pisanu, C., Quadri, G., Santoro, M., Shaw, A., Ramlund, S., Song, J., Tesli, M.S., Tropeano, M., van der Voet, M., Wolfe, K., Cormack, F., DeLisi, L. (2016). Summaries of Plenary, Symposia, and Oral sessions at the XXII World Congress of Psychiatric Genetics, Copenhagen, Denmark, October 12 – 16, 2014. *Psychiatr Genet*. [Epub ahead of print]

Clarke, T-K., **Hall, L.S.**, Fernandez-Pujals, A.M., MacIntyre, D.J., Thompson, P.A., Hayward, C., Smith, B.H., Padmanabhan, S., Hocking, L.J., Deary, I.J., Porteous, D.J., McIntosh, A.M. (2015). Major Depressive Disorder (MDD) and current psychological distress moderate the effect of polygenic risk for obesity on body mass index. *Translational Psychiatry*. **5**. e592. doi:10.1038/tp.2015.83

Clarke, T-K., Smith, A.H., Gelernter, J., Kranzler, H.R., Farrer, L.A., **Hall, L.S.**, Fernandez-Pujals, A.M., MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Hayward, C., Thomson, P.A., Porteous, D.J., Deary, I.J., McIntosh, A.M. (2015). Polygenic risk for alcohol dependence associates with alcohol consumption, cognitive function, education and social deprivation in a population-based cohort. *Addict Biol*. doi: 10.1111/adb.12245. [Epub ahead of print]

Clarke, T-K., Lupton, M., Fernandez-Pujals, A.M., Starr, J., Davies, G., Cox, S., Pattie, A., Liewald, D., **Hall, L.S.**, MacIntyre, D.J., Smith, B.H., Hocking, L.J., Padmanabhan, S., Thomson, P.A., Hayward, C., Hansell, N., Montgomery, G., Medland, S., Martin, N., White, M., Porteous, D.J., Deary, I.J., McIntosh, A.M. (2015). Common polygenic risk for Autism Spectrum Disorders (ASD) is associated with cognitive ability in the general population. *Mol. Psychiatry*. doi: 10.1038/mp.2015.12. [Epub ahead of print]

Whalley, H.C., **Hall, L.S.**, Romaniuk, L., MacDonald, A., Lawrie, S.M., Sussmann, J.E., McIntosh, A.M. (2014). Impact of cross-disorder polygenic risk on frontal brain activation, with specific effect of schizophrenia risk. *Schizophr. Res.* 161 (2-3). 484-489.

Whalley, H.C., Sprooten, E., Hackett, S., **Hall, L.S.**, Blackwood, D.H., Glahn, D.C., Bastin, M., Hall, J., Lawrie, S.M., Sussmann, J.E., McIntosh, A.M. (2013). Polygenic risk and white matter integrity in individuals at high risk of mood disorder. *Biol Psychiatry*. 74(4). 280-6.

Brydges, N.M., **Hall, L.S.** A shortened protocol for assessing cognitive bias in rats. (In review)

Zeng, Y-N., Navarro, P., Shirali, M., Howard, D., Adams, M.J., Hall, L.S., Clarke, T-K, Thomson, P.A., Smith, B.H., Murray, A.D., Padmanabhan, S., Hayward, C., Boutin, T., MacIntyre, D.J., Lewis, C.M., Mehta, D., Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Deary, I.J., Porteous, D.J., Haley, C.S., McIntosh, A.M. Genome-wide regional heritability mapping identifies a locus within the TOX2 gene associated with Major Depressive Disorder. (In review)

Howard, D.M., **Hall, L.S.**, Hafferty, J.D., Zeng, Y-N., Adams, M.J., McIntosh, A.M. The identification of haplotypes associated with major depressive disorder. (In review)

Witt, S.H., Streit, F., Jungkunz, M., Frank, J., Awasthi, S., Reinbold, C.S., Treutlein, J., Degenhardt, F., Forstner, A.J., Heilmann-Heimbach, S., Dietl, L., Schwarze, C.E., Schendel, D., Strohmaier, J., Abdellaoui, A., Adolfsson, R., Air, T.M., Alda, M., AllieyRodriguez, N., Andreassen, O.A., Babadjanova, G., Bass, N.J., Bauer, M., Baune, B.T., Bellivier, F., Bergen, S., Bethell, A., Biernacka, J.M., Blackwood, D.H.R., Boks, M., Boomsma, D.I., Børglum, A.D., Borrmann-Hassenbach, M., Brennan, P., Buttenschøn, H.N., Byrne, E.M., Cervantes, P., Clarke, T-K., Craddock, N., Cruceanu, C., Curtis, D., Czerski, P.M., Dannlowski, U., Davis, T., de Geus, E.J.C., Di Florio, A., Djurovic, S., Domenici, E., Edenberg, H.J., Etain, B., Fischer, S.B., Forty, L., Fraser, C., Frye, M.A., Fullerton, J.M., Gershon, E.S., Giegling, I., Gordon, S.D., Gordon-Smith, K., Grabe, H.J., Green, E.K., Greenwood, T.A., GrigoriuSerbanescu, M., Guzman-Parra, J., **Hall, L.S.**, Hamshere, M., Hauser, J., Hautzinger, M., Herms, S., Hitturlingappa, S., Hoffmann, P., Holmans, P., Hottenga, J-J., Jamain, S., Jones, I., Jones, L.A., Juréus, A., Kahn, R.S., Kammerer-Ciernioch, J., Kirov, G., KittelSchneider, S., Kloiber, S., Knott, S.V., Kogevinas, M., Landén, M., Leber, M., Leboyer, M., Li, Q.S., Lissowska, J., Lucae, S., Martin, N.G., Mayoral-Cleries, F., McElroy, S.L., McIntosh, A.M., McKay, J.D., McQuillin, A., Medland, S.E., Middeldorp, C.M., Milaneschi, Y., Mitchell, P.B., Montgomery, G.Q., Morken, G., Mors, O., Mühleisen, T.W., Müller-Myhsok, B., Nievergelt, C.M., Nurnberger, J.I. Jr, O'Donovan, M.C., Loohuis, L.M.O., Ophoff, R., Oruc, L., Owen, M.J., Paciga, S.A., Penninx, B.W.J.H., Perry, A., Pfennig, A., Potash, J.B., Preisig, M., Reif, A., Rivas, F., Rouleau, G.A., Schofield, P.R., Schulze, T.G., Schwarz, M., Sinnamon, G.C.B., Stahl, E.A., Strauss, J., Turecki, G., Van der Auwera, S., Vedder, H., Vincent, J.B., Willemsen, G., Witt, C.C., Wray, N.R., Xi, H.S., Bipolar Disorders Working Group of the Psychiatric Genomics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Schizophrenia Working Group of the Psychiatric Genomics 2 Consortium, Tadic, A., Dahmen, N., Schott, B.H., Cichon, S., Nöthen, M.M., Ripke, S., Mobascher, A., Rujescu, D., Lieb, K., Roepke, S., Schmahl, C., Bohus, M., Rietschel, M. Genome-wide association study of Borderline Personality Disorder reveals genetic overlap with the Major Psychoses. (In review)

Bigdeli, T., Ripke, S., Bacanu, S., Peterson, R., Abdellaoui, A., Andlauer, T., Beekman, A., Berger, K., Blackwood, D., Boomsma, D., Breen, G., Buttenschoøn, H., Byrne, E., Cichon, S., Clarke, T-K., Couvy-Duchesne, B., Craddock, N., de Geus, E., Degenhardt, F., Dunn, E., Fanous, A., Forstner, A., Frank, J., Gill, M., Gordon, S., Grabe, H., **Hall, L.**, Hamilton, S., Hardiman, O., Hayward, C., Heath, A., Henders, A., Herms, S., Hickie, I., Hoffman, P., Homuth, G., Hottenga, J-J., Ising, M., Jansen, R., Kloiber, S., Knowles, J., Lang, M., Li, Q., Lucae, S., MacIntyre, D., Madden, P., Martin, N., McGrath, P., McGuffin, P., McIntosh, A., Medland, S., Mehta, D., Middeldorp, C., Milaneschi, Y., Montgomery, G., Mors, O., Müller-Myhsok, B., Nauck, M., Nyholt, D., Nöthen, M., Owen, M., Penninx, B., Pergadia, M., Perlis, R., Peyrot, W., Porteous, D., Potash, J., Rice, J., Rietschel, M., Riley, B., Rivera, M., Schoevers, R., Schulze, T., Shi, J., Shyn, S., Smit, J., Smoller, J., Streit, F., Strohmaier, J., Teumer, A., Treutlein, J., Van der Auwera, S., van Grootheest, G., van Hemert, A., Völzke, H., Webb, B., Weissman, M., Wellman, J., Willemsen, G., Witt, S., Levinson, D., Lewis, C., Wray, N., Flint, J., Sullivan, P., Kendler, K. Genetic effects influencing risk for major depressive disorder in China and Europe. (In review)

Wigmore, E.M., Clarke, T-K., Adams, M.J., Fernandez-Pujals, A.M., Gibson, J., Davies, G., **Hall, L.S.**, Zeng, Y-N., Thomson, P.A., Hayward, C., Smith, B.H., Hocking, L.J., Padmanabhan, S., Deary, I.J., Porteous, D.J., Nicodemus, K.K., McIntosh, A.M. Do Regional Brain Volumes and Major Depressive Disorder Share Genetic Architecture: a study in Generation Scotland, UK Biobank and the English Longitudinal Study of Ageing. (In preparation)

Thomson, P.A., Prendergast, J., Huffman, J., Svinti, V., Fernandez-Pujals, A.M., **Hall, L.S.**, Clarke, T-K., Campbell, A., Liewald, D., Vitart, V., Dunlop, M., Taylor, M., MacIntyre, D.J., Generation Scotland, McIntosh, A.M., Deary, I.J., Blackwood, D., Haley, C.S., Hayward, C., Hastie, N., Porteous, D.J., Wright, A. Linkage of early onset major depressive disorder to chromosomes 12 and 18 with burden of rare variants in NINJ2 and ROCK1. (In preparation)

de Jong, S., Diniz, M.J.A., Rodrigues, A.C.S., Gadelha, A., Santoro, M.L., Ota, V.K., Noto, C.S., Curtis, C., Patel, H., **Hall, L.S.**, O'Reilly, P.F., Belangero, S.I., Bressan, R., Breen, G. Polygenic risk scores reveal patterns of assortative mating and anticipation in a large severely affected pedigree. (In preparation)

Gibson, J., Russ, T.C., Adams, M.J., Clarke, T-K., **Hall, L.S.**, Fernandez-Pujals, A.M., Wigmore, E.M., Hayward, C., Davies, G., Murray, A.D., Smith, B.H., Porteous, D.J., Deary, I.J., McIntosh, A.M. Assessing the presence of shared genetic architecture between Alzheimer's disease and major depressive disorder using genome-wide association data. (In preparation)

Neilson, E., Bois, C., Clarke, T-K., **Hall, L.S.**, Johnstone, E.C., Owens, D.G.C., Whalley, H.C., McIntosh, A.M., Lawrie, S.M. Polygenic risk of schizophrenia is associated with transition to illness and cortical gyrification in a familial high-risk sample. (In preparation)

Abstract

Depression is the most common psychiatric disorder and the leading cause of disability worldwide. Despite evidence for a genetic component, the genetic aetiology of this disorder remains elusive. To date, only one association study has identified and replicated risk loci for depression. This thesis focuses on aiding genetic discovery by revisiting the depressed phenotype and developing a quantitative trait, using data from Generation Scotland: The Scottish Family Health Study. These analyses aim to test whether this derived quantitative trait has improved statistical power to identify genetic risk variants for depression, relative to the binary classification of case/control. Measures of genetic covariation were used to evaluate and rank ten measures of mood, personality and cognitive ability as endophenotypes for depression. The highest ranking traits were subjected to principal component analysis, and the first principal component used as a quantitative measure of depression. This composite trait was compared to the binary classification of depression in terms of ability to identify risk loci in a genome-wide association study, and phenotypic variance explained by polygenic profile scores for psychiatric disorders. I also compared the composite trait to the univariate traits in terms of their ability to fulfill the endophenotype criteria as described by Gottesman and Gould, namely: being heritable, genetically and phenotypically correlated with depression, state independent, co-segregating with illness in families, and observed at a higher rate in unaffected relatives than in unrelated controls. Four out of ten traits fulfilled most endophenotype criteria, however, only two traits - neuroticism and the general health questionnaire (a measure of current psychological distress) - consistently ranked highest across all analyses. As such, three composite traits were derived incorporating two, three, or four traits. Association analyses of binary depression, univariate traits and composite traits yielded no genome-wide significant results, with most traits performing equivalently. However, composite traits were more heritable and more highly correlated with depression than their constituent traits, suggesting that analyzing these traits in combination was capturing more of the heritable

component of depression. Polygenic scores for psychiatric disorders explained more trait variance for the composite traits than the univariate traits, and depression itself. Overall, whilst the composite traits generally obtained more significant results, they did not identify any further insight into the genetic aetiology of depression. This work therefore provides support for the urgent need to redefine the depressed phenotype based on objective and quantitative measures. This is essential for risk stratification, better diagnoses, novel target identification and improved treatment.

Lay Summary

Depression is the most common psychiatric disorder, typified by low mood and loss of interest in previously enjoyed activities. Evidence from studies suggests that depression is caused by both genetic and environmental factors. Despite this, the overwhelming majority of studies fail to find any genetic factors related to depression. The purpose of this thesis is to measure depression in a different way to see if this improves our ability to identify genetic factors. In studies of disease, people are categorized as being either with or without the illness of interest. This is problematic in complex disorders like depression as people are not simply ill or well, but on a spectrum of illness. I have created a measurement for depression designed to access information about this spectrum of illness which is lost by only classifying people as ill or well. This measurement was created by combining information from questionnaires about people's mood, personality and cognitive ability. For four out of the ten questionnaire results, how people answered was correlated with whether or not they were depressed. Two questionnaires (which measured neuroticism and current psychological distress) were correlated with having depression and with an elevated risk of having depression (in non-depressed relatives of depressed people). This suggested that these questionnaires would be useful for creating an alternative measure of depression. I made three versions of this measure, incorporating information from two, three or four of the questionnaires. Compared to individual questionnaires, the new measurements had a higher correlation with depression, suggesting that using the questionnaires in combination is more effective than using individual results. Using a technique which estimates an individual's genetic risk for depression, I established that the combined results were more strongly associated with elevated genetic risk than individual results or a diagnosis of depression. However, in analyses designed to identify genetic variants related to depression, none of the measurements had any results that were better than one would expect by chance. These results suggest that "depression" is too nebulous a diagnosis, and future research should focus on looking at different types of depression separately, rather than as one disorder.

Abbreviations

α - significance criterion

A1 - reference allele

A2 - alternate allele

AOO - age of onset

AUT - autism

β - false negative rate

β - beta (regression coefficient)

BDI - Beck Depression Inventory

BIP - bipolar disorder (type 1)

BP - base-pair position

CES-D - Centre for Epidemiological Studies - Depression Scale

cGxE - candidate gene by environment interaction

CHR - chromosome

CI - 95% confidence interval

cov_A - additive covariance

cov_P - phenotypic covariance

DSC - digit symbol coding

DSM-III/IV/V - The Diagnostic and Statistical Manual of Mental Disorders
(3rd/4th/5th edition)

EPQE - Eysenck personality questionnaire: extraversion

EPQN - Eysenck personality questionnaire: neuroticism

FDR - false discovery rate

fMDD - major depressive disorder in females only

GxE - gene by environment interaction

Gen3 - composite trait created using three traits

Gen4 - composite trait created using four traits

GHQ - general health questionnaire

GREML - genomic restricted maximum likelihood

GRM - genetic relationship matrix

GS:SFHS - Generation Scotland: The Scottish Family Health Study

GWAA - genome-wide association analysis

GWAS - genome-wide association study

H_0 - null hypothesis

H_1 - alternative hypothesis

H^2 - broad sense heritability estimate from twin/pedigree studies

h^2 - narrow sense heritability estimate from twin/pedigree studies

h^2_{GWS} - narrow sense heritability estimate from SNPs achieving genome-wide significance

h^2_{SNP} - narrow sense heritability estimate using all SNP data

h_{ij} - coheritability estimate

h_{ij0} - coheritability estimate on the observed scale

ICD-10 - International Statistical Classification of Diseases and Related Health Problems (10th edition)

λ - ratio of observed median χ^2 to that expected by chance

λ_{GC} - genomic control factor

LD - linkage disequilibrium

LDSC - linkage disequilibrium score regression

LM1 - logical memory immediate

LM2D - logical memory delayed

logOR - logarithm of the odds ratio

LRT - likelihood ratio test

$\mu\chi^2$ - mean chi-squared statistic

MAF - minor allele frequency

Mb - megabases

MDD - major depressive disorder

MDQ - mood disorder questionnaire

MDS - multidimensional scaling

MHV - Mill Hill vocabulary

MLM - mixed linear model

MLMA - mixed linear model analysis

MLMe - mixed linear model with candidate marker excluded

mMDD - major depressive disorder in males only

NA - not available (missing data)

OR - odds ratio

P - associated p-value of test statistic

PC - principal component

PCA - principal components analysis

PGC - Psychiatric GWAS Consortium

PGRS - polygenic risk scores

POS - base-pair position

P_T - P-value threshold

QC/QCd - quality control/quality controlled

R^2 - trait variance explained

r^2 - linkage disequilibrium squared correlation coefficient

r_g - genetic correlation

r_p - phenotypic correlation

rMDD - recurrent major depressive disorder

SCID - structured clinical interview for DSM-IV disorders

SCL-90 - Revised Symptom Checklist - 90

SCZ - schizophrenia

SE - standard error

SIMD - Scottish Index of Multiple Deprivation

SLE - stressful life events

SNP - single nucleotide polymorphism

SoS - Sum of Scores, composite trait created using two traits

SPQ - schizotypal personality questionnaire

T2D - type II diabetes

UKB - UK Biobank

VF - verbal fluency

V_A - variance explained by additive genetic effects

V_C - variance explained by common environmental effects

V_E - variance explained by residual effects

V_G - variance explained by all genetic effects

V_P - total phenotypic variance

Contents

Chapter 1: Introduction.....	37
1.1. Background.....	37
1.1.1. Symptoms and diagnosis	37
1.2. Epidemiology of depression.....	38
1.2.1. Evidence for a genetic component	39
1.2.2. Molecular genetics findings	40
1.2.2.1. Linkage studies.....	42
1.2.2.2. Genome-wide association studies	43
1.2.2.3. Candidate gene studies.....	46
1.2.2.4. Gene by environment interactions	46
1.2.2.5. Polygenic profiling.....	49
1.3. Problems facing genetic discovery in depression	52
1.3.1. Statistical power	52
1.3.2. Missing heritability.....	55
1.3.3. Disease heterogeneity	57
1.4. Clinical covariates	62
1.5. The endophenotype concept in psychiatry	64
1.6. Project summary and aims	66
Chapter 2: Methods.....	69
2.1. Sample description	69
2.1.1. Measurement and definition of depression	70
2.1.2. Measurement and definition of mood, personality and cognitive phenotypes.....	71
2.1.3. Quality control	72
2.1.3.1. Phenotypic data	72
2.1.3.2. Genetic data	73

2.1.3.3. Accounting for population stratification	73
2.1.4. Statistical power	78
2.2. Mixed Linear Models Analysis.....	80
2.2.1. Estimating trait heritability.....	81
2.2.1.1. Accounting for shared environment	82
2.2.1.2. Heritability estimates for MDD	86
2.2.1.3. Heritability of quantitative traits.....	86
2.2.3. Estimating genetic and phenotypic between trait correlations.....	87
2.2.3.1. Estimating coheritability between traits.....	87
2.2.4. Estimating between group differences in trait mean	88
2.2.5. Genome-wide association analysis using mixed linear model analysis	88
2.2.6. Polygenic Risk Score Analysis	89
2.3. Multivariate Analysis.....	90
2.4. Replication of results in UK Biobank.....	91
2.4.1. Sample description.....	91
2.4.1.1. Definition of depression phenotype	92
2.4.1.2. Quality control.....	93
2.5.1. LD Score regression	94
Chapter 3: Analysis of depression (as a binary trait) in Generation Scotland and UK Biobank	95
3.1. Introduction	95
3.2. Methods	97
3.2.1. Descriptive statistics.....	97
3.2.2. Heritability of depression.....	97
3.2.2.1. Using pedigree data	98
3.2.2.2. Using GWAA summary statistics	99
3.2.3. Genome-wide association analysis	100

3.2.4. Polygenic profiling analysis	102
3.2.5. Replication in UKBiobank.....	102
3.3. Results	103
3.3.1. Descriptive statistics	103
3.3.2. Heritability of depression.....	106
3.3.2.1. Using pedigree data	106
3.3.2.2. Using GWAA summary statistics	108
3.3.3. Genome-wide association analysis	110
3.3.3.1. Summary statistics from GWAS of MDD	112
3.3.3.2. Summary statistics from GWAS of recurrent MDD	114
3.3.3.3. Summary statistics from GWAS of MDD in females	116
3.3.3.4. Summary statistics from GWAS of MDD in males	118
3.3.3.5. Biological annotation	120
3.3.4. Polygenic profiling analysis	153
3.3.4.1. Association with MDD PGRS	153
3.3.4.2. Association with autism PGRS.....	156
3.3.4.3. Association with bipolar disorder PGRS.....	158
3.3.4.4. Association with schizophrenia PGRS	160
3.3.5. Replication in UKBiobank.....	162
3.3.5.1. Paired samples sign test.....	162
3.3.5.2. Bivariate LD Score regression	163
3.4. Discussion	166
3.4.1. Summary of findings	166
3.4.1.1. Is MDD in GS:SFHS consistent with the literature?.....	166
3.4.1.2. Does basic stratification improve etiological insight?	167
3.4.1.3. Does MDD in GS:SFHS correlate with MDD in UKB?.....	169
3.4.2. Limitations.....	169

3.4.2.1. Model selection.....	169
3.4.2.2. Linear vs. Liability scale	170
3.4.2.3. Calculating variance explained by PGRS.....	171
3.4.3. Conclusions.....	172
Chapter 4: Assessment of mood, personality and cognitive phenotypes as candidate endophenotypes for depression.....	173
4.1. Introduction	173
4.2. Methods	177
4.2.1. Testing endophenotype criteria	177
4.2.1.1. Associated with illness in the population.....	177
4.2.1.2. More extreme in unaffected relatives than in the general population	177
4.2.1.3. State independence	178
4.2.1.4. Heritability of endophenotypes	178
4.2.1.3. Genetic and phenotypic correlation.....	179
4.2.2. Ranking candidate endophenotypes.....	179
4.2.2.1. Using coheritability.....	179
4.2.2.2. Using variance explained by MDD PGRS	180
4.2.2.3. Convergence of ranking methods	181
4.2.3. Performance of endophenotypes relative to binary MDD.....	181
4.2.3.1. Using variance explained by PGRS	181
4.2.3.2. Genome-wide association analysis	181
4.2.4. Replication in UKBiobank	182
4.3. Results	183
4.3.1. Testing endophenotype criteria	183
4.3.1.1. Effect of affection status on trait mean	183
4.3.1.2. Heritability of endophenotypes	185
4.3.1.3. Genetic and phenotypic correlation.....	189

4.3.2. Ranking candidate endophenotypes	190
4.3.2.1. Using coheritability	190
4.3.2.2. Using variance explained by MDD PGRS.....	191
4.3.2.3. Convergence of coheritability measures	195
4.3.3. Performance of endophenotypes relative to binary MDD	197
4.3.3.1. Using variance explained by PGRS	197
4.3.3.2. Genome-wide association analysis.....	209
4.3.4. Replication in UKBiobank.....	230
4.3.4.1. Paired samples sign test.....	230
4.3.4.2. Bivariate LD Score regression	231
4.4. Discussion	233
4.4.1. Summary of findings	233
4.4.1.1. Did quantitative traits fulfill endophenotype criteria?	233
4.4.1.2. Which endophenotypes ranked highest using measures covariation?	234
4.4.1.3. Did endophenotypes "outperform" binary MDD?.....	235
4.4.1.4. Did endophenotypes correlate with MDD in UKB?	236
4.4.2. Limitations.....	237
4.4.2.1. Traits selected for testing endophenotype criteria	237
4.4.2.1. Model selection	237
4.4.3. Conclusions	238
Chapter 5: Using multivariate methods to jointly maximize heritability and genetic correlation with depression	239
5.1. Introduction	239
5.2. Methods.....	240
5.2.1. Creating the composite traits	240
5.2.1.1. Using two traits	240
5.2.1.2. Using more than two traits	240

5.2.2. Testing endophenotype criteria	243
5.2.2.1. Associated with illness in the population	243
5.2.2.2. More extreme in unaffected relatives than in the general population	243
5.2.2.3. State independent	243
5.2.2.4. Heritability of composite traits	243
5.2.2.5. Genetic and phenotypic correlation	244
5.2.3. Performance of composite traits relative to single traits	244
5.2.3.1. Coheritability	244
5.2.3.2. Polygenic risk score analysis	244
5.2.3.3. Genome-wide association analysis	245
5.2.4. Replication in UKBiobank	245
5.3. Results	246
5.3.1. Testing endophenotype criteria	246
5.3.1.1. Effect of affection status on trait mean	246
5.3.1.2. Heritability	247
5.3.1.3. Genetic and phenotypic correlation	248
5.3.2. Performance of composite traits relative to single traits	249
5.3.2.1. Coheritability	249
5.3.2.2. Polygenic risk score analysis	249
5.3.2.3. Genome-wide association analysis	257
5.3.3. Replication in UKBiobank	274
5.3.3.1. Paired samples sign test	274
5.3.3.2. Bivariate LD Score regression	275
5.4. Discussion	276
5.4.1. Summary of findings	276
5.4.1.1. Were composite traits better candidate endophenotypes than single traits?	276

5.4.1.2. Did composite traits rank more highly than single traits using measures of genetic covariation?	276
5.4.1.3. Did composite traits have increased power to detect genetic effects compared to single traits or binary MDD?	277
5.4.2. Limitations.....	277
5.4.2.1. Trait selection for composite traits.....	277
5.4.3. Conclusions	278
Chapter 6: Discussion	279
6.1. Summary.....	279
6.2. Overview of findings	279
6.2.1. Analysis of MDD as a binary trait.....	279
6.2.2. Analysis of quantitative traits	280
6.2.3. Analysis of composite traits	280
6.3. Methodological remarks and limitations	280
6.3.1. Model selection	280
6.3.1.1. Mixed linear model	280
6.3.3.1. Use of genotyped data.....	280
6.4. Implications for the future.....	281
6.4.1. The endophenotype concept in psychiatric genetics	281
6.4.2. Other methods of improving statistical power.....	281
6.4.2.1. Redefining the depressed phenotype.....	281
6.4.2.2. Increasing the sample size	281
6.5. Concluding remarks	282
References.....	283
Appendices.....	303
2.A. Principal components for population stratification	303
2.A.1. Scree plots	303

2.A.2. Significance of principal components	306
2.B. Model selection fitting environmental components	307
2.B.1. Adding common household environment	307
2.B.2. Adding common sibling environment	307
2.B.3. Adding common spouse environment	308
2.B.4. Adding common spouse environment + common household environment	308
2.B.5. Adding common spouse environment + common sibling environment	309
2.B.6. Adding common household environment + common sibling environment	309
2.A.7. Adding common spouse environment + common household environment + common sibling environment	310
3.A. Code for Taylor transformation to liability scale	311
4. Variance component estimates for quantitative traits	312
4.A. Univariate analyses	312
4.A.1. Digit Symbol Coding	312
4.A.2. Extraversion	313
4.A.3. Neuroticism	314
4.A.4. General Health Questionnaire	315
4.A.5. Logical Memory Immediate	316
4.A.6. Logical Memory Delayed	317
4.A.7. Mood Disorder Questionnaire	318
4.A.8. Mill Hill Vocabulary	319
4.A.9. Schizotypal Personality Questionnaire	320
4.A.10. Verbal Fluency	321
4.B. Bivariate analyses between quantitative traits and MDD	322
4.B.1. Digit Symbol Coding and MDD	322
4.B.2. Extraversion and MDD	323
4.B.3. Neuroticism and MDD	324

4.B.4. General Health Questionnaire and MDD.....	324
4.B.5. Logical Memory Immediate and MDD.....	325
4.B.6. Logical Memory Delayed and MDD	325
4.B.7. Mood Disorder Questionnaire and MDD.....	326
4.B.8. Mill Hill Vocabulary and MDD.....	326
4.B.9. Schizotypal Personality Questionnaire and MDD	327
4.B.10. Verbal Fluency and MDD	328
5. Variance component estimates for composite traits	329
5.A. Bivariate analyses between quantitative traits.....	329
5.A.1. GHQ and EPQN	329
5.A.2. GHQ and MDQ	330
5.A.3. GHQ and SPQ.....	330
5.A.4. EPQN and MDQ	331
5.A.5. EPQN and SPQ	331
5.A.6. MDQ and SPQ	332
5.B. Univariate analyses	333
5.B.1. Sum of Z-Scores (GHQ + EPQN)	333
5.B.2. Genetic composite trait 3 (GHQ, EPQN, MDQ).....	333
5.B.3. Genetic composite trait 4 (GHQ, EPQN, MDQ, SPQ)	333
5.C. Bivariate analyses between composite traits and MDD.....	334
5.C.1. Sum of Z-Scores and MDD	334
5.C.2. Genetic composite trait 3 and MDD	334
5.C.3. Genetic composite trait 4 and MDD	335

Chapter 1: Introduction

1.1. Background

Major depressive disorder (MDD) is the most common psychiatric disorder, with an estimated lifetime prevalence of 16.2% (95% CI 15.1 - 17.3%)¹. It is currently one of the leading cause of disability worldwide, accounting for 10.3% of all years lost due to disability². The global economic burden of MDD was estimated to cost \$2.5 trillion(USD) in 2010, which is projected to increase to \$6 trillion(USD) by 2030³. Despite the extensive personal, societal and economic burden of this disorder⁴, no new antidepressants with superior efficacy have been developed for over 3 decades, and 45% of patients do not respond to currently available treatments⁵.

1.1.1. Symptoms and diagnosis

The ICD-10 and DSM-V are the two most commonly used diagnostic tools for MDD. ICD-10 is more sensitive to the mild range of the MDD spectrum, however ICD-10 and DSM-IV (the previous version of DSM-V) are generally concordant in identifying individuals with moderate to severe MDD⁶. Both diagnostic systems utilize the number of criteria symptoms simultaneously present to classify known psychological disorders.

A diagnosis of DSM-V MDD requires an individual to have experienced five or more out of nine specific symptoms (outlined in Box 1.1.1.) to be present for most of the day, every day for more than two weeks. These symptoms must include depressed mood and/or a loss of interest in daily activities⁷.

Box 1.1.1. Symptoms of DSM-IV depression

1. Depressed mood or irritable, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful).
2. Decreased interest or pleasure in most activities
3. Significant weight change (5%) or change in appetite
4. Change in sleep: Insomnia or hypersomnia
5. Change in activity: Psychomotor agitation or retardation
6. Fatigue or loss of energy
7. Feelings of worthlessness or excessive or inappropriate guilt
8. Diminished ability to think or concentrate, or more indecisiveness
9. Suicidality: Thoughts of death or suicide, or has suicide plan

1.2. Epidemiology of depression

The lifetime prevalence of MDD varies substantially across countries, ranging from ~6.6% in Japan to ~21% in France⁸. Other aspects of descriptive epidemiology, such as age of onset (AOO) and persistence, are more consistent across countries, with the median AOO typically occurring in the early to mid 20s⁹. Sex differences are apparent in the epidemiology of MDD, with the lifetime prevalence¹⁰ and hazard ratio¹¹ being approximately two times greater in females than males.

Environmental risk factors, such as unemployment, social isolation, relationship stressors¹², stressful life events (SLE)^{13,14} and childhood trauma¹⁵, also increase the risk of MDD. In particular, childhood trauma has been shown to double the risk for MDD in adulthood¹⁵. Environmental effects are significant when specific to an individual's risk of developing MDD, however the contribution attributable to common environmental effects (e.g. parenting style, socioeconomic status, or local environmental qualities) is negligible¹⁶.

1.2.1. Evidence for a genetic component

Family, twin and adoption studies can be used to establish the role of genetics in the aetiology of disease. MDD aggregates within families, with 1st degree relatives of cases showing a 2-4 fold increase in disease liability, relative to the general population¹⁶.

Family studies cannot generally distinguish genetic factors from potential shared environmental contributions, unless there is information from many different relationships representing varying degrees of relatedness¹⁷. As collecting large family samples can be very time and resource intensive, not many of these studies exist. Adoption and twin studies can identify and separate genetic and environmental effects. There have been no high-quality adoption studies for MDD, therefore evidence for the genetic basis of MDD predominantly comes from twin studies¹⁸.

The proportion of variance in disease liability attributable to additive genetic factors is referred to as the narrow-sense heritability. Trait heritability can be estimated using pedigree information (h^2) from twins or large family studies, or using common genetic polymorphisms (h^2_{SNP}). The advantage of h^2_{SNP} is that it can be estimated in unrelated samples, which can reduce issues pertaining to modeling common environmental effects. However, these methods have their limitations. Twin studies are not reflective of the population, most of whom are not twins. Therefore estimates from twin studies would rely on the environmental component of twins being the same as single born individuals. This is not necessarily the case, given gene by environment interactions and the fact that twins have exact contemporaneity, where single born individuals in a family do not. Moreover the estimates themselves, which are derived by observing the level of concordance between monozygotic twins versus dizygotic twins, assume that both twin types share the same common environment and have been assigned the correct zygosity¹⁷. The first point is problematic, and can result in twin studies incorrectly apportioning variance to additive, environmental and residual components, causing inflation of heritability

estimates. Whilst the latter point, zygosity, is not as much of an issue in the age of genotype data (as the zygosity could be ratified using Identity-By-Descent estimates) this would be a problem for studies, historically. Estimates derived from genotype data are limited by the fact that they only account for the variance explained by common SNPs that are tagged by genotyped variants, and not all causal variants, and as such gives a lower bound heritability estimate. Therefore h^2_{SNP} , unlike h^2 , does not capture the contribution of rare variants, and is not potentially inflated by gene by environment interactions or epistasis.

Meta-analysis of five quality twin studies provided an h^2 estimate of 37% (95% confidence intervals 31–42)¹⁶, whilst mega-analysis of 18,759 unrelated individuals (9,041 MDD cases and 9,381 controls) provided an h^2_{SNP} estimate of 21% (SE 0.021)¹⁹. The heritability of MDD has been shown to differ between the sexes, with twin studies estimating the h^2 of MDD to be significantly higher in females than males (40% versus 30% and 42% versus 29%). These studies also identified a role for sex specific genetic effects, with the genetic correlation (r_g) of MDD between sexes estimated as +0.55 and +0.63^{20,21}. If the same genetics mechanisms were responsible across sexes, r_g would be 1.

1.2.2. Molecular genetics findings

The evidence from twin and SNP based heritability estimates suggest that there is a genetic component contributing to disease liability. Given the familial aggregation of MDD, research has not only focused on identifying genetic variants which associate with disease susceptibility, but also on what these genetic variants can reveal regarding the genetic architecture - that is the number, type, effect size and frequency of susceptibility variants contributing to MDD. In complex disorders, for example schizophrenia, there is evidence for a spectrum of effect sizes, from common, low penetrance variants^{22,23} to rare, high penetrance variants, such as 22q11.2 deletion syndrome²⁴. These findings suggest that rather than adhering to strict common disease common variant or common disease rare variant architectures, many complex traits are likely a blend of these two hypotheses²⁵. Of course, for some complex disorders it may

in fact be disease heterogeneity, with each disease having its own genetic architecture.

The spectrum of allele frequencies and effect sizes for genetic variants are best interrogated using different statistical methodologies for gene discovery. Figure 1.2.2. outlines the relationship between minor allele frequency (MAF) and effect size (odds ratio) with alleles falling between the dotted lines being of most interest and amenable to identification using current gene discovery approaches.

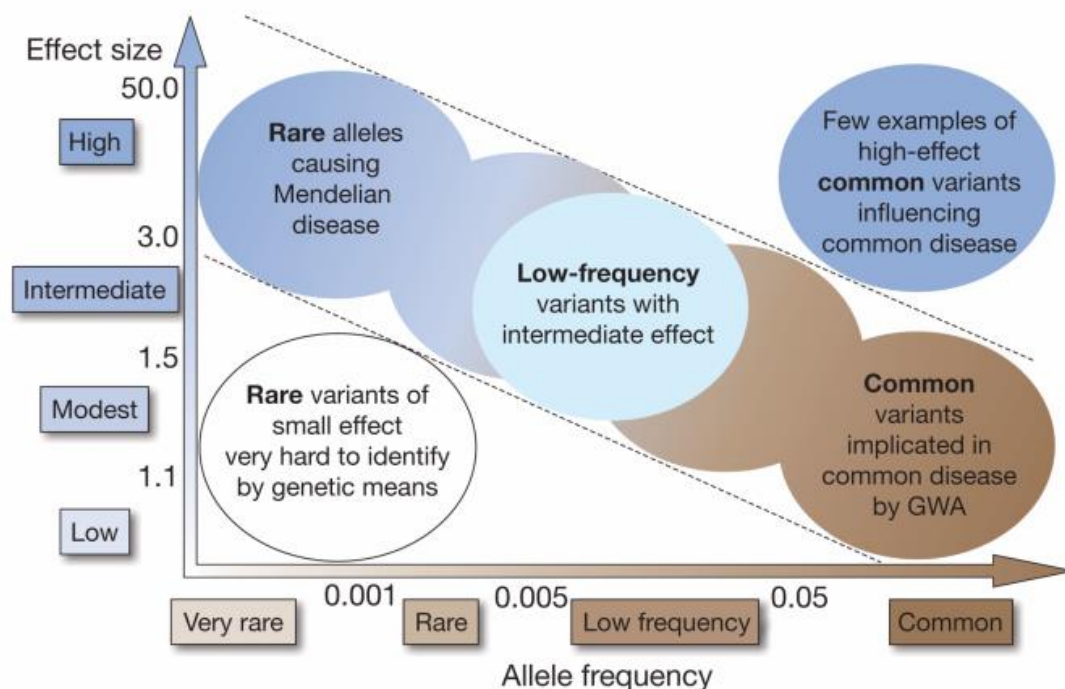


Figure 1.2.2. The relationship between minor allele frequency (MAF) and odds ratio (OR)²⁶. The OR is defined as the odds of disease in individuals with an alternative genotype over the odds of disease in individuals with the reference genotype (adapted from²⁷).

Linkage analysis is most effective at identifying rare variants with large effect sizes, as seen in Mendelian disorders^{28,29} and occasionally in complex traits^{30,31}. Association studies have better power and resolution for identifying common variants with small to moderate effect sizes^{32,33}, with this method establishing susceptibility loci for many complex traits³⁴. Low frequency variants (MAF 0.01<0.05) with intermediate effects sizes have previously been refractory to

conventional gene-discovery approaches due to being poorly tagged by SNPs on genotyping arrays for common variants, and sample sizes in linkage analyses being insufficient to detect these effect sizes. High throughput technologies such as whole genome sequencing provide more in-depth coverage of the genome and may prove invaluable in the detection of these variants in the future, however currently these technologies are still too cost prohibitive in terms of achieving sample sizes required for adequate statistical power.

Rare variants with small effect sizes could contribute to disease liability in common diseases. However, given that the sample size requirements to detect an association scale approximately linearly with $1/\text{MAF}$ and quadratically with $1/|\text{OR}-1|^{35}$, rare variants with small effects are unlikely to be detected with realistic sample sizes. Common variants with large effect sizes are unlikely to contribute to common diseases, as these variants are likely to be older to reach an appreciable frequency in the population, and therefore have been subjected to purifying selection against susceptibility mutations³⁶. Additionally, if such variants did exist, then they would be revealed by association analysis.

1.2.2.1. Linkage studies

Linkage analysis tests for the co-segregation of alleles and the trait of interest within families who have two or more ill relatives. This method utilizes the property of linkage disequilibrium (LD), whereby genetic variants in close physical proximity on the chromosome remain linked during meiosis, resulting in inherited chromosomal fragments called haplotypes. If there is an association between a given haplotype and the disease of interest, all individuals within a pedigree who carry the haplotype should also express the disease phenotype and vice-versa. Linkage analysis is more powerful than association analysis for interrogating rare variants of moderate effect under the assumption that the association differs between families. If LD exists such that the association is consistent across the population, then association analysis is more powerful than linkage analysis for these effects also. The discovery of rare variants of moderate effect would not explain most cases of MDD, however identification of

these variants could provide biologically relevant information regarding the pathophysiology of MDD³⁷. Five genomic regions have been repeatedly found in linkage studies of MDD: chromosome 11, 75-80Mb^{38,39}, chromosome 15, 37-42Mb^{39,40}, chromosome 15, 87-92Mb^{38,41-43}, chromosome 3, 4-9Mb^{38,44}, chromosome 2, 64-68Mb^{44,45}. These results support the role of rare but relatively penetrant variants in contributing to disease risk; although it is possible that these findings are false positives. This is in part dependent on the genetic architecture of MDD - if highly polygenic, as is indicated by genome-wide association studies (GWAS)^{46,47}, then the presence of large-effect loci in a linkage analysis can be falsely indicated⁴⁸.

1.2.2.2. Genome-wide association studies

Association analysis evaluates the difference in allele frequency between unrelated groups of affected and unaffected individuals, or within families. In traits such as MDD where there is an absence of knowledge regarding the underlying biology, methods which evaluate the whole genome in an unbiased manner are favourable. To date there have only been two robustly replicated genome-wide association hits ($p < 5 \times 10^{-8}$) for MDD - rs12415800 (chromosome 10, 69.9Mb; 5' to the *SIRT1* gene) and rs35936514 (chromosome 10, 126.2Mb; in an intron of the *LHPP* gene). These were both identified in the CONVERGE consortium GWAS of Han Chinese women (5,053 cases and 5,337 controls), but did not replicate in the European population⁴⁹. The largest MDD GWAS using individuals of recent European ancestry (9,240 cases and 9,519 controls) failed to identify any replicable genome-wide significant associations⁴⁷, despite successes in similarly sized studies of schizophrenia⁵⁰ and bipolar disorder⁵¹. These results are most likely due to substantial attempts to reduce known sources of phenotypic and genetic heterogeneity (discussed in Section 1.3.3.) in the CONVERGE study. The study only included females (due to the sex differences outlined in section 1.2.1.), cases were recurrent (mean number of episodes = 5.6) and clinically ascertained, both of which are subtypes of MDD that genetic epidemiological studies suggest are more familial or have a higher heritability^{16,52}; controls were aged over 40 to increase the likelihood that they

would remain controls and not experience an episode of MDD in the future⁴⁹. The MDD Working Group of the Psychiatric GWAS Consortium (PGC) mega-analysis combined data from 9 MDD GWAS of European ancestry, including single episode, recurrent and recurrent early onset cases, both sexes, and ascertainment from clinical, population and voluntary sources⁴⁷. However meta-analysis of other heterogeneous diseases, such as Type II Diabetes⁵³, Inflammatory Bowel Diseases⁵⁴ and some cancers^{55,56}, have been successful in identifying genetic variants. This suggests that depression as it is presently diagnosed, descriptively and without reference to any underlying biology, is a sub-optimal phenotype which could be extremely heterogeneous or have non-genetic phenocopies⁴⁷. Figure 1.2.2.2. shows Manhattan plots from CONVERGE and PGC MDD GWAS.

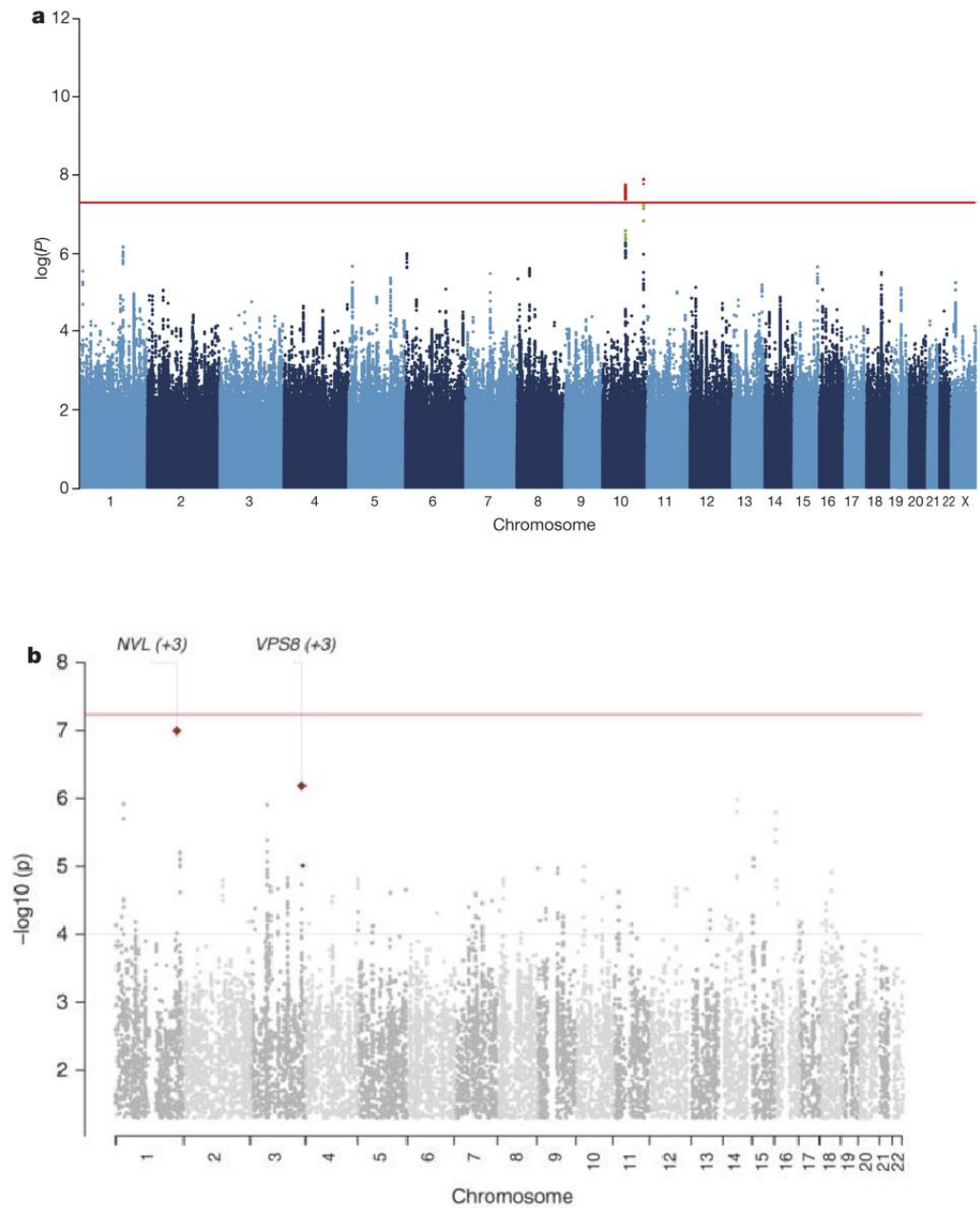


Figure 1.2.2.2. Manhattan plots from the largest available GWAS of MDD showing genetic variants grouped by chromosome and within-chromosome position along the X axis, and the association of each variant with case/control status, given as $-\log_{10}$ of the p-value. a) shows association results from the CONVERGE consortium⁴⁹, b) from the PGC mega-analysis⁵⁷.

Despite the fact that the overwhelming majority of GWAS studies have failed to identify robustly MDD-associated genetic variants, these investigations are still informative regarding the genetic architecture of MDD. Findings to date imply

that common variants of intermediate or large effect do not have main effects in the genetic architecture of MDD, and that MDD is therefore likely to be highly polygenic with multiple loci of low-penetrance acting additively to confer liability to the disorder^{46,47,58}.

1.2.2.3. Candidate gene studies

Candidate gene studies focus on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. Candidate genes are most often selected for study based on *a priori* knowledge of the gene's biological functional impact on the trait or disease in question. Despite a notable lack of *a priori* knowledge regarding the biological underpinnings of MDD, almost 200 genes have been subject to testing via this method. 26 genes have been meta-analyzed which yielded 7 significant ($p < 0.05$) genes: 5HTTLPR, APOE, DRD4, GNB3, HTR1A, MTHFR and SLC6A3. These candidate variants all had a mean OR > 1.35 and MAF $> 10\%$ (mean MAF = 38%) and therefore should be detectable by published GWAS, especially if imputation is used to obtain data from markers not present on the array. Given that the largest published MDD GWAS⁴⁷ has $> 80\%$ to detect an OR greater than 1.2, this suggests that these candidate gene findings are false positives¹⁸.

1.2.2.4. Gene by environment interactions

Gene by environment interactions (GxE), whereby an individual's genotype influences the effect of an environmental exposure on the risk of disease liability, could also play an important role in MDD. The involvement of this mechanism of action seems highly likely, as although there are environmental exposures associated with MDD, for example stressful life events, only one in five occurrences of severely negative life events lead to depression⁵⁹ and not all MDD cases have a precipitating stressful life event. Research in twins concluded that individuals at a higher genetic predisposition to MDD (affected monozygotic co-twin) had an increased probability of onset of MDD (14.6%) relative to individuals with a lower genetic predisposition (6.2%) in response to recent stressful life events. High risk individuals also had a higher probability of

MDD (1.1%) in the absence of recent negative life events relative to low risk individuals (0.5%)⁶⁰. Results from a study of adolescent girls are consistent with this finding, demonstrating that individuals with a higher genetic predisposition to MDD (one or both parents suffering an emotional disorder) who had also experienced a life event in the past year scored highest in a measure of depression⁶¹ derived from the Child and Adolescent Psychiatric Assessment⁶². Furthermore, life events did not have a significant effect when there was no emotional disorder in either parent⁶¹. These findings give weight to the hypothesis that an individual's genotype not only alters their average risk but impacts on their sensitivity to depressogenic environmental exposures. However, it is worth noting that this is also consistent with a threshold model, making it possible for GxE, and epistatic interactions, to exist on the observed scale but not contribute to disease liability on the underlying scale. Consistent with this, despite evidence implicating the contribution of GxE to the aetiology of MDD on the observed scale, identifying specific causal interactions which could lead to biological insight has proved problematic. Interaction terms suffer from issues pertaining to statistical power to a greater extent than studies seeking to identify causative genetic main effects, such as in linkage and association studies. This is due to interaction terms being derived from the product of two first order predictors, and measurement errors in these individual predictors being exacerbated when multiplied to form the interaction term⁶³. Therefore much larger sample sizes are required to reduce this measurement error for both predictors in order to reliably detect an interaction effect, relative to an individual main effect. This, in combination with an elevated multiple testing burden for interaction terms, has hindered progression towards a more unbiased, systematic GxE approach. As such, published GxE studies are predominantly candidate gene GxE (cGxE) studies which hypothesize both the genetic and environmental exposure *a priori*. Table 1.2.2.4. summarizes the candidate GxE literature for MDD⁶⁴.

Gene	Environmental Exposure	Original Study	Successful Replication	Replication Study
BDNF	Stressful life events	Kim et al (2007) ⁶⁵	Yes	Brown et al (2014) ⁶⁶
CRHR1	Childhood maltreatment	Bradley et al (2008) ⁶⁷	Yes	Polanczyk et al (2009) ⁶⁸
HTR2A	Parenting	Jokela et al (2007) ⁶⁹	No	-
MAOA	Childhood adversity	Cicchetti et al (2007) ⁷⁰	Yes	Melas et al (2013) ⁷¹
NR3C1	Childhood adversity	Bet et al (2009) ⁷²	No	-
SLC6A4	Childhood maltreatment	Caspi et al (2003) ⁷³	Yes	Karg et al (2011) ⁷⁴
SLC6A4	Stressful life events	Caspi et al (2003) ⁷³	Yes/No	Karg et al (2011) ⁷⁴ , Brown et al (2013) ⁷⁵ Fergusson et al (2011) ⁷⁶

Table 1.2.2.4. cGxE identified for MDD, including references for the original study and replication attempts where appropriate (adapted from⁶⁴).

Although cGxE have been identified for MDD, measurement of the environmental variable is often problematic in terms of replication, dependent on whether the replication is direct or indirect. Direct replications use the same statistical model on the same outcome variable, genetic polymorphism, and environmental variable tested in the original report. However, replication is invariably performed in an independent cohort which has already collected data. Therefore environmental outcomes are often similar but not identical, resulting in an indirect replication and, as such, testing a different hypothesis to the original study⁷⁷. This can give the impression that the original finding has replicated successfully when, in fact, it has not. This is observed in replication attempts of the role of the serotonin transporter gene 5-HTTLPR in moderating the association between adverse life events and mental disorders. The original study found that individuals carrying one or two copies of the relatively low-expressing short allele had a higher risk of MDD after being exposed to stressful life events or childhood maltreatment than homozygotes for the long allele⁷³.

Many studies have sought to replicate this finding, with results being surprisingly consistent. Meta-analysis of studies which are consistent with the original study design⁷⁶, i.e. direct replication attempts of the original finding, have generally not been supportive. Meta-analysis which included studies incorporating a broader range of environmental stressors, alternative outcome measures and alternative statistical models⁷⁴, i.e. indirect replication attempts, have supported the original finding by Caspi *et al*⁷⁷.

1.2.2.5. Polygenic profiling

Polygenic profiling estimates an individual's burden of common risk variants for a given trait by summing trait-associated alleles, weighted by their marker effects, across the genome to create a polygenic risk score (PGRS) within each subject²². This process requires GWAS summary statistics from an independent training subset, from which the marker weights (generally the logarithm of the odds ratio) and a SNP-list ranked by their evidence for association (usually p-values) can be obtained. PGRS can then be derived for individuals in the replication dataset and used to evaluate the explanatory power of common variants not individually achieving genome-wide significance, incorporating SNPs from a range of P-value thresholds (P_T). This is invaluable for interrogating currently existing data for disorders such as MDD, where the majority of genetic variants are sub-genome-wide significant. PGRS can be used to assess common genetic architecture for related disorders and to aid prediction of disease liability or quantitative trait values by regressing the target sample phenotype onto the PGRS, after accounting for other known covariates. Utilizing the aggregate effects of markers for predictive purposes could result in better discriminative properties than using only established markers, due to the small proportion of phenotypic variance (R^2) that each individual marker explains⁷⁸.

The first successful application of polygenic profiling was in schizophrenia, using the International Schizophrenia Consortium (ISC) GWAS as a training dataset. Schizophrenia cases had a higher mean PGRS compared to controls in

the replication datasets, demonstrating that scoring alleles identified in the discovery GWAS were significantly enriched among independently ascertained cases. Schizophrenia PGRS also showed association with bipolar disorder case/control status, but not six clinically distinct non-psychiatric traits, suggesting a substantial shared genetic component between schizophrenia and bipolar disorder²². Figure 1.2.2.5. shows the phenotypic variance in schizophrenia, bipolar disorder and six non-psychiatric traits explained by schizophrenia PGRS using different P_T .

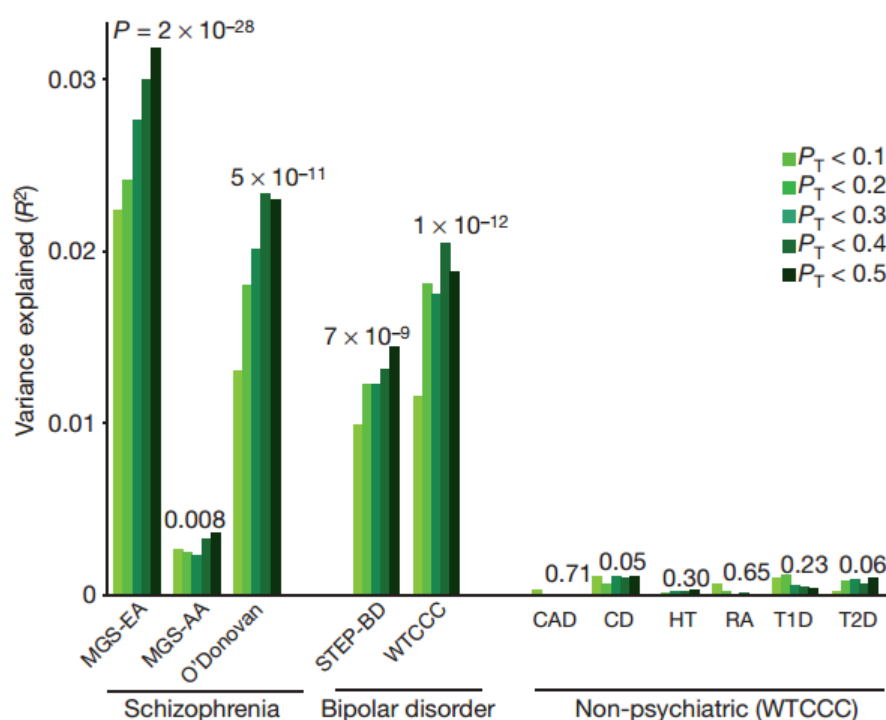


Figure 1.2.2.5. Variance explained (R^2) by ISC derived schizophrenia PGRS using different p-value thresholds in target datasets of schizophrenia, bipolar disorder and six non-psychiatric phenotypes. The y-axis indicates Nagelkerke's pseudo- R^2 ; the P-value above the bars pertains to the association statistic between target phenotype and PGRS using SNPs with a P-value ≤ 0.5 in the ISC GWAS²².

Since this seminal paper, research utilizing PGRS in psychiatric disorders has expanded rapidly. PGRS have limited predictive capacity for extensively polygenic traits using currently available sample sizes^{78,79}. For example, in the PGC MDD mega-analysis, MDD PGRS was highly associated with MDD

case/control status ($p < 10^{-6}$), but only explained 0.6% of the phenotypic variance⁴⁷, demonstrating poor predictive ability. As such, the majority of research has focused on identifying associations between psychiatric disorders and other traits, which has provided some interesting insights regarding common genetic architecture. Research by the PGC Cross-disorder working group has demonstrated a significant overlap of polygenic risk between three adult psychiatric disorders (MDD, bipolar disorder (BD) and schizophrenia (SCZ)), but not between MDD and childhood psychiatric disorders (autism and attention-deficit-hyperactivity disorder)⁸⁰. This common basis to psychiatric disorders has been further exploited to dissect sub-types of BD⁸¹, SCZ⁸² and MDD⁸³. Significant PGRS associations have been observed between psychiatric disorders and a number of non-psychiatric traits including creativity⁸⁴, aggression⁸⁵, cognitive decline⁸⁶, the personality traits neuroticism and extraversion⁸⁷; and neuroimaging measures such as increased gyrification⁸⁸, decreased white matter intensity⁸⁹ and frontal brain activation^{90,91}.

PGRS have an ongoing role in demonstrating a strong genetic basis for traits with an absence of identified genetic variants, and in assessing the common genetic architecture between related traits. However, there are limitations to note regarding this methodology. The scores do not take into account the distribution of SNP effects in the genome⁹² and do nothing to elucidate the biological mechanisms correlating two disorders. Scores are generated using a SNP-set in approximate linkage equilibrium - accounting for LD between SNPs requires SNP selection, based on arbitrary thresholds, which can bias R^2 estimates if the replication sample is not completely independent (i.e. if a dataset has been split into training and replication samples)⁹². P_T are also based on arbitrary thresholds, with the optimum P_T for a given trait dependent on the proportion of null markers (π_0). If π_0 is high (i.e. only a small number of markers contribute to trait variance) then a more stringent P_T is appropriate as this will eliminate a lot of sampling error for the estimated score⁷⁸. In traits where π_0 is low (such as MDD), and a more lenient P_T is appropriate, errors attributable to sampling variance can be considerable due to the small errors for individual

marker effects accumulating across the multiple markers contributing to the PGRS. The number of risk loci underlying a complex disease and their effect size are constrained by the disease parameters of prevalence and heritability⁹³, therefore applying broad P_T across complex traits with different prevalence and heritability estimates may result in substantial information loss. However, software has been developed to counteract this particular problem by testing many incremental P_T ⁹⁴. Regarding the predictive capacity for PGRS, a few thousand cases and controls could provide a clinically useful area under the ROC curve (AUC ~ 0.75) if the trait heritability is explained by approximately 1000 markers. However, for traits like MDD where heritability and the proportion of null markers are low, tens of thousands of subjects would be required. Both association and prediction using PGRS will improve as sample sizes increase⁷⁸.

1.3. Problems facing genetic discovery in depression

As is probably evident from this introduction so far, there are several problems facing genetic discovery for psychiatric disorders, namely statistical power, missing heritability, clinical heterogeneity and clinical covariates. These problems are interlinked, and mostly come back to the pervasive issue of statistical power. These problems are exaggerated for depression relative to other adult psychiatric disorders due to its comparatively low heritability and high prevalence. However, it is worth noting that whilst high prevalence is a disadvantage in a case control GWAS due to cases being less genetically extreme, it can be advantageous in a population sample, as for a given underlying heritability a higher disease prevalence increases the heritability on the observed scale.

1.3.1. Statistical power

Power is a statistical concept used in hypothesis testing to determine the probability of the null hypothesis (H_0) being rejected when the alternative hypothesis (H_1) is true, given a test statistic from the observed data. In genetics

studies H_0 typically states that the genetic variant being tested has an effect size of zero and therefore does not have an effect on the phenotype. H_1 therefore states, as the logical complement of H_0 , that the genetic variant has a non-zero effect size and has either a positive or negative effect on the phenotype (in a two tailed test). Each genetic variant tested will have a corresponding test statistic, generally a p-value. If the p-value is less than or equal to the chosen significance criterion (denoted α), the test suggests that the observed data is inconsistent with H_0 , so H_0 must be rejected. The significance criterion α is also known as the Type I error rate, or false positive rate, and represents the probability of incorrectly rejecting a true H_0 . The Type II error rate, or false negative rate (denoted β), represents the probability of failing to reject a false H_0 . Power, defined as $1-\beta$, is inversely proportional to β therefore the probability of a Type II error is reduced as power increases. Using a more lenient α can increase power by reducing β , but this also increases the probability of Type I error. For whole genome GWAS under the additive model, an α of 5×10^{-8} has been widely adopted for studies of European populations, irrespective of the number of SNPs being tested. Therefore β must be reduced by addressing other factors which affect power such as the sample size, selection of subjects, definition and measurement of the phenotype, accounting for possible confounders and statistical methodologies used⁹⁵. Disease prevalence, and the effect size and MAF of the variant being tested have a substantial impact on power. Whilst these latter factors are outwith the investigator's control, if the underlying genetic architecture is known then a genetic power calculator can utilize estimates of these parameters to indicate the requisite sample size to achieve adequate power.

As previously stated, the higher prevalence of MDD relative to other adult psychiatric disorders suggests that cases are less extreme in the population compared with controls. Disease prevalence has an impact on statistical power. A sample size of 10,000 cases and 10,000 controls would have 50% power to identify a variant that explains 0.08% of the trait variance in liability for SCZ (prevalence $\sim 1\%$), but only 4% power for an equivalent variant in MDD (when

the prevalence is assumed to be $\sim 10\%$)¹⁸. In order to detect a variant that explains the same proportion of total variance in disease liability, a sample size of 2.4 times greater would be required for MDD compared to SCZ. The extensive polygenicity implicated by results from MDD GWAS^{46,47} combined with the success of similarly sized GWAS of other psychiatric disorders^{50,51} suggests that effect sizes may be smaller for MDD than in other psychiatric disorders. This can also greatly impact statistical power. Assuming the number and frequency of risk alleles are the same between these disorders, then sample size requirement increases to 5 times greater due to the differences in heritability (MDD $h^2 \sim 0.37$, SCZ $h^2 \sim 0.81$)^{46,96}. Therefore MDD as it is currently defined could require sample sizes approaching 200,000 cases to achieve the GWAS success of SCZ²³, where 108 loci have now been identified (as shown in Figure 1.3.1.).

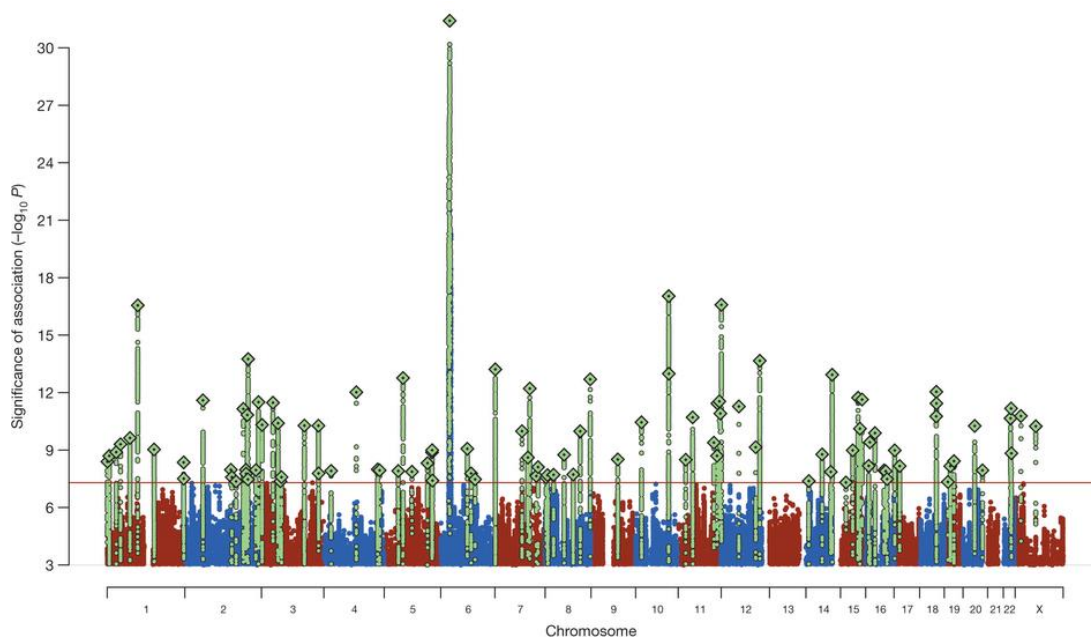


Figure 1.3.1. Manhattan plot showing 108 loci achieving genome-wide significance from the largest available GWAS of schizophrenia (36,989 cases and 113,075 controls)²³.

1.3.2. Missing heritability

Missing heritability is the discrepancy between the proportion of phenotypic variance explained by twin study heritability estimates (h^2) and DNA variants identified by genome-wide association analysis (h^2_{GWS})²⁶. Using MDD in populations of European descent as an example, h^2 is ~ 0.37 ¹⁶ and $h^2_{GWS} = 0$, given that no genome-wide significant loci have been discovered⁴⁷. The variance explained by all genotyped SNPs (h^2_{SNP}) can be estimated using linear mixed models to perform Genomic Restricted Maximum Likelihood (GREML) analysis^{97,98}. Although the estimated h^2_{SNP} for MDD is ~ 0.21 ¹⁹, and therefore the discrepancy between h^2 and h^2_{SNP} is less than that of h^2 and h^2_{GWS} , considerable unexplained heritability remains and is termed "still missing heritability". A schematic of the different heritabilities is shown in Figure 1.3.2.

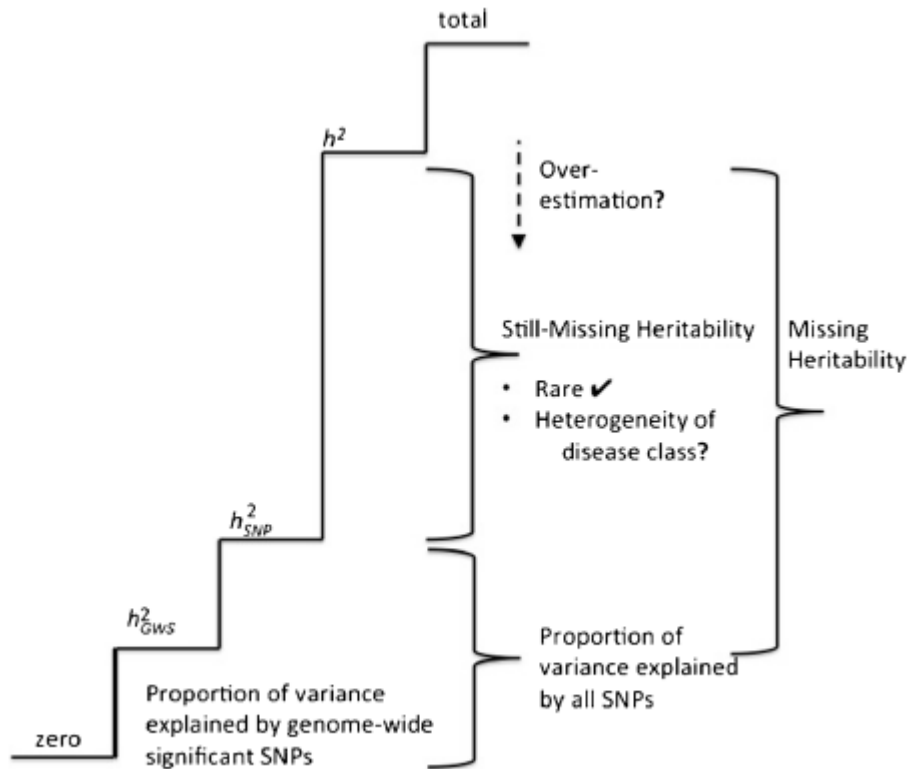


Figure 1.3.2. A schematic of heritabilities showing total phenotypic variance (total), and the discrepancies between the proportion of phenotypic variance attributable to additive genetic effects as estimated from twin studies (h^2), all SNPs (h^2_{SNP}) and genome-wide significant SNPs (h^2_{GWS})⁹⁹.

Many possible explanations have been suggested regarding missing heritability, with proposed reasons pertaining to common variants, rare variants, parent of origin effects, structural variants (e.g. insertions, deletions, inversions, translocations), interaction terms, failing to adequately model environmental effects, and disease heterogeneity (covered in section 1.3.3.). Explanations attributable to common variants include the existence of undiscovered variants whose effect sizes are too small to detect with currently existing sample sizes, as most common variants individually or in combination confer relatively small increments of risk (1.1-1.5 fold)²⁶. Another reason, attributable to both common and rare variants, is variants not being captured by currently available genotyping platforms due to incomplete LD between causal markers and genotyped SNPs⁹⁸. Lack of complete LD can occur if causal variants have lower MAF or allelic heterogeneity, where different mutations at the same locus cause a similar phenotype. If allelic heterogeneity is at play, it might be the case that the missing heritability of common, complex disorders is caused by rare combinations of common variants, rather than an aggregation of yet undiscovered individual common and/or rare variants¹⁰⁰. Parent of origin effects could account for some of the missing heritability. Using Type 2 Diabetes (T2D) as an example, susceptibility variants were shown to confer risk only when inherited from a specific parent. In one instance a variant had opposing effects - conferring risk or reducing risk - depending on the parent of origin. In total, ~13% of the total heritability explained for T2D in this study could be attributed to parent of origin effects¹⁰¹. These effects can be modeled in family studies using phased data but in population studies, where individuals are genetically unrelated and information on the parental genotypes is not available, this may remain an unresolved issue for partitioning the trait heritability. Regarding structural variants, gene-by-gene interactions, gene-by-environment interactions and environmental effects, it has been argued that because non-additive effects do not contribute to the narrow-sense h^2 , arguments based on non-additive effects are not relevant to the missing heritability problem⁹⁸. This is not necessarily true. If missing heritability is considered as $1-(h^2_{\text{known}}/h^2_{\text{all}})$, where h^2_{known} is the proportion of the phenotypic

variance explained by known variants and h^2_{all} is the proportion of the phenotypic variance explained by all variants that affect a trait, then non-additive effects could inflate the denominator term. This results in phantom heritability where, even if all the additive genetic variation for a trait were discovered, the h^2_{SNP} estimate would still be less than h^2 due to h^2 being overestimated by incorrectly modeling the trait as strictly additive¹⁰².

All of these explanations are valid possibilities, and not mutually exclusive. Resolving the issue of the missing heritability problem requires two main solutions. Firstly, a shift from genotyping to whole genome sequencing would allow the investigation of all genetic contributions, including common, rare and structural variants. This would take care of explanations pertaining to additive genetic effects, which are likely to be common in complex phenotypes¹⁰³, and would also allow for long-range phasing to address issues pertaining to allelic heterogeneity. Limitations here are logistical, in terms of sequencing costs and obtaining sufficient sample sizes. Secondly, it has become apparent that methods such as linkage and GWAS are not wholly sufficient to deal with the genetic architecture of complex traits. More appropriate methods must be developed to model genetic architectures which are not strictly additive and can incorporate interaction terms. This second solution is of particular importance as heritability (and therefore solving the missing heritability problem) is not necessarily a key determinant for the identification of strong and replicable genetic associations¹⁰⁴, however the underlying genetic architecture is. If this is not modeled correctly, then hopes of unraveling the genetic basis of complex traits to facilitate prevention, diagnosis, prognosis and treatment are misplaced.

1.3.3. Disease heterogeneity

Disease heterogeneity refers to the variability in expression of the trait of interest. This can prove problematic for genetics studies depending on whether or not the symptomatic variability has a common genetic basis. Evidence to-date suggests that MDD is an aggregation of independent, albeit correlated, subtypes which differ in terms of heritability, genetic background, risk factors, and

symptom presentation¹⁰⁵. Higher heritabilities have been reported for cases who are female^{20,21}, clinically ascertained⁵², recurrent¹⁶ or have an early age of onset (≤ 30 years)¹⁰⁶. It could be argued that these subtypes are simply representative of the more extreme tail of the MDD phenotype, but this begs the question - what is the MDD phenotype? As outlined in Section 1.1.1., a diagnosis of DSM-V MDD requires an individual to have experienced five or more out of nine specific symptoms (Box 1.1.1.), which must include depressed mood and/or a loss of interest in daily activities⁷. All of the symptoms are compounds that include at least 2 sub-symptoms, with three somatic symptoms allowing opposing direction of effects. Therefore, two individuals who meet the DSM-V criteria for MDD may not have a single symptom in common. Considering the nine questions as dichotomous outcomes results in 227 unique potential symptom profiles. When all of the possible permutations of compound scores and opposing direction of effects are incorporated, this increases to 16,400 unique possible profiles. However, just because a permutation is possible does not necessarily mean that it is likely to manifest in reality. To test this, cases from the STAR*D consortium (n=3,703) were used to examine the variability of symptom profiles. This revealed 1030 unique symptom profiles, the most common of which was exhibited by only 1.8% of the sample. The thirty most common profiles described only 24% of subjects, and 14% of the sample exhibited a unique symptom profile¹⁰⁵. Using sum-scores of many disparate symptoms to collapse individuals into one undifferentiated category assumes that symptoms are interchangeable and equal indicators of a single condition with a common genetic basis¹⁰⁷. This assumption is debatable, as individual depressive symptoms have been shown to be differentially heritable. Symptom domains extracted from the Beck Depression Inventory (BDI)¹⁰⁸, the Centre for Epidemiologic Studies - Depression Scale (CES-D)¹⁰⁹, and the Revised Symptom Checklist-90 (SCL-90)¹¹⁰, using Principal Components Analysis (PCA), had heritability estimates ranging from 0.00-0.35. Factors that had a heritable basis described endogenous or physiological functions such as loss of appetite or libido, as well as cognitions such as guilt and hopelessness (possibly reflecting heritable personality traits). Symptoms such as tearfulness did not have a

heritable basis, suggesting that these symptoms are responses to negative life experiences or a learned association to changes in physiologic function. Table 1.3.3.A. shows the heritability estimates for all symptom domains¹¹¹.

Scale	r_{MZ}	r_{DZ}	h^2	95% CI	c^2	95% CI	e^2	95% CI
<i>Beck Depression Inventory</i>								
Component 1. Cognitive-Affective	0.13	0.25	0.00		0.19	(0.09–0.30)	0.81	(0.70–0.91)
Component 2. Performance	0.14	0.11	0.00		0.13	(0.02–0.23)	0.88	(0.77–0.98)
Component 3. Somatic Complaints	0.21	0.03	0.18	(0.05–0.31)	0.00		0.82	(0.69–0.95)
<i>Centre for Epidemiologic Studies—Depression Scale</i>								
Component 1. Depressed Affect	0.17	0.02	0.15	(0.02–0.28)	0.00		0.85	(0.72–0.98)
Component 2. Positive Affect	0.27	–0.02	0.22	(0.10–0.35)	0.00		0.78	(0.65–0.90)
Component 3. Interpersonal	0.19	–0.09	0.13	(0.00–0.26)	0.00		0.87	(0.73–1.00)
Component 4. Somatic/Retarded Activity	0.26	0.09	0.24	(0.12–0.38)	0.00		0.76	(0.63–0.88)
<i>Revised Symptom Checklist-90</i>								
Component 1. Depressed Affect	0.32	0.22	0.34	(0.22–0.47)	0.00		0.66	(0.53–0.78)
Component 2. Suicidal Thoughts	0.21	0.07	0.20	(0.07–0.33)	0.00		0.80	(0.67–0.93)
<i>All Items</i>								
Component 1. Loneliness	0.27	0.22	0.00		0.25	(0.15–0.35)	0.75	(0.65–0.85)
Component 2. Phobias	0.08	0.09	0.00		0.09	(0.00–0.19)	0.92	(0.80–1.00)
Component 3. Insomnia	0.35	0.18	0.35		0.00		0.65	(0.53–0.75)
Component 4. Loss of Libido/Pleasure	0.25	0.05	0.22	(0.01–0.36)	0.00		0.78	(0.65–0.90)
Component 5. Positive Affect	0.20	0.04	0.18	(0.06–0.32)	0.00		0.82	(0.69–0.94)
Component 6. Backaches and Pains	0.16	0.12	0.00		0.14	(0.04–0.25)	0.86	(0.75–0.96)
Component 7. Tearfulness	0.19	0.23	0.00		0.21	(0.11–0.32)	0.79	(0.68–0.89)
Component 8. Loss of Appetite	0.20	0.09	0.20	(0.07–0.33)	0.00		0.80	(0.67–0.93)
Component 9. Interpersonal Problems	0.21	0.24	0.00		0.22	(0.12–0.33)	0.78	(0.67–0.89)
Component 10. Guilt and Hopelessness	0.30	0.17	0.30	(0.18–0.43)	0.00		0.70	(0.57–0.82)
Component 11. Nausea and Headaches	0.04	0.21	0.00		0.10	(0.00–0.21)	0.90	(0.79–1.00)
Component 12. Psychomotor Retardation	0.24	0.32	0.00		0.28	(0.17–0.39)	0.72	(0.62–0.95)
Component 13. Suicidal Thoughts	0.22	–0.02	0.18	(0.05–0.31)	0.00		0.82	(0.69–0.95)
Component 14. Physical Anxiousness	0.24	0.02	0.20	(0.08–0.40)	0.00		0.80	(0.67–0.92)

n = 180 MZ and 163 DZ pairs.

Table 1.3.3.A. Heritability estimates for symptom domains derived from PCA of the BDI, CES-D and SCL-90 showing differential heritability estimates for individual depressive symptoms¹¹¹.

If all depressive symptoms are caused by a common underlying disorder, symptoms should have similar risk factors, yet risk factors exert differential influences on individual symptoms. In a study of 1,015 twins, the association of 25 risk factors with the nine "A" symptom criteria were analyzed. Neurovegetative symptoms tended to differ qualitatively from cognitive and emotional symptoms, sometimes having opposing effects on risk factors such as neuroticism and substance abuse. Criteria also varied substantially over the demographic characteristics of age, educational attainment and income¹¹². A longitudinal study of 1,289 medical interns demonstrated that an individual's prior history of depression, higher number of SLE, childhood stress and sex

predicted unique subsets of three to seven of the symptoms, with neuroticism and work hours significantly impacting on all nine symptoms over time¹¹³.

Furthermore, the reliability of DSM-V is lower (inter-rater kappa = 0.28) than that of the preceding versions, DSM-IV and DSM-III, with psychiatrists under study conditions only agreeing that a patient was suffering from DSM-V depression between 4-15% of the time¹¹⁴. These findings all indicate considerable heterogeneity, which can substantially impact upon statistical power. For example, a GWAS of 10,000 cases and 10,000 controls for a homogenous disease with a prevalence of 1% has >75% power to detect risk variants with heterozygous relative risk of 1.15 and MAF of 0.2-0.8. However, if the 10,000 cases are comprised of two phenotypically and genetically distinct diseases which are diagnosed as a single disease due to indistinguishable clinical presentations, the power becomes non-existent. Figure 1.3.3.B. shows the power curves for both of these scenarios in addition to the power curve for a homogenous disease with only 5,000 cases to verify that the reduction in power is due to heterogeneity rather than the reduced sample size for the disease⁹⁹.

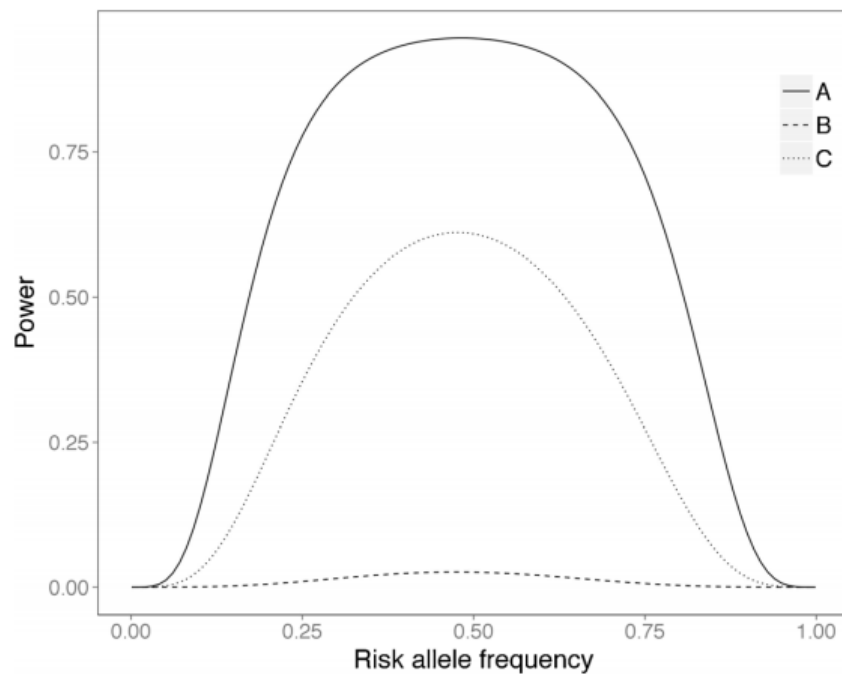


Figure 1.3.3.B. Power of a GWAS to detect risk variants with relative risk of 1.15 when there are A) 10,000 cases of a homogeneous genetic disease and 10,000 controls B) 10,000 cases of a composite disease and 10,000 controls C) 5,000 cases of a homogeneous genetic disease and 10,000 screened controls⁹⁹.

MDD does not reflect a single dimension of genetic liability¹¹⁵. This pervasive heterogeneity impacts on statistical power to detect causative variants. This contributes to the missing heritability problem, as analyzing associations between symptom sum-scores and genetic markers can only capture the shared genetic variance of all symptoms, which may be low. Phenotypic heterogeneity is a difficult problem to solve. The obvious route forward is to stratify the phenotype into more homogenous subsets using what is known regarding differential heritabilities, clinical covariates and environmental exposures. For example, depression-associated genetic factors are largely shared with generalized anxiety disorder, whereas environmental determinants seem to be distinct^{14,60,116}. However, it would be more difficult to achieve sufficiently large sample sizes using more refined phenotypes⁴⁷, and there is not a consensus on what constitute valid subtypes for MDD¹¹³. Another possibility would be to increase the level of detail regarding individual symptoms, as not all existing

diagnostic tools (for example the Patient Health Questionnaire, a routinely used nine question screening instrument) have information on the direction of compound symptoms. In addition, research has consistently confirmed that semi-structured interviews are more accurate and effective than simple questionnaires¹¹⁷, therefore more detailed questions on complex constructs such as sadness could be worthwhile. For the purposes of case control studies, it would also be invaluable for control participants to fully complete structured interviews, as a lot of valuable information is lost due to questions being skipped once the initial screening questions have been administered. This additional information could be used to inform stratification, with multivariate methods such as item response theory and structural equation modeling being employed to analyze the exact relationships between symptoms and underlying dimensions¹⁰⁷. Lastly, MDD could be stratified by the presence or absence of clinical covariates.

1.4. Clinical covariates

Depression covaries with numerous psychiatric and somatic phenotypes. It is not often clear whether this covariance is caused by genuine pleiotropy, by diagnostic misclassification due to disease heterogeneity, by environmental factors, or whether there is a directional effect of one phenotype upon the other. However, looking at phenotypes which covary with MDD could provide insight into common genetic mechanisms and/or the identification of more homogenous subtypes. Regarding other psychiatric disorders, as stated in section 1.2.2.5., polygenic risk between MDD, bipolar disorder and schizophrenia show significant overlap⁸⁰. These genetic relationships have also been demonstrated using GREML methods to estimate the coheritability between five psychiatric disorders (ADHD, autism, bipolar disorder, MDD and schizophrenia). Genetic correlations (r_g) were significant between MDD and bipolar disorder ($r_g=0.47$; $se=0.06$), schizophrenia ($r_g=0.43$; $se=0.06$), and (unlike the PGRS analysis) attention deficit hyperactivity disorder ($r_g=0.32$; $se=0.07$), but not autism ($r_g=0.05$; $se=0.09$)¹⁹.

MDD is also associated with worse clinical outcomes when co-morbid with somatic health problems, notably coronary heart disease¹¹⁸, stroke¹¹⁹, cancer¹²⁰, type 2 diabetes¹²¹ and dementia¹²². However, most of these observations have been based on phenotypic correlations, which do not necessarily indicate a shared genetic basis. A recent study which utilizes genome-wide summary statistics to infer correlations between traits supported the significant genetic correlation between MDD, bipolar disorder and schizophrenia, but did not indicate shared genetic architecture between MDD and Alzheimer's, Type 2 Diabetes or Coronary Artery Disease¹²³. Figure 1.4. shows the genetic correlations between these traits.

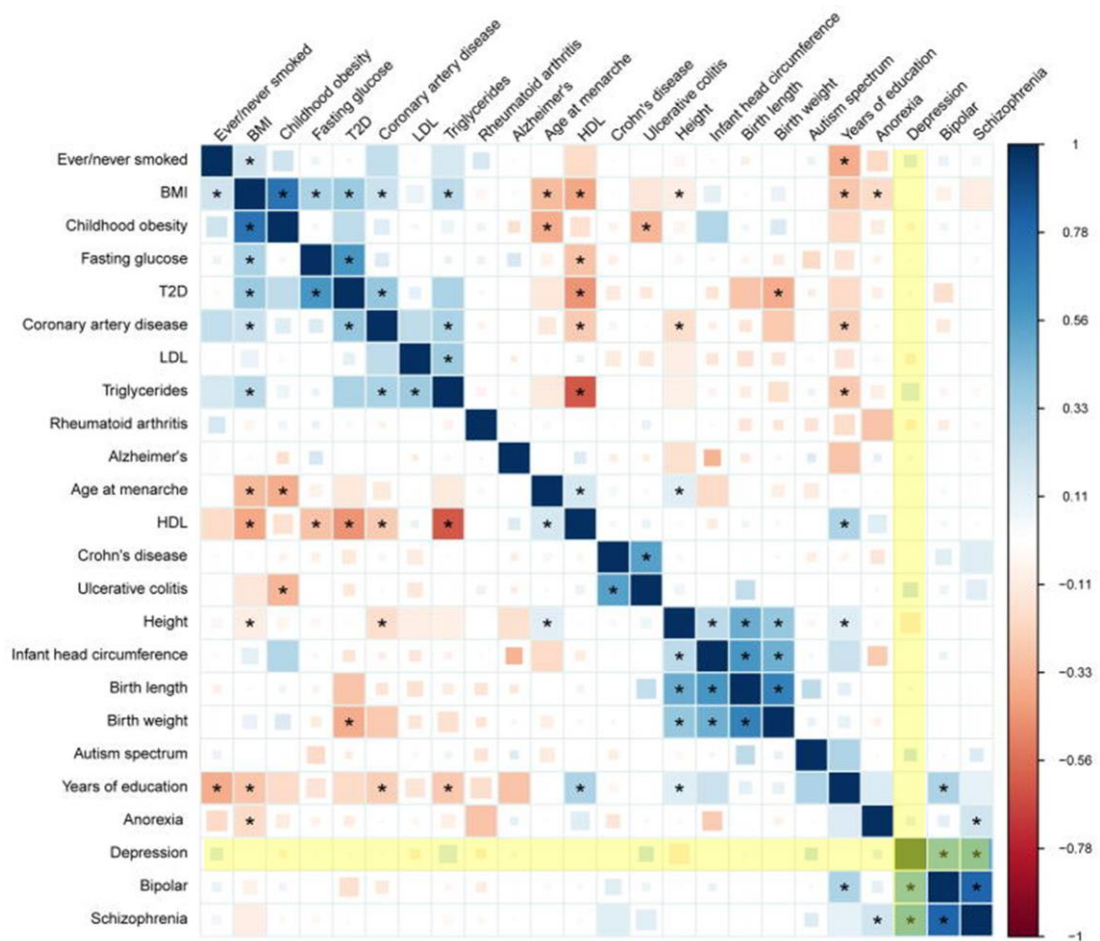


Figure 1.4. An atlas of genetic correlations between pairs of traits demonstrating a shared genetic basis between MDD (highlighted in yellow) and other adult psychiatric traits, but not frequently comorbid somatic symptoms¹²³.

However, somatic symptoms could prove useful for stratifying the MDD phenotype, as these could be representative of environmental factors, could elucidate directionality of effect or prove to be phenocopies. All of this information is useful in the pursuit of reducing heterogeneity in the depressed phenotype. One disadvantage of clinical covariates is that they tend to also be disease states, and therefore binary. Even greater analytic power can be obtained through joint analyses of the dichotomous diagnostic phenotype and a quantitative correlated trait, such as endophenotypes^{124,125}.

1.5. The endophenotype concept in psychiatry

The endophenotype concept was introduced to the field of psychiatry over 40 years ago to bridge the gap between high-level clinical phenotypes and low-level genetic variability, for example SNPs. This is achieved via the identification of measurable traits that are not observable by the unaided eye, and that lie more proximal to the underlying genetics of a disorder than the clinical phenotype itself. These traits should fulfill certain criteria to assert their validity as a candidate endophenotypes. Traits should be heritable, co-segregate with illness, be state independent (i.e. present even when the disease is not), and be more extreme in non-affected family members than in the population¹²⁶. These traits may have increased statistical power to detect genetic variants by being genetically simpler, closer to the level of gene action, having larger genetic effect sizes, more reliably measured or providing added statistical power through their ability to quantitatively rank people within diagnostic category¹²⁷. In practice however, it is often only the last two of these potential benefits that are realized. In psychiatry, candidate endophenotypes are often neuroimaging, electrophysiological, behavioural or cognitive variables. But, the endophenotype approach has had success in other complex phenotypes such as alcohol dependency^{128,129}, heart disease^{130,131}, obesity^{132,133}, and osteoporosis^{134–136}. A schematic of how the endophenotype concept might look for neuroimaging variables in MDD is shown in Figure 1.5.

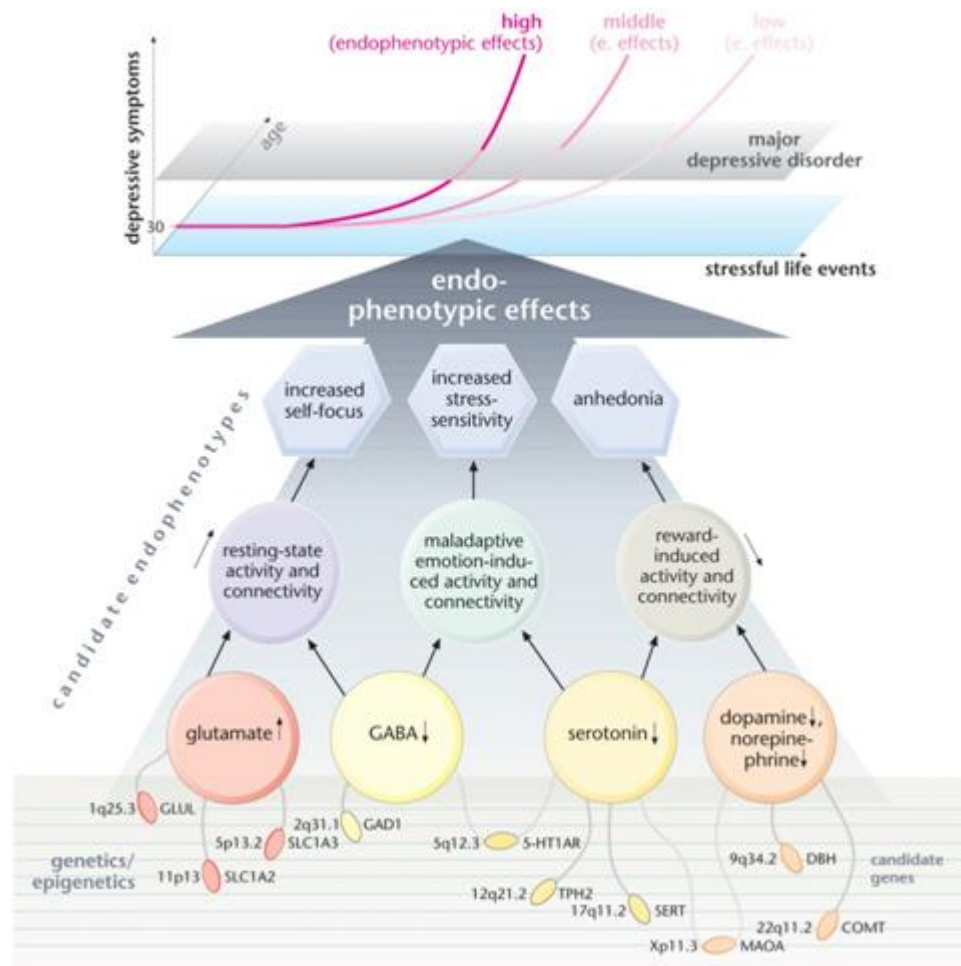


Figure 1.5. A heuristic model of liability to MDD, showing a continuum between candidate genes and the MDD phenotype. This image postulates the relationship between candidate genes for MDD, neurochemical abnormalities, functional neuroimaging abnormalities, and the MDD phenotype in response to age-related stressful life events¹³⁷.

Some candidate endophenotypes are less genetically complex and potentially closer to the level of gene action, such as transcriptomic and proteomic phenotypes. Using unbiased methods for assessing and ranking thousands of measures such as the Endophenotype Ranking Value, which is the absolute value of the coheritability between traits, makes it possible to consider the vast numbers of potential 'omic endophenotypes. This method has been used to assess over 11,000 behavioural, neurocognitive, neuroanatomical and transcriptomic traits as candidate endophenotypes for recurrent MDD. Top ranking traits (BDI, bilateral ventral diencephalon volume, and expression levels of the RNF123 transcript) were utilized in conjunction with disease state

in bivariate linkage analysis. This resulted in a linkage peak on chromosome 4p15 (LOD=3.5) for RNF123 transcript, but no significant peaks for behavioural or neuroanatomical measures, suggesting that candidate endophenotypes closer to gene action could prove fruitful for MDD¹²⁴.

The failing of many endophenotype studies to date has been focusing on phenotypic correlations between traits, when the shared genetic component between traits is what is fundamental to improved gene discovery. Therefore, pedigree-based observations, in conjunction with large population samples, can be used to efficiently test whether a trait fulfills endophenotype criteria, and also to estimate the coheritability of the trait and disease trait.

1.6. Project summary and aims

Genetic association studies of MDD have failed to identify any robust, replicable genetic risk loci in the European population. This is due to issues regarding statistical power. Methods of increasing statistical power include further increasing the sample size, stratifying the depressed phenotype into more homogenous subtypes based on clinical covariates, and revisiting the phenotype to develop a quantitative trait. The last of these methods shall be the focus of this thesis.

Data on MDD case/control status and ten measures of mood, personality and cognitive ability are available in Generation Scotland: The Scottish Family Health Study (GS:SFHS), a large population- and family-based study. These measures will be evaluated and ranked as endophenotypes for MDD, primarily, using measures of covariation from coheritability and polygenic risk score analysis. The convergence between these measures of covariation will be assessed, as polygenic risk score analysis has not previously been used to rank candidate endophenotypes. Given that this method utilizes only common trait-associated variants, selecting quantitative traits for further genetic analysis based on this method may prove advantageous for gene-discovery in genome-wide association studies. Secondly, these measures will be tested for their

adherence to the endophenotype criteria as outlined by Gottesman and Gould, namely: being heritable, genetically and phenotypically correlated with MDD, state independent, co-segregating with illness in families, and observed at a higher rate in unaffected relatives than in unrelated controls. GWAS of these measures will reveal whether individual quantitative traits associated with MDD have improved power to identify genetic risk variants relative to the binary classification of MDD.

To jointly maximize the heritability and genetic correlation with MDD, a quantitative measure of MDD incorporating information from all traits which fulfill endophenotype criteria shall be derived using principal component analysis. The performance of this composite trait will be assessed in terms of ability to identify risk loci in a genome-wide association study, and phenotypic variance explained by polygenic profile scores for psychiatric disorders relative to its individual constituent traits and the binary classification of MDD.

Collaboration not only across institutes but across continents is becoming ever more commonplace, facilitating the pooling of resources to create cohorts of previously unimaginable sizes. As more population studies start to perform intensive phenotyping as part of their design, the possibilities of genetic research in psychiatry are ever expanding. Therefore, if quantitative measures of depression do provide improved etiological insight in a population study of 20,000, combining this technique with publicly available data to increase the sample size and inform stratification could finally take the genetics of MDD a step closer to clinical utility.

Chapter 2: Methods

This Chapter will firstly provide some background on Generation Scotland: The Scottish Family Health Study, and how MDD and extended phenotypes were measured. Secondly, quality control of phenotypic and genetic data will be discussed. This will be followed by a description of the model selection process, which will describe how the population structure and enrichment of families were accounted for in subsequent analyses. Next, an outline of analyses which are in common between results Chapters, such as heritability based analyses, genome-wide association analysis and polygenic risk score analysis will be provided. Lastly, the replication sample used in this thesis, UKBiobank, will be described, as will the replication methods. This Chapter aims to serve as a general reference point for analyses, with any Chapter specific analyses discussed in more detail in the relevant results Chapter.

2.1. Sample description

The analyses presented here were conducted using data from the Generation Scotland: The Scottish Family Health Study cohort (GS:SFHS). All components of GS:SFHS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS:SFHS has also been granted Research Tissue Bank status by the Tayside Committee on Medical Research Ethics (REC Reference Number: 10/S1402/20), providing generic ethical approval for a wide range of uses within medical research. GS:SFHS is a family and population-based study consisting of over 24,000 participants recruited via general medical practices across Scotland between 2006 and 2011. Potential participants were identified from the registers of collaborating general practices via their Community Health Index, a unique identifying number allocated to nearly every individual in Scotland registered with a general practitioner. Probands ($n = 7,953$) were aged between 35 and 65 years and were not ascertained on the basis of having any particular disorder. Eligibility criteria specified that participants were over 18 years of age and had one first-degree relative also willing to participate. A shared family identity number was

given to groups where each member was a first-degree relative of at least one other person, resulting in 5,573 families. Family size ranged from 2 - 36 members, with a mean family size (excluding 1,400 singletons without any relations in the study) of 4.05 members and a median family size of 3 (IQR 2–5)^{138,139}. GS:SFHS is more affluent, educated, in better health and has a different age-sex profile relative to the true Scottish population¹⁴⁰, however, the range of socio-demographic features and detailed clinical information available make it an invaluable study population. The analyses presented in this thesis focus on MDD status, and measures of mood, personality and cognitive ability.

2.1.1. Measurement and definition of depression

A diagnosis of MDD was made using the structured clinical interview for DSM-IV disorders (SCID)¹⁴¹; researchers administered the screening questions (“Have you ever seen anybody for emotional or psychiatric problems?” and “Was there ever a time when you, or someone else, thought you should see someone because of the way you were feeling or acting?”) after a period of training and reliability assessment. Participants who answered yes to either of these questions (21.7% screened positive) were invited to continue the interview, which provided information on the presence or absence of a lifetime history of MDD, age of onset and number of depressive episodes. Participants who answered no to both of these questions were assigned control status. Those who completed the SCID but did not meet the criteria for MDD were also defined as controls. Participants who declined to complete the screening questionnaire or the SCID were set to missing. Participants who fulfilled criteria for a diagnosis of Bipolar Disorder (n=76) were excluded from the current study. Case definition was further refined through NHS data linkage. Controls with a history of antidepressants or who had been referred to a secondary psychiatric care centre (n=1,072) were excluded, as were cases who had received a previous diagnosis of schizophrenia or bipolar disorder (n=47). Inter-rater reliability for the presence or absence of a lifetime diagnosis of major depressive disorder was high (Kappa = 0.86, $p < 0.001$, 95%CI 0.7 to 1.0)¹⁴².

2.1.2. Measurement and definition of mood, personality and cognitive phenotypes

Verbal declarative memory was assessed using the Wechsler logical memory test, which tests participant's immediate (LM1) and delayed (LM2D) recall of one paragraph¹⁴³. Executive function was assessed using the letter-based phonemic verbal fluency (VF) test, where participants stated as many words as they could beginning with letters C, F and L, for one minute each¹⁴⁴. Processing speed was measured using the Wechsler digit symbol substitution task (digit symbol coding (DSC)), consisting of nine digit-symbol pairs followed by a list of digits. Participants were required to write the corresponding symbol under each digit and complete as many substitutions as possible in 120 seconds¹⁴⁵. Acquired verbal knowledge and vocabulary were assessed using the Mill Hill Vocabulary Scale (MHV), junior and senior synonyms. The MHV consists of 88 words equally divided between two sets (A and B). Participants are asked to explain the meaning of each stimulus word in Set A, and then to choose the most accurate synonym in Set B from a choice of six alternatives provided for each stimulus¹⁴⁶. The personality traits of neuroticism (EPQN) and extraversion (EPQE) were assessed using the relevant sections of the Eysenck Personality Questionnaire^{147,148}. Participants were screened for symptoms of bipolar spectrum disorders using the Mood Disorder Questionnaire (MDQ). The MDQ includes 13 questions plus items assessing clustering of symptoms and functional impairment^{149,150}. The schizotypal personality questionnaire-brief (SPQ) assesses the three main factors associated with schizotypal personality disorder: cognitive-perceptual deficits, interpersonal deficits and disorganization, and is used to measure schizotypal features in the general population¹⁵¹. Current levels of MDD and psychological distress were assessed with the 28-item General Health Questionnaire (GHQ), which consists of 4 subscales designed to assess: (A) somatic symptoms, (B) anxiety and insomnia, (C) social dysfunction and (D) 'severe depression'. The Likert scoring method (0 for 'not at all', 1 for 'no more than usual', 2 for 'rather more than usual' and 3 for 'much more than usual') was used¹⁵².

2.1.3. Quality control

2.1.3.1. Phenotypic data

Impossible values, as dictated by the maximum theoretical scores for mood, personality and cognitive traits¹⁴⁰, were set to missing. Quantitative phenotypes were transformed to improve normality using the Box-Cox power transformation for linear models¹⁵³, implemented in the MASS package¹⁵⁴ in R¹⁵⁵. This method utilizes maximum likelihood to identify an optimal power (λ) to which all data should be raised. Optimal power transformations are presented in table 2.1.3.1. To simplify the interpretation of effect sizes, transformed quantitative traits were standardized using $Y=(X-\mu)/\sigma$ to have a mean of zero and a unit standard deviation.

Phenotype	λ	Optimal Transformation
DSC	1	DSC
EPQE	1	EPQE
EPQN	0	$\log(\text{EPQN}+1)$
GHQ	0	$\log(\text{GHQ}+1)$
LM1	1.5	$\text{LM1}^{1.5}$
LM2D	1.5	$\text{LM2D}^{1.5}$
MDQ	0	$\log(\text{MDQ}+1)$
MHV	1.5	$\text{MHV}^{1.5}$
SPQ	0	$\log(\text{SPQ}+1)$
VF	1	VF

Table 2.1.3.1. Optimal Box-Cox power transformation for quantitative traits.

2.1.3.2. Genetic data

Autosomal genotype data were available for 20,023 individuals, 2,659 of whom were cases. Blood samples were obtained using standard operating procedures and were stored at the Wellcome Trust Clinical Research Facility Genetics Core (www.wtcrf.ed.ac.uk). DNA extraction was performed following methods described elsewhere¹⁵⁶. Genotyping was performed using the Illumina HumanOmniExpressExome-8 v1.0 DNA Analysis BeadChip and Infinium chemistry. Genotypes were processed using GenomeStudio Analysis Software v2011.1¹⁵⁷. Analyses were performed using raw genotype data. Quality control (QC) inclusion thresholds used were missingness per individual $\leq 2\%$, genotype call rate of 98%, minor allele frequency $>0.5\%$ and Hardy-Weinberg equilibrium p-value $>1e-6$. Monozygotic twins (19 pairs) were identified using a genomic relationship matrix (GRM) generated from raw QCd SNPs. Twin pairs were retained for genetic analyses, but one from each pair was removed for pedigree-based analyses, with cases preferentially retained. 561,125 genotyped SNPs and 19,994 individuals passed QC criteria and were used in subsequent analysis.

2.1.3.3. Accounting for population stratification

Population stratification refers to the presence of a systematic difference in allele frequencies between subpopulations in a study due to ancestry difference between study subjects. This is a major problem in genetics studies as failing to control for these differences may result in false positives or negatives and ultimately result in true association signals being obscured¹⁵⁸. GS:SFHS is a relatively homogenous population - 87% of participants were born in Scotland and 96% of participants born in the UK or Ireland. Furthermore, approximately 82% of parents and 75% of grandparents were born in Scotland¹⁴⁰. However, as GS:SFHS is a combination of multiple families and singleton data, it is not robust to population stratification¹⁵⁹ and has an enrichment of familial relationships which could upwardly bias test statistics if not properly accounted for. As such, two levels of analysis were performed to address this issue. Multidimensional scaling (MDS) analysis was employed to evaluate discrete population stratification, where populations are remotely related (e.g. Europeans, Africans

and Asians)¹⁵⁸, to establish where GS:SFHS fits in a global genomic context and identify population outliers. To account for population stratification and the enrichment of familial relationships in GS:SFHS in analysis using genotype data, two strategies utilizing additional fixed effects and random effects were employed depending on which program was used in the analysis.

2.1.3.3.1. MDS analysis to identify population outliers

MDS analysis was performed in accordance with the ENIGMA 1000 Genomes protocol¹⁶⁰. Briefly, further to the genetic QC performed as per section 2.1.3.2., strand ambiguous SNPs were removed to avoid strand mismatch between local data and data from the 1000 Genomes Project¹⁶¹. An NxN matrix of genome-wide IBS pairwise distances was generated using PLINK software¹⁶², from which the first 20 MDS components were extracted. The first few components should account for a substantial proportion of the genetic variation in the population and when visualized, allow the identification of any outliers¹⁶³. Plotting the first two MDS components (Figure 2.1.3.3.1.) demonstrated that GS:SFHS overlaps with the populations of European descent (TSI and CEU), as would be expected.

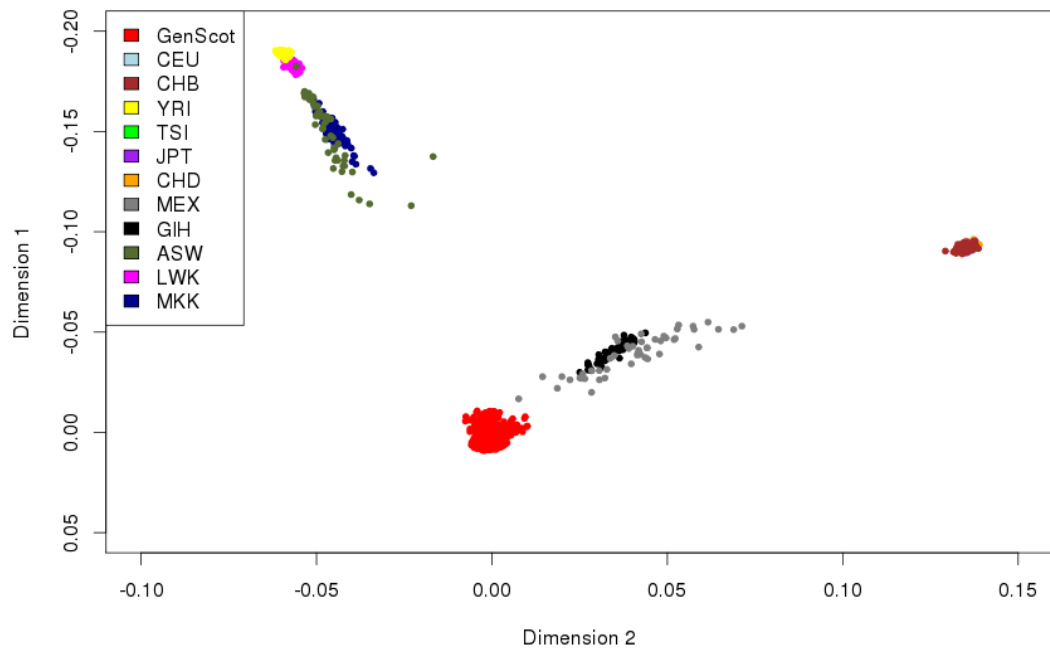


Figure 2.1.3.3.1. MDS plot of Generation Scotland (GenScot) and populations from the 1000 Genomes Project: Utah residents with ancestry from northern and western Europe (CEU); Chinese in Beijing (CHB); Yoruba in Ibadan, Nigeria (YRI); Tuscany in Italy (TSI); Japanese in Tokyo (JPT); Chinese in metropolitan Denver (CHD); people of Mexican ancestry in Los Angeles (MEX); Gujarati Indians in Houston (GIH); people of African ancestry in the southwestern United States (ASW); Luhya in Webuye, Kenya (LWK); and Maasai in Kinyawa, Kenya (MKK)¹⁶¹. Generation Scotland overlays samples of European descent, CEU and TSI.

Further work by Dr. Carmen Amador using principal component analysis (the pipeline for which is described elsewhere¹⁶⁴) identified 91 individuals more than six standard deviations from the mean of the first principal component, indicating population outliers. These individuals were excluded as part of the QC pipeline.

2.1.3.3.2. Accounting for population stratification and enrichment of relationships within GS:SFHS

Two types of analysis in this thesis use individual level genetic data: genome-wide association analysis (GWAA) and polygenic risk score analysis (PGRS). To account for population stratification and enrichment of relationships in GWAA, model fit was testing by running GWAA on MDD as a binary trait and assessing the distribution of test statistics using univariate LD score regression, a statistical method which can distinguish confounding from polygenicity¹⁶⁵. GWAA were conducted using mixed linear model based association analysis (MLMA), implemented in GCTA (v1.25.)^{166,167}. MDD was fitted as the dependent variable; sex, age and age² were fitted as fixed effects; and two Genomic Relationship Matrices (GRMs) were fitted as random effects. The first GRM included pairwise relationship coefficients for all individuals. The second GRM had off-diagonal elements of pairs of individuals who had a relationship coefficient < 0.025 set to 0, therefore excluding pairs of individuals that have a most recent common ancestor of approximately four generations distant, assuming no inbreeding¹⁶⁸. This method has been demonstrated by Zaitlen *et al* to account for potential upward biases due to excessive relationships, thus allowing the inclusion of closely and distantly related individuals in genetic analyses¹⁶⁹. GRMs were created using the mixed linear model with candidate marker excluded (MLMe) approach¹⁶⁷. The MLMe approach involves removing SNPs from the chromosome in question, from the base set of 561,125 genotyped autosomal SNPs, therefore creating an MLMe GRM for each chromosome. This was to prevent loss of power through double fitting of the candidate SNP (and those in LD with it) in the GRM as a random effect, while testing each SNP as a fixed effect. As MLMA is performed on the linear scale, effect sizes are calculated as betas, therefore betas and their corresponding standard errors were transformed to odds ratios and their corresponding standard errors using the Taylor transformation^{170,171}. Summary statistics were then merged with the European population from the 1000 Genomes Project¹⁶¹, as this sample provided the LD Scores and regression weights required for LD score

regression. A sample prevalence of 0.13 and a population prevalence of 16.2%¹⁷² for MDD were supplied to obtain a heritability on the liability, rather than the observed, scale¹⁶⁵. If the model adequately accounts for confounding, the deviation of the LD Score regression intercept from 1 divided by its standard error should be non-significant, and the LD Score Ratio, $(\text{intercept}-1)/(\mu\chi^2-1)$, should be close to 0. The LD Score regression intercept was estimated at 0.99, with a standard error of 0.007. Therefore the intercept was not significant (<1.96). The LD Score Ratio was NA, which usually indicates genomic correction. Therefore, as this model appeared to adequately account for population stratification and enrichment of relationships in this sample, two MLMe GRMs were fitted as random effects for subsequent GWAA.

Polygenic risk score analysis was conducted using restricted maximum likelihood methods, implemented in ASReml-R¹⁷³. Due to computational limitations, the model used for accounting for population structure in GWAA could not be used for PGRS analysis. Therefore in PGRS analysis using mixed linear models, an additive genetic component, common household environment and common sibling environment (discussed in Section 2.2.1.1.) were fitted as random effects, and sex, age, age² and principal components (PCs) were fitted as fixed effects. The first 20 PCs were derived using GCTA¹⁶⁶ software, incorporating data from all autosomal SNPs (n=561,125) and all genotyped individuals who passed QC (n=19,994). To determine the number of PCs required, linear regression models of the standardized traits on age, age², sex, and up to 20 PCs were run for major depressive disorder, mood, personality and cognitive traits. PCs are included in the model in their order of importance based on the variance they explain in the kinship matrix. As the optimum number of PCs required across traits was not immediately apparent from the visualizations of log-likelihood and residual variance (scree plots are shown in Appendix 2.A.1.), the best fitting model for each trait was evaluated using a stepwise approach, adding one PC at a time, and using a likelihood-ratio test (LRT), the output of which was assessed against a mixed $0.5(\chi^2)+0.5(0)$ distribution¹⁷⁴. The significant PCs for each trait are shown in Appendix 2.A.2.

For continuity, the same number of PCs were incorporated into the mixed linear model across all traits. As the number of significant PCs was so variable between traits, the first 20 PCs were fitted as fixed effects for PGRS analysis to ensure adequate correction across traits.

2.1.4. Statistical power

Statistical power is the probability of rejecting the null hypothesis when a specified alternative hypothesis is true. Power analysis, which can be used to calculate the minimum sample size required to detect a given effect size at a given minor allele frequency, was performed using the Genetic Power Calculator¹⁷⁵. For MDD, the Case-Control for Discrete Traits option was used. Disease prevalence was set to 0.13 (as this was the sample prevalence for MDD in GS:SFHS). Power for a genetic variant of Genotypic Relative Risk (GRR) 1.2 and 1.44.minor at minor allele frequencies (MAF) of 1%, 5%, 10%, 15%, 20% and 25% were calculated. This was assuming the multiplicative model, whereby each risk allele multiplies the risk factor by 1.2. Results for these calculations are shown in Table 2.1.4.A.

Trait	MAF (0-1)	Power (0-1)	N _{cases} for 80% power
MDD n _{cases} = 2659; n _{controls} = 17,237	0.01	0.00005	43,894
	0.05	0.02	9,294
	0.10	0.19	5,002
	0.15	0.48	3,600
	0.20	0.71	2,924
	0.25	0.84	2,543

Table 2.1.4.A. Power estimates for MDD for genetic variants at 1%, 5%, 10%, 15%, 20% and 25% MAF which have a genotypic relative risk of 1.2 and 1.44. The estimated number of cases required to provide 80% power at significance level 5×10^{-8} is provided.

Power calculations for quantitative traits were performed using the Case-Control for Quantitative Traits option, considering GHQ (n=19,693) and MDQ (n=10,357) under a strictly additive model. Effects for quantitative traits are expressed in terms of variance components rather than genotypic relative risks, therefore power was estimated at MAF of 1% for an allele which explained 0.01-0.05% of the phenotypic variance. As an increase in trait value for GHQ and

MDQ indicates the presence of more symptoms, the threshold for caseness was one-sided, with cases defined as being one standard deviation above the trait mean. Results are shown in Table 2.1.4.B.

Trait	% Variance Explained	Power (0-1)	N_{cases} for 80% power
GHQ n _{cases} = 3000; n _{controls} = 16,693	0.1	0.01	11,500
	0.2	0.23	5,316
	0.3	0.69	3,348
	0.4	0.94	2,400
	0.5	0.995	1,848
MDQ n _{cases} = 2,360; n _{controls} = 10,253	0.1	0.0004	20,630
	0.2	0.01	9,686
	0.3	0.06	6,183
	0.4	0.19	4,487
	0.5	0.40	3,496

Table 2.1.4.B. Power estimates for quantitative traits, where cases are defined as having a trait value 1 standard deviation above the mean, for genetic variants at 1% MAF which explain 0.1-0.5% of the phenotypic variance. The estimated number of cases required to provide 80% power at significance level 5×10^{-8} is provided.

With the sample sizes available, there is >80% power to detect a variant of GRR 1.2 and 1.44 with a MAF of 25% for MDD. Quantitative traits with a larger number of observations, such as GHQ, are well powered to detect variants with MAF of 1% which explain >0.4% of the phenotypic variance. Quantitative traits with a smaller number of observations (MDQ and SPQ) do not approach sufficient power to detect variants with a MAF 1%.

It is worth noting that standard power calculators rely on the assumption that the sample is unrelated. The mixed family- and population-based design of GS:SFHS is therefore likely to cause an upwardly biased estimate of power. Furthermore, the Case-Control for Discrete Traits option in the GPC assumes that controls have been selected to be disease free or are a random sample from the population. This can be problematic for disorders such as MDD that show evidence of a non-linear relationship with age^{176,177} as, although controls were screened for depression, some individuals may go on to develop depression.

Especially as relatives of cases are included in the sample who, in accordance with the epidemiological literature, have an increased disease liability. Therefore these power calculations are presented to provide a pragmatic benchmark.

2.2. Mixed Linear Models Analysis

Mixed linear models (MLM) are statistical models that contain both fixed effect and random effects. Fixed and random effects respectively refer to known population-average effects and subject-specific effects, where the latter are generally assumed to be unknown, latent variables. The linear mixed model can be written as:

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

where y is a known vector of observations, β is an unknown vector of fixed effects, u is an unknown vector of random effects, X and Z are design matrices of regressors which relate the observations of y to β and u respectively, and ε is unknown vector of random errors¹⁷³. This technique has been applied in animal breeding for many years. More recently MLM have been applied in human genetics in preference to the fixed effects linear models, as MLM increase the power to detect causal variants and reduce the risk of false positive associations by applying a correction that is specific to the sample structure. Additive genetic effects, the variance/covariance structure of which can be described by relationship matrices, can be fitted as a random effects. This allows the phenotypic variance attributable to these factors to be regressed out before testing the fixed effect of interest. Therefore, MLM are effective in accounting for sample structure such as family relatedness (genetic and environmental), cryptic relatedness and geographic population structure.

A MLM framework was employed for heritability analysis, GWAS and PGRS analysis. Age, age² and sex were fitted as fixed effects across models. Age and sex were fitted due to their differential heritabilities and prevalence, as noted in sections 1.2. and 1.3.3. Age² was fitted to account for the non-linear relationship

between MDD and age which has been observed, with average symptom scores being higher amongst young (<35) and very old (>75)^{176,177}. Different additional fixed effects and random effects were fitted in these analyses dependent on whether they used pedigree data, genomic data, or both.

2.2.1. Estimating trait heritability

Heritability is the proportion of the phenotypic variance that is attributable to genetic effects. Estimates vary from zero to one, with higher heritabilities representing stronger genetic effects¹⁷. Classically, heritability has been estimated from correlations among family members. Close relatives, such as siblings, should be more phenotypically correlated than more distant relatives, such as cousins, and the correlation should fall off in a predictable manner with the degree of family relationship. Heritability estimated from twin samples is 'broad sense' heritability. This estimate measures the full contribution of genetic effects, including dominance and epistatic components. It is defined as $H^2 = V_G/V_P$, where V_G is the total variance due to genetic effects, and V_P is the total phenotypic variance, or more specifically $V_G = V_P - V(P|G)$ where $V(P|G)$ is the phenotypic variance between genetically identical individuals¹⁰². However, generally when studies refer to heritability estimates, they are only considering the extent to which phenotypes are determined by the genes transmitted from parents, otherwise known as 'narrow sense' or additive genetic heritability¹⁷. This estimate is the maximum variance that can be explained by a linear combination of the allele counts, defined as $h^2 = V_A/V_P$; where $V_P = V_A + V_C + V_E$, where V_A is the additive genetic variance, V_C is the variance attributable to common environmental factors and V_E is the residual variance. Residual variance encompasses random noise, unmeasured environmental influences and the effects of genetic variants that are not captured by the measured variants.

2.2.1.1. Accounting for shared environment

To ensure that trait heritability was not overestimated, potential sources of environmental resemblance were modeled by fitting household, sibling and spouse variables as random effects.

2.2.1.1.1. Shared household environment

Household environment was modeled based on parent-offspring relationships. Two household environment variables were created, representing historic and recent shared environment. This was to avoid inferring a common environment between grandparents and grandchildren in three generation families, where the individuals in the mid-parent generation are parents and offspring concurrently. Household variables were created by Ms. Yanni Zeng using the following methodology: Three generation families were split into historic (grandparent - mid-parent) and recent (mid-parent - offspring) groups. Two generation families were then assigned to one of these groups. To establish which group each two generation family should be assigned to, an optimal age cut-off was calculated. To determine this value, three generation (n=721) and two generation (n=8,798) families were identified and the age range of the parents in these groups visualized, as shown in figure 2.2.1.1.1. The mean age for each group was 74 (grandparents from three generation families), 49 (mid-parents from three generation families) and 60 (parents from two generation families).

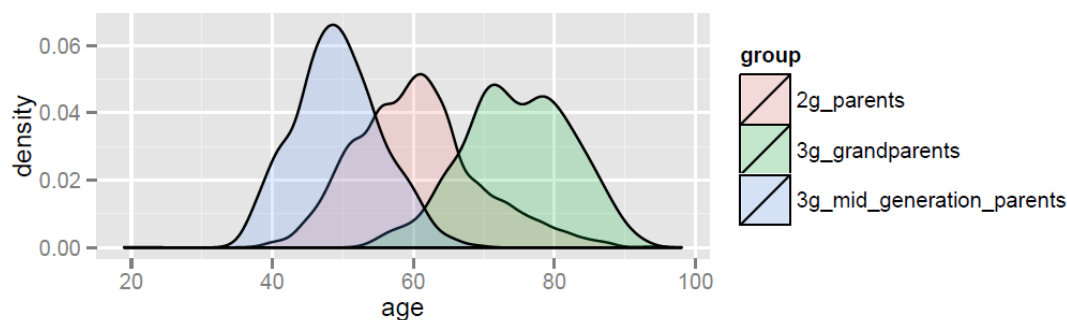


Figure 2.2.1.1.1. Age distributions for parents and mid-parents from two and three generation families.

All possible age cut-offs in this range (49 - 74 years) were tested, using the average age of each parent couple to ensure that parents belonged to the same generational group. Using age 59 as the age cut-off minimized the variance between the two groups, and resulted in approximately equal group numbers. Parents and offspring were then assigned their own unique identifier for either the historic or recent shared environment. In analyses which fit common household environment as a random effect, both old and young household are fitted simultaneously. This is because the two household environmental variables were generated for the purpose of more accurately modeling parent-offspring environmental effects in multi-generational families, rather than specifically seeking to model these generations separately for hypothetical purposes.

2.2.1.1.2. Shared sibling environment

A shared sibling environment was calculated as the effect of common parents, whereby individuals who shared two parents were assigned a same unique identifier, reflecting the shared environment of full-sibs.

2.2.1.1.3. Shared spouse environment

A shared spousal environment was determined by coding individuals with a common child as having an unique identifier.

Individuals who did not share a common environment variable with any other participants were assigned their own unique identifier, demonstrating that they had a common household, sib and/or spouse environment with themselves. These environmental variables do not account for more nuanced modeling of shared environments (e.g. half-sib, grandparent, maternal or paternal effects), however they serve the purpose of broadly accounting for shared environment so as not to overestimate genetic effects.

2.2.1.1.4. Shared environment model selection

The best fitting model for each trait was identified using a stepwise approach by comparing null and alternative models using a likelihood-ratio test (LRT), the

output of which was assessed against a mixed $0.5(\chi^2)+0.5(0)$ distribution. All models fitted sex, age and age² as fixed effects. The initial null model included only an additive genetic effect (ped) as a random effect. Three alternative models were generated, each fitting an individual environmental variable as an additional random effect (e.g. ped + household, ped + sibling, ped + spouse). Additional environmental effects were then added until all four random effects were incorporated into the model. Model selection was implemented in ASReml-R with MDD or standardized traits as the dependent variable, and age, age² and sex fitted as fixed effects. Detailed results tables from the likelihood ratio test for each stage of this stepwise model fitting approach are provided in Appendix 2.B. Variance component estimates are also provided in Appendix 4.A. (as these tables are referred to predominantly in the context of heritability estimation in Chapter 4).

The addition of common spouse environment improved model fit for seven traits (GHQ, MDQ, SPQ, DSC, EPQE, LM2D and VF). The LRT indicated that adding common spouse environment improved model fit for MHV, however other components from the model output indicated that this is likely to be erroneous. Initially, the spouse and ped parameters changed by ~75% in the final iteration (a warning message is issued if >1%), therefore the `update.asreml` function was applied. This function uses the variance structures of the random and error terms from the previous model as starting values to give a better estimate for the parameter space. Whilst this resolved the issues pertaining to percentage parameter change, and the log-likelihood converged, the constraint changed from positive to "?". The default parameter space constraint is positive. If the algorithm pushes the parameter space negative for two successive iterations it is constrained to a small positive value and reported as "boundary". Therefore "?" suggests that, despite model convergence, the parameter space is unknown and therefore this model output is likely to be erroneous¹⁷³. The addition of common household improved model fit for four traits (GHQ, SPQ, MHV and VF), and common sibling environment improved model fit for three traits (LM1, LM2D, and MHV). Therefore, for the next step of model selection,

null models fitted ped and spouse, and ped and household as random effects. The alternative models for this next stage of model selection involved adding household or sibling.

Regarding models with ped + spouse as a null model, adding household environment significantly improved model fit for two traits (SPQ and VF). This model still suffered constraint issues for MHV. Adding sibling environment as an additional effect also significantly improved model fit for two traits (LM1 and LM2D). However, this model caused singularity for MHV, meaning that the model did not converge due to the residual variance component being constrained to zero, caused by two successive iterations having negative values. This error is usually caused by model overspecification¹⁷³, therefore fitting the inverted relationship matrix, spouse environment and sibling environment is soaking up all of the variance resulting in no residual variance. Regarding models with ped + household environment as the null model, adding sibling environment significantly improved model fit for five traits (DSC, LM1, LM2D, MHV, VF), notably all variables pertaining to cognitive ability.

Lastly, model fit including all three environmental variables as random effects was assessed.

Although the effect of common spouse environment significantly improved model fit for seven out of eleven traits, without longitudinal data it is difficult to establish whether this effect is indeed common environment, or whether it is assortative mating. Given that there is strong assortative mating for intelligence, trying to fit a common spouse environmental effect in the presence of assortative mating would incorrectly partition some of the variance from the genetic effect. As the purpose of fitting environmental sources of resemblance in this study is to ensure that heritability estimates are not overestimated, spouse effects were not modeled in subsequent analyses.

Although common household and common sibling environment only improved model fit individually for four and three traits respectively, these effects could influence the heritability estimate if ignored. Therefore common household and

sibling environments, in addition to an additive component, were modeled as random effects in subsequent analyses.

2.2.1.2. Heritability estimates for MDD

Heritability analysis of MDD as a binary trait was performed by Ms. Ana Maria Fernandez-Pujals. In order to overcome the limitations of restricted maximum likelihood methods with regard to non-Gaussian binary response data¹⁷⁸, the h^2 of MDD was estimated from variance components generated using Bayesian methods implemented in MCMCglmm-R¹⁷⁹. MDD status (0 = absent, 1 = present, NA = missing or excluded) was fitted as the dependent variable using a logit link function. Age and sex were fitted as fixed effects, and an additive genetic effect was fitted as a random effect by creating an inverse of the additive relationship matrix using pedigree kinship information. The model ran for 600,000 iterations and used a 200,000 burn in period to achieve acceptable parameter space sampling. This analysis was repeated four times to ensure satisfactory model convergence, which was confirmed by visually comparing the sampling distributions from each run overlapped and testing whether they were indistinguishable^{142,180}.

2.2.1.3. Heritability of quantitative traits

Heritability estimates for quantitative traits were estimated using restricted maximum likelihood methods, implemented in ASReml-R¹⁷³. Standardized traits were fitted as the dependent variable, and sex, age and age² were fitted as fixed effects. An inverse of the additive relationship matrix using pedigree kinship information, and common household and sibling environment were fitted as random effects. Theoretical relationships based upon the recorded pedigree were used for the additive random effect rather than a genomic relationship matrix for computational efficiency. This is appropriate, and will not result in a loss of power, given the sample size and likely polygenic architecture of these traits¹⁸¹. As these analyses used theoretical relationships, principal components were not fitted to adjust for population stratification. It is worth noting however that PCs could influence the heritability estimate.

2.2.3. Estimating genetic and phenotypic between trait correlations

Bivariate genetic analyses between each trait and MDD were conducted in ASReml-R with standardized trait and MDD as dependent variables, and age, age² and sex fitted as fixed effect interaction terms. An inverted additive relationship matrix, and common household and sibling environment were fitted as random effects. Using kinship information was favourable for bivariate analyses involving the MDD phenotype, as heritability estimates using pedigree information do not suffer downward biases as a product of phenotypic heterogeneity to the same extent as estimates using genomic information⁹⁹. The genetic correlation on the observed scale is equivalent to that on the liability scale due to an approximate linear relationship existing between the random additive genetic effects on different scales¹⁸². Genetic correlations (r_g) between traits and MDD were calculated using $r_g = \text{cov}_A / \sqrt{(V_{A_i} * V_{A_j})}$ where cov_A is the additive covariance, V_{A_i} is the additive variance of the quantitative trait and V_{A_j} is the additive variance of MDD. The significance of r_g was calculated by generating a null model assuming no covariance between traits, comparing the alternative and null models using a LRT, the output of which was assessed against a mixed $0.5(\chi^2) + 0.5(0)$ distribution. Phenotypic correlations (r_p) between traits and MDD were calculated using $r_p = \text{cov}_P / \sqrt{(V_{P_i} * V_{P_j})}$ where cov_P is the phenotypic covariance, V_{P_i} is the phenotypic variance of the quantitative trait and V_{P_j} is the phenotypic variance of MDD. As ASReml-R does not provide a test statistic for phenotypic correlations, the significance of r_p was estimated from the ratio of the phenotypic covariance component to its standard error compared to zero using a Z-test.

2.2.3.1. Estimating coheritability between traits

Coheritability assesses the covariation between two traits and evaluates the ability of quantitative traits to predict disease liability. Coheritability (h_{ij}) between traits and MDD was calculated using $h_{ij} = |\sqrt{h_i^2} \sqrt{h_j^2} r_g|$ where h_i^2 is the estimated heritability of the quantitative trait, h_j^2 is the estimated heritability

for MDD and r_g is the genetic correlation between trait and MDD. The method of calculating the significance of h_{ij} is identical to that used for r_g in section 2.2.3. The above formula returns the absolute value, making these estimates equivalent to the Endophenotype Ranking Value¹²⁴, described in section 1.5.

2.2.4. Estimating between group differences in trait mean

Groups were defined as cases, unaffected first degree relatives and unrelated controls - as endophenotype trait values should be more extreme among cases and their unaffected family members relative to the general population¹²⁶. This resulted in 2,659 cases, 2,106 unaffected first degree relatives and 11,898 unrelated controls. Group differences in mean trait scores were calculated using restricted maximum likelihood methods, implemented in ASReml-R¹⁷³. Standardized traits were fitted as the dependent variable; age, age², sex and group (with unrelated controls as the reference group) were fitted as fixed effects; an inverse of the additive relationship matrix, common household and common sibling environment were fitted as random effects. F statistics, degrees of freedom, effect sizes, Z-scores and P-values were derived using the Wald Conditional F-test¹⁸³.

2.2.5. Genome-wide association analysis using mixed linear model analysis

GWAS were conducted using a restricted maximum likelihood methods, implemented in GCTA^{166,167}, using raw genotype data (nSNPs=561,125). Chromosomes were analyzed separately due to computational limitations. Sex, age and age² were fitted as fixed effects. Two GRMs created using the MLMe approach were fitted per analysis as random effects to account for population structure and the enrichment of relationships in GS:SFHS^{169,165}, as described in Section 2.1.3.3.2. Where necessary, betas and their corresponding standard errors were transformed to odds ratios and their corresponding confidence intervals using the Taylor transformation^{170,171}. Estimates of genomic inflation factor λ and its standard error were obtained using the `estlambda` function in

the R package GenABEL¹⁸⁴. GWAS results were visualized using quantile-quantile plots and Manhattan plots, generated using the qqman package¹⁸⁵ in R.

2.2.6. Polygenic Risk Score Analysis

Polygenic profile scores were created by Ms. Ella Wigmore and Ms. Jude Gibson using PLINK, in accordance with methods described by Purcell *et al*²². Summary statistics from the most recent Psychiatric Genomics Consortium (PGC) GWAS of MDD⁴⁷, schizophrenia²³, bipolar disorder⁵¹ and autism⁸⁰ were used to provide marker weights and evidence of association for SNPs. SNP lists containing SNP rs ID, reference allele and log of the odds ratio were generated for each of these disorders at five p-value thresholds (P_T) of $p \leq 1$, $p \leq 0.5$, $p \leq 0.1$, $p \leq 0.05$, $p \leq 0.01$. In GS:SFHS, genotyped SNPs which passed QC were used. In addition, strand ambiguous SNPs were removed. SNPs common to PGC and GS:SFHS datasets were extracted from GS:SFHS and subjected to clump-based linkage disequilibrium pruning, using an LD r^2 cut off of 0.25, 300kb sliding window and variance inflation factor of 2, to create a five SNP sets in approximate linkage equilibrium for each disorder. Polygenic profile scores were obtained by summing the number of reference alleles (0, 1, or 2) at that SNP multiplied by the logarithm of odds ratio for that SNP across the SNP list.

The association between trait values and polygenic profile scores were calculated using the Wald Conditional F-test¹⁸³ in ASReml-R¹⁷³. Trait value was fitted as the dependent variable; the first 20 PCs, sex, age, age², the number of missing genotypes per individual (to control for potential differences in genotyping rate and DNA quality between samples), and polygenic profile score were fitted as fixed effects; an inverse of the additive relationship matrix, common household environment and sibling environment were fitted as random effects. Twenty PCs were fitted in favour of a second relationship matrix due to the computational limitations associated with fitting additional GRMs in ASReml-R. As outlined in Table 2.1.3.3.2., 20 PCs should be sufficient to account for structural biases in GS:SFHS. Variance in trait value explained by polygenic profile score was estimated by multiplying the profile score by its

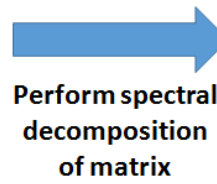
corresponding regression coefficient and estimating its variance. This value was then divided by the variance of the observed phenotype to yield a coefficient of determination between 0 and 1¹⁸⁶, and multiplied by 100 to convert R^2 to percentage variance explained.

2.3. Multivariate Analysis

Three composite measures of depression were derived incorporating the two, three or four traits which ranked most highly across endophenotype criteria. The two trait measure was constructed by simply summing the two standardized variables, as the contributing trait weights cannot be estimated via PCA for fewer than three traits, so this method assigns equal weight to each trait. When incorporating three or four traits, measures were generated using genetic and phenotypic correlations between traits. Genetic correlations were derived using the methods outlined in section 2.2.3. For measures derived from genetic correlations, a genetic correlation matrix between traits was constructed. This matrix was then subjected to spectral decomposition using the eigen function in R to compute eigenvalues and eigenvectors. The loading for the first principal component of each trait was then extracted, multiplied by its corresponding standardized trait value, and the resulting values summed together across traits to create the composite trait. This is visualized in Figure 2.3.

Create genetic correlation matrix

Disorder	GHQ	MDQ	EPQN
GHQ	1	0.76	0.78
MDQ	0.76	1	0.61
EPQN	0.78	0.61	1



Extract loadings from 1st unrotated PC

Disorder	1 st	2nd	3rd
GHQ	-0.60	0.04	0.80
MDQ	-0.56	-0.73	-0.39
EPQN	-0.57	0.68	-0.46

$$(zGHQ * \text{loading}) + (zMDQ * \text{loading}) + (zEPQN * \text{loading}) \rightarrow \text{Composite trait}$$

Figure 2.3. The process by which composite traits using genetic correlations were derived, using a three trait example.

Alternative composite measures were also constructed by performing principal components analysis on the traits, and using the first unrotated principal component as the trait. This was in part to observe the convergence between traits derived using the eigen function versus the prcomp function, as the latter method is preferred for numerical accuracy.

These traits were then subjected to the same set of analyses as univariate traits, using the same methodology as outlined in Sections 2.2.

2.4. Replication of results in UK Biobank

To assess whether quantitative measures of MDD had a genetic correlation with MDD status in an independent sample, quantitative trait GWAS summary statistics were assessed against MDD GWAS summary statistics from UK Biobank using bivariate LD Score regression.

2.4.1. Sample description

UK Biobank is a health research resource that aims to improve the prevention, diagnosis and treatment of a wide range of illnesses. Approximately 500,000 participants were recruited between 2006 and 2010 from across the UK. Autosomal genotype data were available for 112,151 (58,914 females; 53,237 males) aged 40 to 73 years (mean = 56.91 years, SD = 7.93). UK Biobank

received ethical approval from the Research Ethics Committee (reference: 11/NW/0382)¹⁸⁷; data used here were released to study investigators as part of project 4844. Study design and recruitment has been described previously¹⁸⁸, in brief, participants were asked to complete a touchscreen questionnaire and additional data were collected by nurse interview.

2.4.1.1. Definition of depression phenotype

Questions on depressive and manic symptoms were added to the touchscreen questionnaire during the last two years of data collection, resulting in data from 172,751 individuals. Current and previous depressive symptoms were assessed by items relating to the lifetime experience of minor and major depression, items from the Patient Health Questionnaire and items on help-seeking for mental health. 149,847 individuals provided sufficient data to allow an assessment of probable bipolar disorder and/or probable major depression. Of these, 26,847 did not meet the diagnostic criteria for a mood disorder but were excluded because of sub-threshold manic or depressive symptoms, however, it could be argued that exclusion of these individuals could lead to bias if these individuals were part of the disease spectrum. The remaining 123,000 participants were coded within mood disordered and non-mood disordered groups as shown in Figure 2.4.1.1. Further details on the UK Biobank putative depression phenotype are described in Smith *et al* 2013¹⁸⁹.

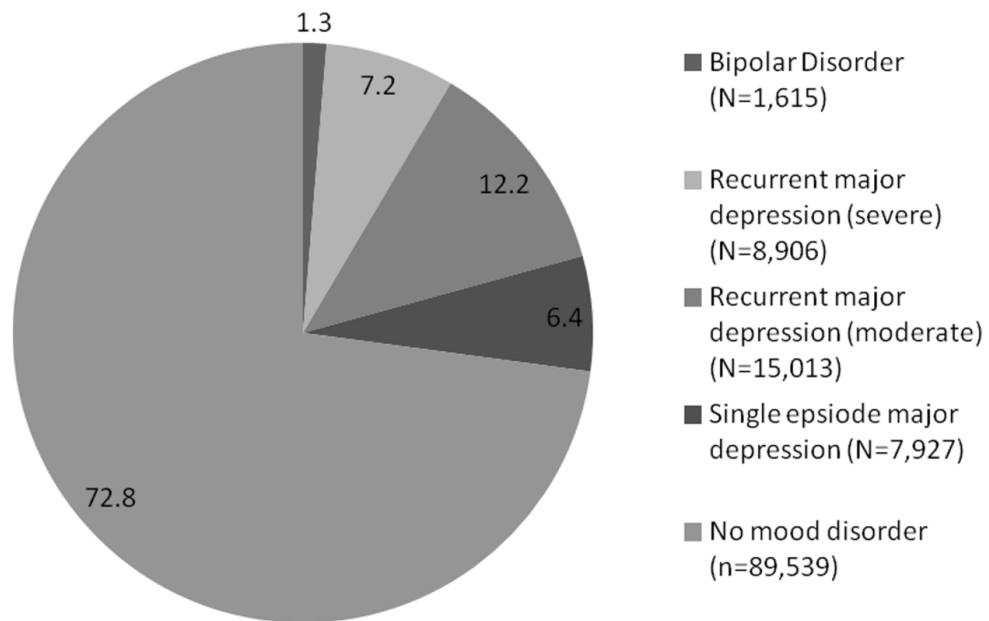


Figure 2.4.1.1. The proportion of 123,000 phenotyped participants who fell within different mood disorder diagnostic categories in UK Biobank¹⁸⁹.

2.4.1.2. Quality control

Genotyping was completed utilizing two Affymetrix arrays: BiLEVE (n=49,979) and the UK Biobank axion (n=102,750). Details of QC performed by UK Biobank have been described previously¹⁹⁰. Additional QC was applied by Dr. Mark Adams. SNPs were excluded if the MAF <1% or the imputation INFO score <0.9. Participants were retained if they were of white British ethnicity (identified via self-report and genotype data), a genetic sex match, passed UK Biobank QC criteria, were unrelated within the UK Biobank sample ($r < 0.0442$) and had array genotype missingness < 0.02. Individuals from GS:SFHS and their relatives who participated in UK Biobank were identified using a GRM and excluded from the UK Biobank sample (n=35,752).

Four depression variables were created, with varying degrees of specificity regarding case definition. These variables are outlined in table 2.4.1.2. and were created by Dr Mark Adams, who also performed the GWAS of these definitions. GWAS were performed using logistic regression in PLINK, with MDD definition as the dependent variable and age, age², sex, centre, batch, and the first 15 PCs fitted as fixed effects.

Variable	Description	Cases	Controls
depr	Putative MDD category: missing (not enough data), absent-self (MDD absent based only on self-report), absent (symptoms assessed), unspecified (have symptoms but severity unknown), single (episode), moderate (recurrent episode, seen a GP);		
mdd_strict	Putative MDD (measured): absent = 0; single, moderate, recurrent = 1	35960	91879
mdd_narrow	Putative MDD narrow definition: absent, unspecified, single = 0; moderate, recurrent = 1, else = NA	27393	136714
mdd_broad	Putative MDD broad definition: absent, absent-self = 0; unspecified, single, moderate, recurrent = 1, else = NA	72228	106014

Table 2.4.1.2. Descriptions of different case definitions, with varying levels of specificity in UK Biobank.

2.4.2. LD Score regression

The bivariate implementation of LD Score regression has been touched upon briefly in Section 1.4. Clinical Covariates. This method uses the information from GWAS summary statistics, relying on the fact that the GWAS effect-size estimate for a given SNP incorporates the effects of all SNPs in LD with that SNP, to compute genetic correlations between traits of interest. For a single study of a polygenic trait, SNPs with high LD have a higher χ^2 statistic on average than SNPs with low LD¹²³. This relationship is equivalent for the product of z-scores from two studies of traits with non-zero genetic correlation. Therefore the genetic correlation between two traits is calculated by regressing the product of z-scores onto the LD score and the slope of the regression used to estimate genetic covariance¹²³. Summary statistics from GWAS of MDD, its subtypes and quantitative measures of MDD from GS:SFHS and MDD definitions from UK Biobank were subjected to bivariate LD Score regression to assess their genetic correlation as a means of replicating results from GS:SFHS in an independent sample.

Chapter 3: Analysis of depression (as a binary trait) in Generation Scotland and UK Biobank

3.1. Introduction

In this chapter depression as a binary trait will be analyzed to address several questions. Firstly, the descriptive statistics of MDD cases and controls will be presented, as will a summary of work by Fernandez-Pujals *et al*¹⁴². These findings establish that MDD in GS:SFHS is consistent with the epidemiological literature in terms of the prevalence and heritability estimates for MDD and its more familial subtypes. Heritability estimates derived from genome-wide association analysis summary statistics, using LD Score software¹⁶⁵, will also be presented to assess whether these estimates are consistent with the literature and to confirm that the mixed model used sufficiently accounted for population stratification and the enrichment of families in GS:SFHS across MDD subtypes.

Secondly, results from genome-wide association analysis (GWAA) of depression, recurrent depression and sex-specific depression (analyzing females and males separately) will be presented. These analyses are vastly underpowered to detect genetic variants of moderate effect for MDD (GRR of 1.2 and 1.44) unless they are common in the population (MAF >20%), as stated in Section 2.1.4. However, GWAA of MDD and its subtypes are of interest due to the complicated yet relatively homogenous population structure in GS:SFHS, as outlined in Section 2.1¹⁴⁰. Most published GWAA of MDD, including a sizable meta-analysis in individuals of European descent⁴⁷, have not yielded replicable genome-wide significant findings. The only successful published GWAA of MDD to date identified two replicable genome-wide significant hits. This success was due to substantial attempts to reduce known sources of phenotypic and genetic heterogeneity⁴⁹, as described in Section 1.2.2.2. Although potential sources of heterogeneity are reduced to some extent in GS:SFHS due to the enrichment of families and moderately homogeneous ancestry, as GS:SFHS is a family- and population-based cohort, participants purposefully represent a wide range of

sociodemographic values. Data are included from participants of both sexes, ages 18-99, single and recurrent episode MDD with varying age of onset, and individuals with other illnesses which are associated with worse clinical outcomes when co-morbid with MDD¹⁴⁰ such as coronary heart disease¹¹⁸ and type 2 diabetes¹²¹. This reintroduces the phenotypic, and possibly genetic, sources of heterogeneity which have hindered progress in gene discovery in MDD to date^{18,47,99}. Also, unlike most prospective case-control studies, due to ascertainment for GS:SFHS not being on the basis of any particular illness(es) the sample is not enriched for MDD cases. However, this is not necessarily problematic as all analyses in this chapter are performed in a mixed linear model framework which can suffer a severe loss of power due to oversampling in case-control ascertainment. This is because methods for mixed-model association analysis assume that study samples are randomly ascertained with respect to the phenotype of interest¹⁸¹. Results from GWAA of MDD subtypes will be analyzed to assess whether this level of basic stratification is sufficient to yield improved genetic insights relative to undifferentiated MDD.

Next, results from polygenic profiling analysis will be presented. Since polygenic profile scores have limited predictive capacity for extensively polygenic traits using currently available sample sizes^{78,79}, this set of analyses focuses on identifying associations between psychiatric disorders and subtypes of MDD rather than prediction. This work explores whether there is an enrichment of variants identified in independent samples for MDD and other psychiatric disorders in MDD cases in GS:SFHS, and whether different subtypes of MDD have differentially shared genetic architecture with these psychiatric disorders.

Lastly, results will be presented from analyses seeking to assess how well the phenotypic definition of MDD in GS:SFHS correlates with the phenotypic definitions of MDD in UK Biobank^{187,189} - another large study conducted utilizing data from individuals of European descent. This will be achieved using GWAA summary statistics from both samples to perform a paired-samples sign test and bivariate LD score regression¹²³. This work should provide an indication as to whether it would be expected for analyses conducted in subsequent results

chapters, focusing on univariate and multivariate quantitative traits which are correlated with MDD, to have utility in populations other than GS:SFHS.

Ultimately, the results presented in this chapter serve to provide a benchmark against which quantitative traits can be assessed.

3.2. Methods

3.2.1. Descriptive statistics

Descriptive statistics for MDD were calculated in R (v3.2.3)¹⁵⁵. The sample prevalence of MDD, recurrent MDD and sex-specific MDD were calculated as proportions. Work by Fernandez-Pujals *et al*¹⁹¹ inferred an estimate of the Scottish population prevalence from the GS:SFHS sample prevalence by reweighting the sample based on population frequencies of age, sex, and the Scottish Index of Multiple Deprivation (SIMD). The distribution of sex in MDD cases and controls was estimated by calculating the proportion of each sex in each group. To assess whether there was a significant group difference, the proportion of females in each group was subjected to a Pearson's χ^2 test. Descriptive statistics for the distribution of age in MDD cases and controls were estimated using the `describeBy` function, available as part of the `psych` package¹⁹². Restricted maximum likelihood methods, implemented in `ASReml-R173, were employed to test the significance of age differences between groups. Case-control status was fitted as the dependent variable; the intercept and age were fitted as fixed effects; an additive genetic effect, common household and common sibling environment were fitted as random effects. F statistics, degrees of freedom, effect sizes, Z-scores and P-values were derived using the Wald Conditional F-test183.`

3.2.2. Heritability of depression

Historically, as stated in Section 1.2.1., the majority of studies which provide evidence for a genetic component in MDD have been conducted using twin samples¹⁸. In recent years, estimates have also been derived from unrelated

samples using genotype data^{47,98}. These methods have their limitations and can result in the overestimation (twin studies) or underestimation (SNP-based studies) of heritability estimates⁹⁹. Family studies can also be used but, unless the pedigree is large and incorporates a variety of familial relationships, may not be able to correctly partition variance into genetic and common environmental components¹⁸. Therefore estimating the heritability using pedigree information from a large mixed family- and population-based study such as GS:SFHS is useful as it provides a balance between the precision and bias of the estimate, with closer relatives and increased sample size providing greater precision and the inclusion of more distant relatives and unrelated individuals decreasing bias.

3.2.2.1. Using pedigree data

Heritability estimates for MDD, recurrent MDD (rMDD) and sex-specific MDD in females (fMDD) and males (mMDD) on the liability scale were determined by Fernandez-Pujals *et al*¹⁴² from variance components generated using Bayesian methods implemented in MCMCglmm-R¹⁷⁹, as described in Section 2.2.1.2.

For undifferentiated MDD, three models were used to incorporate different levels of model complexity. An unadjusted heritability estimate was determined using a model which fitted age and sex as fixed effects, and an additive genetic component as a random effect. This was to provide an approximate upper boundary for the heritability of MDD. This model was then adjusted to include family ID fitted as an additional random effect to model a broadly defined common family environment effect. Family ID was assigned to all known relatives in a family group (with family groups ranging from 2-36 members), including married-in individuals. The last model fitted all known sociodemographic correlates of depression which had been phenotyped (income, education, occupation, the SIMD, smoking status, alcohol use status, and cohabitation with a partner) as additional fixed effects, but only the additive genetic component as a random effect.

The heritability of MDD stratified by sex (female, male) and illness course (absent, single, recurrent) fitted age and sex as fixed effects and an additive genetic component and common family environmental as random effects. These models were also used to calculate the genetic and shared environment correlations within each stratified subset (i.e. female vs. male, single vs. recurrent). Further details of these analyses are described in Fernandez-Pujals *et al*¹⁴².

3.2.2.2. Using GWAA summary statistics

Heritability estimates for MDD, rMDD, fMDD and mMDD on the liability scale using GWAA summary statistics were determined by regressing the χ^2 value for each variant onto its LD Score, using LD Score (v1.0.0.)¹⁶⁵. Pre-computed LD Scores, which measure the amount of genetic variation tagged by each genetic variant, were used. These scores were estimated from the European-ancestry samples in the 1000 Genomes Project¹⁶¹, and are appropriate for basic LD Score analyses such as calculating the LD Score regression intercept, heritability estimates and the genetic correlation between traits¹⁶⁵. Summary statistics from GWAA in GS:SFHS were harmonized with the LD scores using the munge.py function. To obtain heritability estimates on the liability scale, rather than the default of the observed scale, sample and population prevalence estimates were provided. Sample prevalence estimates for MDD, rMDD, fMDD and mMDD were as described in Section 3.2.1. This yielded sample prevalence estimates in GS:SFHS of 13.4% for MDD, 7.1% for rMDD, 16.2% for fMDD and 9.3% for mMDD. The population prevalence estimates were derived from the literature. For MDD, the prevalence estimated by Kessler *et al*¹ of 16.2% was used. Extensive literature searching demonstrated a paucity of information regarding estimates for the lifetime prevalence of rMDD. Therefore the prevalence estimated, albeit from a very small population sample (n=493), by Altamura *et al*¹⁹³ of 6.9% was used. For sex specific MDD, information from Weissman *et al*¹⁰ - which estimated the prevalence of MDD in females and males across ten different sites, was used. Prevalence estimates across the ten sites are shown in Figure 3.2.2.2. As this study did not provide information on the racial

distribution in these samples and prevalence estimates varied substantially between countries, prevalence estimates for France, Germany and Italy were averaged to derive an approximate prevalence estimate for sex-specific MDD in the European population. This resulted in an estimated prevalence of 17.8% for fMDD and 7.0% for mMDD.

Table 2.—Lifetime and Annual Rates and Onset Age for Major Depression, Ages 18 to 64 Years*

	Annual Rate/ 100 (SE)†	Lifetime Rate/100 (SE)				Mean Age at Onset, y (SE)
		Overall	Females	Males	F/M Ratio	
United States	3.0 (0.18)	5.2 (0.24)	7.4 (0.39)	2.8 (0.26)	2.6 (.11)	25.6 (0.30)
Edmonton, Alberta	5.2 (0.45)	9.6 (0.60)	12.3 (0.93)	6.6 (0.73)	1.9 (.13)	24.8 (0.52)
Puerto Rico	3.0 (0.49)	4.3 (0.59)	5.5 (0.91)	3.1 (0.72)	1.8 (.29)	29.5 (1.19)
Paris, France	4.5 (0.65)	16.4 (1.16)	21.9 (1.80)	10.5 (1.39)	2.1 (.16)	29.2 (0.52)
West Germany‡	5.0 (1.13)	9.2 (1.50)	13.5 (2.46)	4.4 (1.56)	3.1 (.39)	29.7 (1.18)
Florence, Italy	...	12.4 (1.33)	18.1 (2.16)	6.1 (1.40)	3.0 (.26)	34.8 (1.12)
Beirut, Lebanon	...	19.0 (1.76)	23.1 (2.63)	14.7 (2.25)	1.6 (.19)	25.2 (1.00)
Taiwan	0.8 (0.09)	1.5 (0.12)	1.8 (0.19)	1.1 (0.16)	1.6 (.17)	29.3 (1.04)
Korea	2.3 (0.22)	2.9 (0.24)	3.8 (0.38)	1.9 (0.29)	2.0 (.18)	29.3 (0.88)
Christchurch, New Zealand	5.8 (0.70)	11.6 (0.96)	15.5 (1.51)	7.5 (1.14)	2.1 (.18)	27.3 (0.58)

*Figures standardized to US age and sex distribution.

†Ellipses indicate data not available.

‡Data from former Federal Republic of Germany (West Germany) based on ages 26 to 64 years.

Figure 3.2.2.2. Lifetime prevalence estimates for MDD and sex specific MDD across ten different population sites from Weissman et al¹⁰.

This analysis was also performed to assess the regression intercept and the ratio returned by LD Score, which measures the proportion of inflation in the mean χ^2 that the LD Score regression intercept ascribes to causes other than polygenic heritability, i.e. confounding caused by stratification or enrichment of relatives¹⁶⁵. This was to ensure that the mixed linear model used was adequately controlling for population structure and enrichment of relatives in all MDD subtypes.

3.2.3. Genome-wide association analysis

GWAA of MDD, rMDD, fMDD and mMDD were conducted using mixed linear model based association analysis (MLMA), implemented in GCTA (v1.25.)^{166,167}. An overview of how this analysis was performed is given in Section 2.2.5.; some additional details shall be provided here. The mixed linear model used can be written as:

$$y = X\beta + Zu_1 + Zu_2 + \varepsilon$$

where y is the vector of binary observations for MDD, rMDD, fMDD or mMDD. β is a matrix of fixed effects: sex, age, age² and the genotype of the SNP being tested for association (coded additively as 0, 1, 2 reflecting the number of copies of the minor allele). u_1 and u_2 are random effects taking into account the genomic relationships, following a multivariate normal distribution of mean 0, and variance $G\sigma_u^2$, where G was the genomic relationship matrix. X , Z , and Z are design matrices of regressors which relate the observations of y to β , u_1 and u_2 respectively. ε is an unknown vector of random errors¹⁷³, and assumed to follow a multivariate normal distribution of mean zero and variance $I\sigma_\varepsilon^2$, where I is an identity matrix.

Two GRMs, as described in Section 2.1.3.3.2., were used to account for population structure and potential upward biases due to excessive relationships, thus allowing the inclusion of closely and distantly related individuals in genetic analyses¹⁶⁹. This is invaluable as using only genetically unrelated individuals would reduce the sample size from 19,994 to 7,388 individuals. The GRMs were constructed using the MLMe approach, also described in Section 2.1.3.3.2. Whilst GCTA does have a function to run MLMe analysis (`--mlma-loco`), this function does not presently support the use of multiple GRMs. Therefore this analysis was performed using the standard mixed linear model association analysis function (`--mlma`)¹⁶⁷, with GRMs constructed manually for each chromosome. As MLMA employs restricted maximum likelihood methods, the regression is conducted on the linear scale, rather than the liability scale. As such, test statistics (betas and their corresponding standard errors) were transformed to Odds Ratios (ORs) and their corresponding 95% Confidence Intervals (CI) on the liability scale using the Taylor transformation^{170,171}. The code for this transformation was provided by Naomi Wray and Jian Yang (personal communication), and is supplied in Appendix 3.A.

3.2.4. Polygenic profiling analysis

Polygenic profile, or polygenic risk, scores (PGRS) for major depressive disorder, autism, bipolar disorder type 1 and schizophrenia were derived for all genotyped individuals as described in Section 2.2.6. The association between PGRS and MDD, rMDD, fMDD and mMDD was tested by regressing the PGRS onto the phenotype, covarying for age, age², sex (for MDD and rMDD), and the first twenty PCs. An additive genetic component, common household and common sibling environment were fitted as random effects. F statistics, degrees of freedom, effect sizes, Z-scores and P-values were derived using the Wald Conditional F-test¹⁸³. As this analysis was performed on the linear scale, again, test statistics were transformed to the liability scale as described in Section 3.2.3. Variance in trait value explained by polygenic profile score was estimated by multiplying the profile score by its corresponding regression coefficient and estimating its variance. This value was then divided by the variance of the observed phenotype to yield a coefficient of determination between 0 and 1¹⁸⁶, and multiplied by 100 to convert R² to percentage variance explained.

3.2.5. Replication in UK Biobank

How well the MDD phenotype in GS:SFHS correlated with the MDD phenotype in UKBiobank (UKB) was determined by a paired-samples sign test and bivariate LD score regression¹²³.

ORs of SNPs with a p-value <1e-04 from GWAS of MDD, rMDD, fMDD and mMDD in GS:SFHS were extracted from GWAS summary statistics of broad, narrow, and strict MDD definitions in UKB. These SNPs were subjected to a paired-samples sign test using the binom.test function in R (v3.2.3)¹⁵⁵ to assess whether the ORs from UKB were in the same direction as in GS:SFHS more or less frequently than expected by chance (p=0.05).

The genetic correlation between pairs of GWAS summary statistics was estimated using the rg function in LD Score (v1.0.0.)¹²³. This was performed on the liability scale, to ensure results were comparable across studies with different prevalence estimates. The population prevalence and sample

prevalence in GS:SFHS for MDD, rMDD, fMDD and mMDD are as described for the univariate LD Score Regression in Section 3.2.2.2. Sample prevalence for UKB was estimated as 40.5% for the broad MDD definition, 16.7% for the narrow MDD definition, and 28.1% for the strict MDD definition (as outlined in Table 2.4.1.1.).

3.3. Results

3.3.1. Descriptive statistics

Descriptive statistics for the full GS:SFHS (n=23,960) sample are described by Smith *et al*¹⁴⁰. Statistics here refer to the genotyped sample after quality control measures, as described in Section 2.1.3. (n=19,994).

19,896 individuals provided data on MDD status, or screened negative for depression (2,659 cases; 17,237 controls), therefore 13.4% of the genotyped sample had a diagnosis of undifferentiated depression, either single or recurrent episode. This was estimated by Fernandez-Pujals *et al* to correspond to an estimated population prevalence of 12.2% (CI: 11.4-13.1%)¹⁹¹, which is within the range of prevalence estimates (8.2% - 16.9%) inferred from recent studies in continental Europe, the USA, and Canada^{1,194,195}. The sample prevalence for rMDD was 7.1% (1,327 cases; 17,237 controls). An inferred population prevalence was not calculated for rMDD, however the sample prevalence is consistent with the population prevalence estimated by Altamura *et al*¹⁹³, as stated in Section 3.2.2.2. The sample prevalence for fMDD was 16.2% (1,894 cases; 9,817 controls), with an inferred population prevalence of 15.8% (CI: 14.7-16.8%)¹⁴². The sample prevalence for mMDD was 9.3% (765 cases; 7,420 controls), with an inferred population prevalence of 9.1% (CI: 8.3-9.9%)¹⁴². Sex-specific sample and inferred population prevalence estimates for MDD in GS:SFHS are consistent with other sex-specific population prevalence estimates as outlined in Section 3.2.2.2. Overall these results suggest that GS:SFHS is congruent with other European population samples of MDD.

GS:SFHS has a higher proportion of females (58.9% of the 19,992 genotyped individuals who had information on sex) than the Scottish population (51%)¹⁹⁶. Sex distributions within the full genotyped sample, and MDD controls and cases are shown in Table 3.3.1.A. The case group had a significantly higher proportion of females (71.2%) than the control group (60.0%), as determined by the Pearson's χ^2 test ($\chi^2 = 193.3$, $df = 1$, $p\text{-value} < 2.2e-16$). This is consistent with the epidemiological literature, with the lifetime prevalence being approximately two times greater in females than males¹⁰. The proportion of females in rMDD cases was the same as in undifferentiated MDD.

Sex distributions				
	Total n	Males	Females	Prop. Female
GS:SFHS	19,992	8,220	11,772	58.9%
Controls	17,237	7,420	9,817	60.0%
MDD	2,569	765	1,894	71.2%

Table 3.3.1.A. Distribution of males and females in the genotyped GS:SFHS sample, and controls and MDD cases.

The median age in GS:SFHS (49 years) is higher than the median age of the Scottish population (41 years)¹⁹⁶. This is in part due to GS:SFHS being recruited primarily at middle age, and the requirement that participants must be over 18 years old¹⁴⁰. The age of participants in the control group ranged from 18 to 99 years (mean age = 47.6 years, $SD = 15.3$), whilst age in the case group ranged from 18-84 years (mean = 46.4 years, $SD = 12.9$). Age in both controls and cases was non-normally distributed, with data negatively skewed more than 2 standard errors of skewness from the mean (controls = -0.11, cases = -0.18) and demonstrating platykurtosis more than 2 standard errors of kurtosis from the mean (controls = -0.69, cases = -0.62). Age distributions are detailed in Table 3.3.1.B, and Figures 3.3.1.B and 3.1.1.C.

Age distributions						
	Min.	1st Qu.	Median	Mean (S.D.)	3rd Qu.	Max.
Control	18	36	49	47.6 (15.3)	59	99
MDD	18	37	47	46.4 (12.9)	57	84

Table 3.3.1.B. Distribution of age in controls and MDD cases is GS:SFHS.

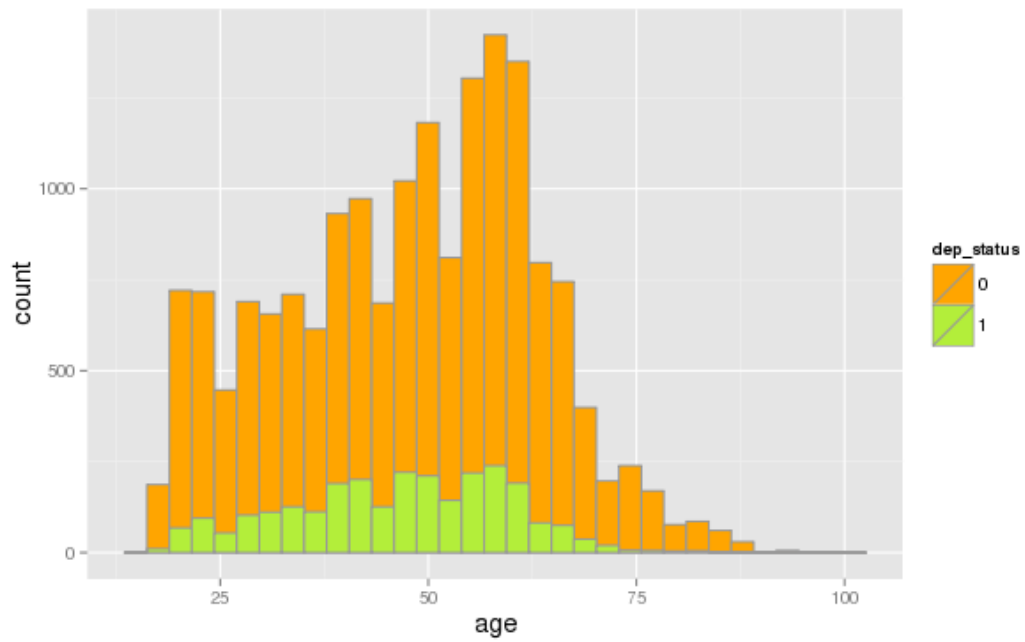


Figure 3.3.1.B. Histogram of the distribution of age in controls (dep_status = 0) and MDD cases (dep_status = 1), demonstrating negative skew and platykurtosis.

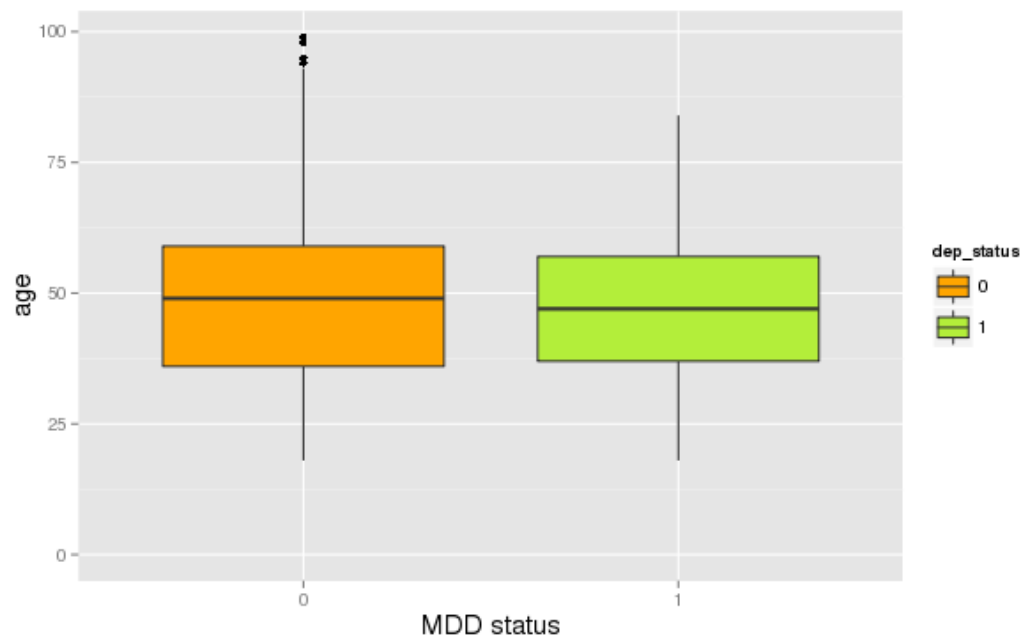


Figure 3.3.1.C. Box and whiskers plot showing the distribution of age in controls (dep_status = 0) and MDD cases (dep_status = 1).

There was a small but significant difference in age distribution between cases and controls, as determined by Wald's Conditional F-Test: $\beta(\text{SE}) = -0.0006(0.0002)$, $F_{(1,18475)} = 16.5$, $p = 4.87 \times 10^{-5}$, indicating that cases are significantly younger than controls. This could be due to underreporting of depression in older subjects¹⁹⁷.

3.3.2. Heritability of depression

3.3.2.1. Using pedigree data

This Section (3.3.2.1.) presents work from Fernandez-Pujals *et al*¹⁹¹. These results are relevant to the work presented in this Chapter which examines MDD and its subtypes using analyses which incorporate genotype data. Therefore it is useful to be able to draw comparisons between this work, and work which has been performed using the pedigree data. Also, the heritability of MDD estimated from pedigree data is used in the estimation of coheritability between quantitative traits and MDD in subsequent chapters. Lastly, as outlined in the introduction to this Chapter, these results put GS:SFHS into the context of the existing literature. Table 3.3.2.1.A. shows the number of informative pairwise familial relationships used in the pedigree-based heritability analysis.

Relationship dyad	N
Total	28,040
Mother-Offspring	14,441
Father-Offspring	13,648
Full siblings	9,038
Half siblings with shared mother	9,126
Half sibling with shared father	9,056
Maternal grandmother-grandchild	2,980
Maternal grandfather-grandchild	2,913
Paternal grandmother-grandchild	1,627
Paternal grandfather-grandchild	1,642

Table 3.3.2.1.A. Number of informative family relationships for MDD case control status in GS:SFHS (adapted from Fernandez-Pujals *et al*¹⁹¹).

The unadjusted heritability of MDD in GS:SFHS was 0.44 (CI: 0.37-0.52) on the liability scale. Adjusting for known sociodemographic correlates of MDD as

additional fixed effects did not substantially change the heritability estimate, reducing it to 0.41 (CI: 0.32-0.50). However, adding a broadly defined common environmental effect substantially reduced the heritability to 0.28 (CI: 0.12-0.47). Sex-specific heritability was not significantly different ($p=0.58$) for females ($h^2 = 0.44$, CI:0.25-0.61) and males ($h^2 = 0.35$, CI:0.08 - 0.63). Recurrent MDD was significantly ($p < 0.0005$) more heritable ($h^2 = 0.41$, CI: 0.20 - 0.60) than single episode MDD ($h^2 = 0.28$, CI: 0.14 - 0.41). MDD in females and males were highly genetically correlated ($r_g = 0.75$, CI: 0.43 - 0.99), as were single and recurrent MDD ($r_g = 0.86$, CI:0.68 - 0.97)¹⁹¹. These results are shown in Table 3.3.2.1.B.

Trait	h^2 (CI)	r_g (CI)	c^2 (CI)	r_c (CI)
MDD ^a	0.44 (0.37, 0.52)	-	-	-
MDD ^b	0.28 (0.12, 0.47)	-	0.07 (0.01, 0.15)	-
MDD ^c	0.41 (0.32, 0.50)	-	-	-
MDD in females	0.44 (0.25, 0.61)	0.75 (0.43, 0.99)	0.04 (0.00, 0.13)	0.52 (-0.70, 1.00)
MDD in males	0.35 (0.08, 0.63)		0.08 (0.00, 0.20)	
Single episode	0.28 (0.14, 0.41)	0.86 (0.68, 0.97)	0.03 (0.00, 0.08)	0.50 (-0.48, 0.93)
Recurrent episode	0.41 (0.20, 0.60)		0.07 (0.00, 0.16)	

Table 3.3.2.1.B. Point estimates and 95% confidence intervals (CI) of the contribution of additive genetic variance (h^2) and common family environment (c^2) for undifferentiated MDD, and MDD stratified by sex and illness course; genetic (r_g) and environmental correlations (r_c) within stratified subsets. MDD^a - unadjusted model, MDD^b - model adjusting for common family environment, MDD^c - model adjusting for known sociodemographic correlates.

Heritability estimates from GS:SFHS using pedigree data are in a similar range to heritability estimates from the Sullivan *et al*¹⁶ meta-analysis estimate of 37% (CI: 31–42), with confidence intervals overlapping between all estimates. Although recurrent MDD is frequently cited as being more familial or heritable,

a heritability estimate for this subset could not be identified in the literature. However, it did prove to be more heritable than single episode MDD in GS:SFHS. Sex-specific heritability estimates differed to a similar extent comparative to estimates from Kendler et al (40% versus 30% and 42% versus 29%)^{20,21}, although the genetic correlation between sexes in GS:SFHS (+0.75) was higher than in Kendler et al (+0.55 and +0.63^{20,21}), perhaps suggesting fewer sex specific mechanisms for MDD in GS:SFHS. Therefore GS:SFHS is relatively consistent with the literature in terms of heritability estimates from pedigree information.

For subsequent analyses incorporating the h^2 of MDD, the estimate of 0.44 from the unadjusted model was taken forward. This model was favoured over the model which included a broad common family environment effect, as this variable was assigned to all known family members and therefore assumes that all close, distant and married-in relatives are subject to the same common environment. This could dilute true family environmental covariance by including many relationships outside the nuclear family which do not share the environment and hence have zero environmental covariance. This coding could also incorrectly partition variance away from the additive component, resulting in a downward bias in heritability estimates. Additionally, Section 2.2.1.1.4. demonstrated that shared household and shared sibling environment do not significantly improve model fit for MDD, further supporting the use of the estimate from the unadjusted model for coheritability analysis.

3.3.2.2. Using GWAA summary statistics

For undifferentiated MDD, $h^2_{\text{SNP}} = 0.05$, although the standard error was 0.06, therefore the estimate was not significantly different from zero. This was also the case for mMDD, where $h^2_{\text{SNP}} (\text{SE}) = 0.07 (0.16)$. However, for mMDD, perhaps this is in part due to the reduced sample size ($n=8,185$, of which 765 were cases). Both rMDD and fMDD had low SNP heritability estimates, although in both instances these results were significantly different from zero with h^2

(SE) for rMDD = 0.14 (0.10), and fMDD = 0.19 (0.10). Results from univariate LD Score regression of MDD and its subsets are shown in Table 3.3.2.2.

Trait	h² (SE)	λ_{GC}	μχ²	Intercept (SE)	Ratio (SE)
MDD	0.05 (0.06)	1.002	1.00	0.990 (0.007)	NA
rMDD	0.14 (0.10)	1.011	1.012	0.9998 (0.008)	<0
fMDD	0.19 (0.10)	1.011	1.013	0.997 (0.008)	<0
mMDD	0.07 (0.16)	1.005	0.998	0.995 (0.007)	NA

Table 3.3.2.2. Output from univariate LD Score regression using summary statistics from MDD, rMDD fMDD and mMDD GWAS in GS:SFHS. h² (SE): SNP based heritability estimate on the liability scale with its standard error; λ_{GC}: genomic control, estimated using median(chi²)/0.4549; μχ²: mean chi-squared statistic; Intercept (SE): LD Score regression intercept with its standard error; Ratio (SE): proportion of inflation ascribed to causes other than polygenic heritability with its standard error.

Only two SNP based heritability estimates were found in the literature, h²_{SNP}=0.21¹⁹, and h²_{SNP}=0.05¹⁹⁸, both from samples analyzing undifferentiated MDD with a sample prevalence of 0.49 and 0.25 respectively, and implementing a population prevalence of 0.15. Higher heritabilities in rMDD and fMDD are consistent with the results from the heritability analysis using pedigree data, although admittedly the estimates are much lower, as to be expected. These estimates may also be low due to the analysis only incorporating genotyped SNPs, rather than imputed data, as this resulted in 465,678 SNPs in common between GS:SFHS and the LD score SNP list. For all analyses, the mean chi-squared statistic (μχ²) was under 1.02, meaning that there is very little polygenic signal to work with. This is reflected in the ratio (intercept - 1)/(μχ² - 1), which should be close to zero, but is less than zero across subtypes. This usually indicates that genomic correction has been performed¹⁹⁹, however that is not the case in our data. Possible reasons for the low μχ² and ratio include using raw genotype data and the mixed linear model could be too conservative. These points shall be discussed further in subsequent chapters after assessing results from univariate LD Score regression on quantitative traits.

3.3.3. Genome-wide association analysis

Genome-wide association analysis of MDD, rMDD, fMDD and mMDD yielded no genome-wide significant hits. The ratio of the observed median χ^2 to that expected by chance, denoted λ , indicated that the observed $-\log(P)$ for MDD and mMDD were slightly deflated to those expected under the null, with λ (SE) values for MDD = 0.997 (7.97e-06) and mMDD = 0.995 (7.72e-06). Observed $-\log(P)$ for rMDD and fMDD were slightly inflated compared to those expected under the null with λ values of 1.013 (7.90e-06) for rMDD and 1.015 (4.97e-06) for fMDD. However, as stated in Section 3.3.5., the ratio from LD Score regression was <0 for rMDD and fMDD, indicating that this inflation is owing to polygenicity rather than confounding by population and family structure¹⁶⁵ in GS:SFHS. Quantile-quantile plots of observed $-\log(P)$ against expected $-\log(P)$ and Manhattan plots of genomic coordinates against $-\log(P)$ are shown for MDD, rMDD, fMDD and mMDD in Figures 3.3.3.1. - 3.3.3.4. The 20 most significantly associated SNPs for each MDD definition are shown in Tables 3.3.3.1. - 3.3.3.4.

3.3.3.1. Summary statistics from GWAS of MDD

CHR	SNP	BP	A1	A2	MAF	OR (CI)	P
11	rs882151	75904137	A	G	0.21	1.19 (1.12,1.26)	1.19e-06
10	rs1331328	18532672	A	C	0.17	0.81 (0.72,0.90)	4.98e-06
5	rs39700	66162044	G	A	0.38	0.85 (0.79,0.92)	5.32e-06
13	rs7991353	72545858	G	A	0.24	0.84 (0.76,0.92)	6.83e-06
9	rs7867936	132314191	A	G	0.38	1.15 (1.08,1.21)	8.08e-06
12	rs10771350	27760130	G	A	0.22	0.84 (0.76,0.92)	8.99e-06
13	rs7983339	72539392	G	A	0.22	0.84 (0.76,0.92)	1.51e-05
9	rs7868056	132314273	A	G	0.38	1.14 (1.08,1.20)	1.64e-05
2	rs6741418	191723719	A	G	0.16	1.19 (1.11,1.27)	1.92e-05
8	rs10086390	132131683	G	A	0.08	1.26 (1.15,1.36)	2.36e-05
5	rs4868136	171472575	G	A	0.41	0.87 (0.80,0.94)	2.85e-05
10	rs1757236	18527454	G	A	0.17	0.83 (0.74,0.92)	3.05e-05
6	rs9942465	35324213	A	C	0.02	1.52 (1.32,1.71)	3.10e-05
2	rs7593100	183694499	A	G	0.07	1.26 (1.15,1.37)	3.16e-05
6	rs9490199	121677165	G	A	0.13	1.20 (1.11,1.28)	3.22e-05
5	rs13172324	7368845	A	G	0.09	1.23 (1.13,1.33)	3.46e-05
20	rs6093898	42566577	G	A	0.10	0.79 (0.67,0.90)	3.65e-05
20	rs6031245	42559531	A	G	0.10	0.79 (0.67,0.90)	3.73e-05
7	rs17711148	14332000	A	G	0.24	0.85 (0.78,0.93)	4.13e-05
4	rs6843595	4351730	G	A	0.25	1.15 (1.08,1.22)	4.39e-05

Table 3.3.3.1. Top 20 SNPs from GWAS of Major Depressive Disorder, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), odds ratio with 95% confidence intervals (OR (CI)), and association p-value (P).

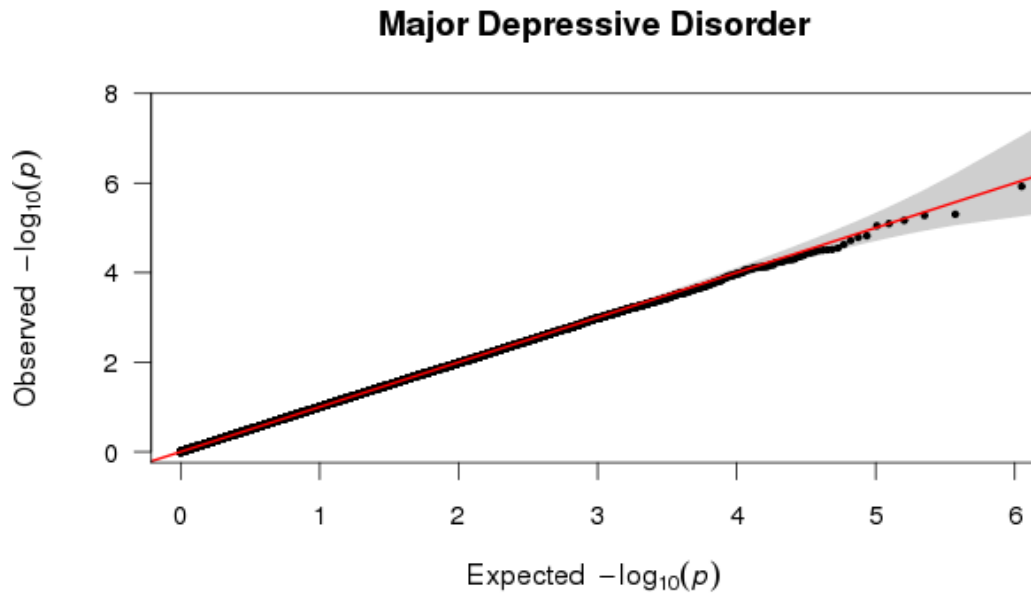


Figure 3.3.3.1.A. Quantile-quantile plot for Major Depressive Disorder, showing expected distribution of GWAS test statistics, $-\log_{10}(p)$, versus the observed distribution. $\lambda(\text{SE}) = 0.997$ ($7.97\text{e-}06$).

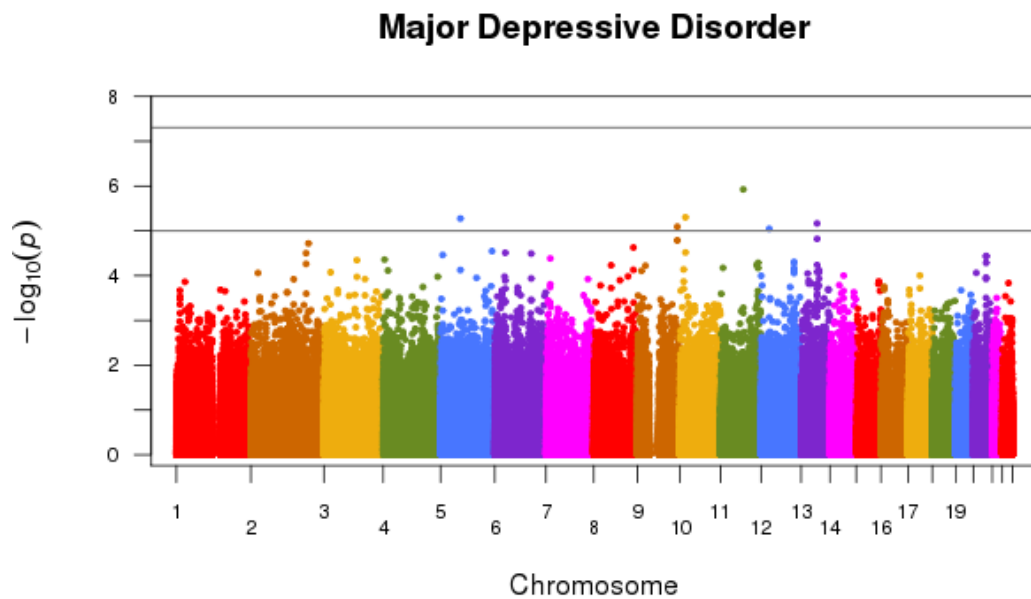


Figure 3.3.3.1.B. Manhattan plot of genomic coordinate against test statistic for association from GWAS of Major Depressive Disorder, showing no genome-wide significant hits.

3.3.3.2. Summary statistics from GWAS of recurrent MDD

CHR	SNP	BP	A1	A2	MAF	OR (CI)	P
16	rs11076064	55033694	G	A	0.49	0.80 (0.70,0.89)	1.68e-06
6	rs9942465	35324213	A	C	0.02	1.79 (1.55,1.03)	2.43e-06
1	rs6668724	14623470	G	A	0.05	1.46 (1.30,1.63)	5.91e-06
6	rs10946731	24844443	A	G	0.13	0.72 (0.57,0.86)	7.64e-06
6	rs6901565	24836161	C	A	0.13	0.72 (0.58,0.86)	8.12e-06
12	rs1005374	3436701	G	A	0.41	1.20 (1.12,1.27)	1.04e-05
9	rs795085	78331204	C	A	0.31	1.21 (1.12,1.29)	1.10e-05
1	rs12725447	113810931	A	G	0.24	0.78 (0.67,0.89)	1.15e-05
6	rs7766847	169398467	A	G	0.40	1.19 (1.11,1.27)	1.19e-05
4	rs2024530	181707386	A	G	0.40	0.81 (0.72,0.90)	1.21e-05
5	rs13172324	7368845	A	G	0.09	1.33 (1.20,1.46)	1.52e-05
13	rs9515125	110548423	G	A	0.44	1.19 (1.11,1.27)	1.61e-05
12	rs10735017	3228973	G	A	0.17	1.25 (1.15,1.35)	1.69e-05
6	rs4499887	79979700	G	A	0.47	1.19 (1.11,1.27)	1.71e-05
1	rs4047801	183415318	A	G	0.49	1.19 (1.11,1.26)	1.78e-05
13	rs7983339	72539392	G	A	0.22	0.78 (0.67,0.89)	1.88e-05
1	rs7529857	213811721	C	A	0.06	0.62 (0.40,0.84)	1.90e-05
6	rs2322217	79959909	G	A	0.47	1.19 (1.11,1.26)	2.00e-05
4	rs7685363	96640243	A	G	0.45	1.19 (1.11,1.26)	2.01e-05
17	rs227756	54752709	A	G	0.28	1.21 (1.12,1.29)	2.11e-05

Table 3.3.3.2. Top 20 SNPs from GWAS of recurrent Major Depressive Disorder, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), odds ratio with 95% confidence intervals (OR (CI)), and association p-value (P).

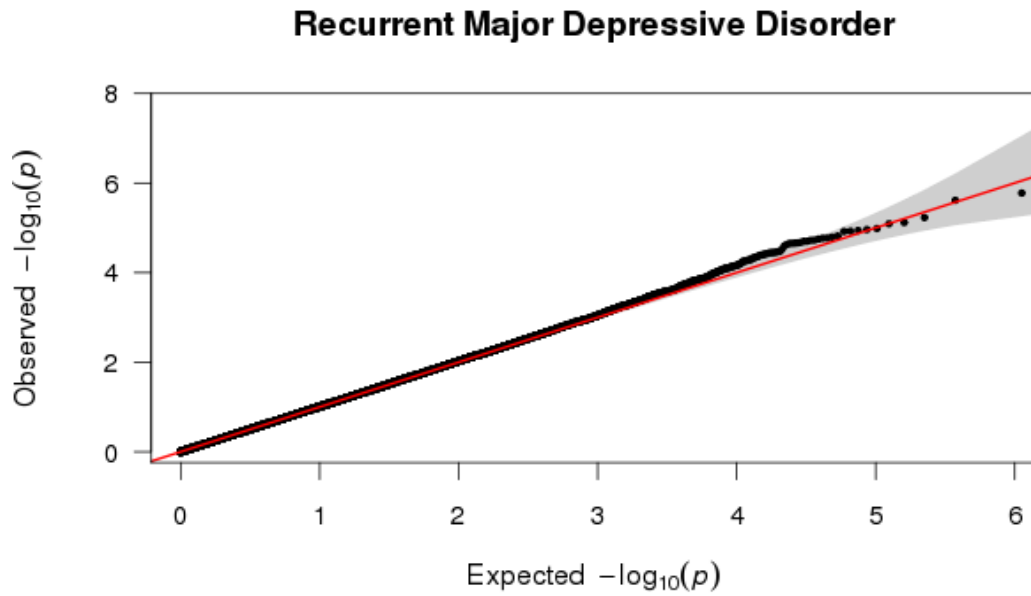


Figure 3.3.3.2.A. Quantile-quantile plot for recurrent Major Depressive Disorder, showing expected distribution of GWAS test statistics, $-\log_{10}(p)$, versus the observed distribution. $\lambda(\text{SE}) = 1.013$ ($7.90\text{e-}06$).

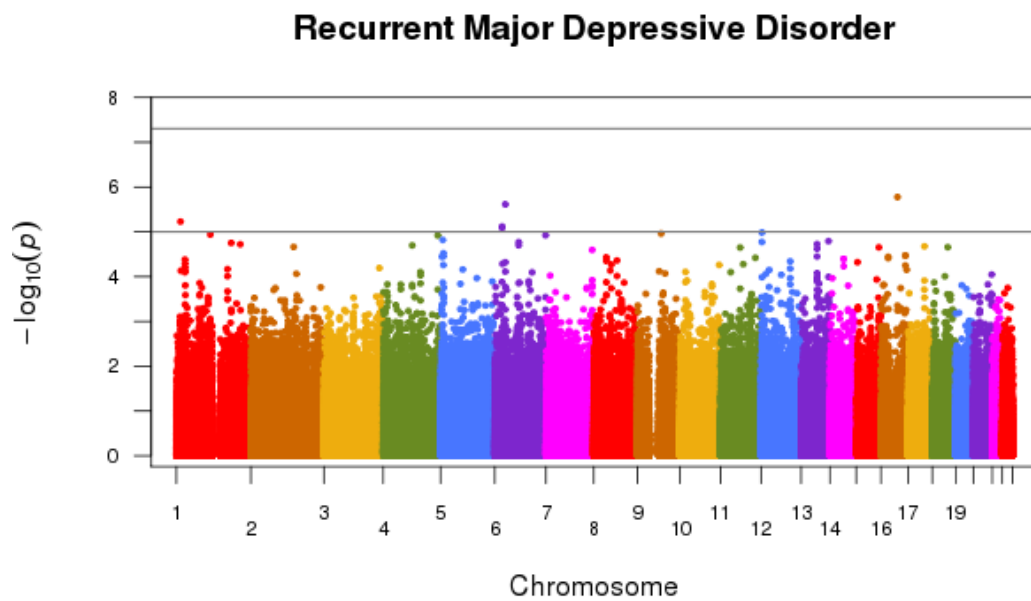


Figure 3.3.3.2.B. Manhattan plot of genomic coordinate against test statistic for association from GWAS for recurrent Major Depressive Disorder, showing no genome-wide significant hits.

3.3.3.3. Summary statistics from GWAS of MDD in females

CHR	SNP	BP	A1	A2	MAF	OR	P
6	rs1745580	114619795	G	A	0.34	1.19 (1.12,1.26)	2.01e-06
6	rs785129	114612895	A	G	0.34	1.19 (1.12,1.26)	2.45e-06
6	rs785144	114628139	C	A	0.34	1.18 (1.11,1.26)	3.59e-06
6	rs9942465	35324213	A	C	0.02	1.70 (1.47,1.93)	5.23e-06
2	rs1530046	23623283	A	G	0.02	1.59 (1.39,1.80)	5.90e-06
7	rs12699544	13780997	A	G	0.16	1.23 (1.14,1.32)	6.84e-06
5	rs4868136	171472575	G	A	0.41	0.84 (0.76,0.92)	1.15e-05
17	rs17764251	4214005	A	G	0.49	0.84 (0.76,0.92)	1.20e-05
12	rs598013	4072162	A	G	0.40	0.84 (0.76,0.92)	1.37e-05
7	rs6949771	13792450	A	G	0.11	1.26 (1.16,1.37)	1.41e-05
8	rs2055322	129138272	G	A	0.45	1.17 (1.10,1.23)	1.47e-05
5	rs39700	66162044	G	A	0.38	0.84 (0.76,0.92)	1.61e-05
12	rs7309194	39207946	A	G	0.07	1.33 (1.20,1.46)	1.78e-05
11	rs1630030	125301587	A	G	0.15	1.23 (1.13,1.32)	2.16e-05
3	rs9817638	103298212	G	A	0.19	1.20 (1.12,1.29)	2.50e-05
6	rs785143	114645237	G	A	0.35	1.17 (1.09,1.24)	2.55e-05
11	rs645172	125302722	G	A	0.15	1.22 (1.13,1.32)	2.78e-05
12	rs10771350	27760130	G	A	0.22	0.82 (0.73,0.91)	3.23e-05
10	rs1133400	134459388	G	A	0.22	1.19 (1.11,1.27)	3.53e-05
13	rs1555544	53825161	G	A	0.06	0.71 (0.54,0.87)	3.87e-05

Table 3.3.3.3. Top 20 SNPs from GWAS of Major Depressive Disorder in females only, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), odds ratio with 95% confidence intervals (OR (CI)), and association p-value (P).

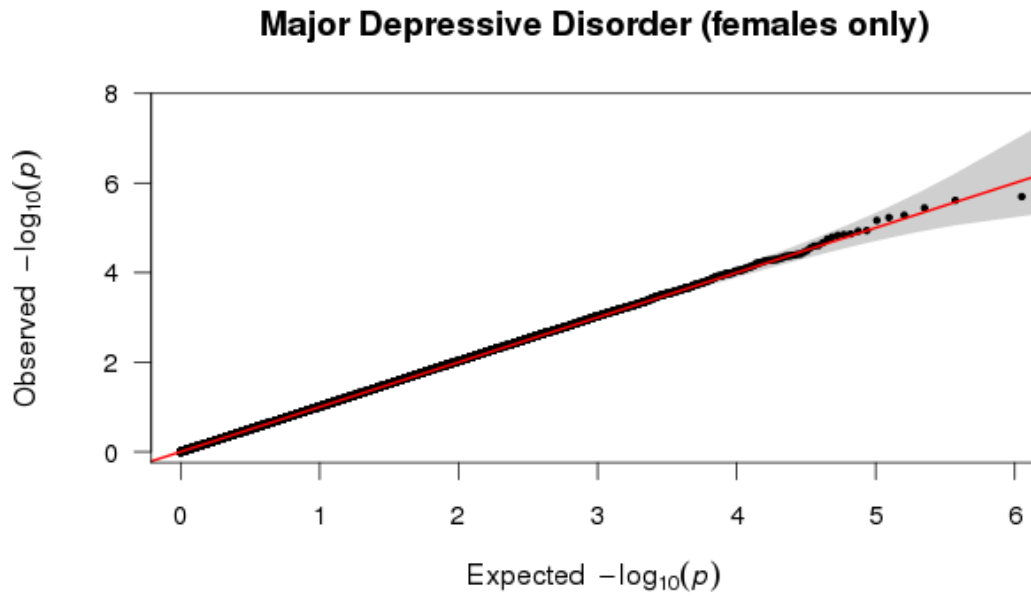


Figure 3.3.3.3.A. Quantile-quantile plot for Major Depressive Disorder in females only, showing expected distribution of GWAS test statistics, $-\log_{10}(p)$, versus the observed distribution. $\lambda(\text{SE}) = 1.015$ ($4.97\text{e-}06$).

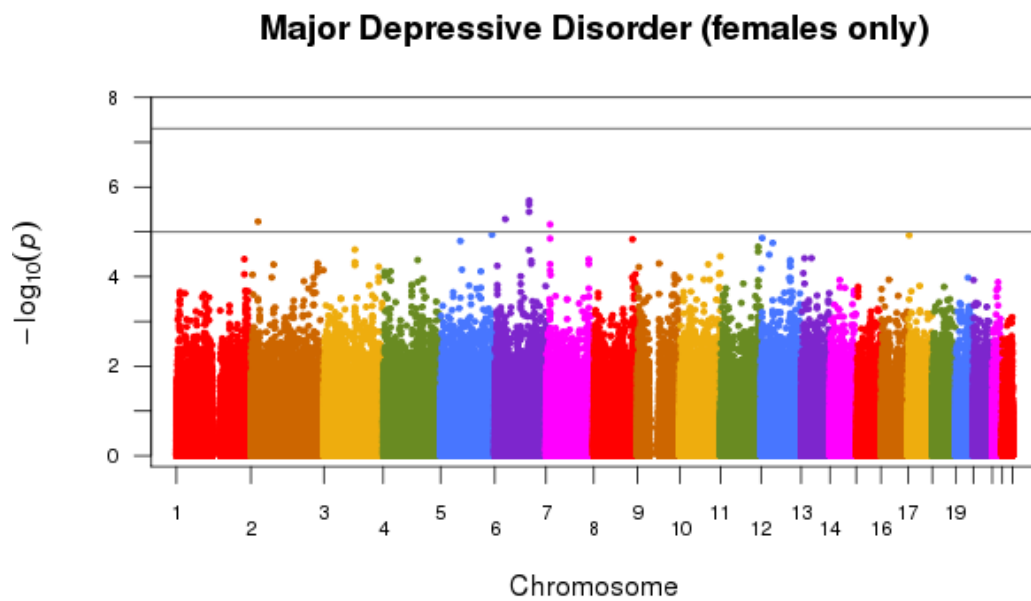


Figure 3.3.3.3.B. Manhattan plot of genomic coordinate against test statistic for association from GWAS for Major Depressive Disorder in females only, showing no genome-wide significant hits.

3.3.3.4. Summary statistics from GWAS of MDD in males

CHR	SNP	BP	A1	A2	MAF	OR (CI)	P
19	rs2307187	44079611	A	G	0.01	2.44 (2.08,2.79)	8.47e-07
9	rs7035578	78745177	A	G	0.16	1.39 (1.25,1.52)	1.33e-06
19	rs10409248	28531174	A	G	0.03	1.89 (1.62,2.15)	2.92e-06
12	rs1849854	79381967	G	A	0.05	1.65 (1.43,1.86)	6.71e-06
1	rs4272665	81891826	G	A	0.08	1.47 (1.30,1.65)	1.25e-05
14	rs2295306	24765684	G	A	0.49	0.76 (0.63,0.88)	1.48e-05
8	rs7822917	32190881	A	C	0.28	0.74 (0.60,0.88)	1.81e-05
14	rs2416024	47585469	A	G	0.29	1.27 (1.16,1.38)	2.37e-05
7	rs11761864	104218575	G	A	0.27	0.74 (0.60,0.88)	2.51e-05
8	rs4284046	22712077	A	G	0.16	1.33 (1.19,1.46)	2.67e-05
12	rs1514920	79396782	G	A	0.05	1.59 (1.38,1.81)	2.71e-05
7	rs2215781	153530887	A	G	0.26	1.27 (1.16,1.39)	3.15e-05
2	rs4972969	231556663	A	G	0.37	1.25 (1.14,1.36)	3.34e-05
2	rs4674406	220360683	A	G	0.35	0.76 (0.63,0.89)	3.67e-05
10	rs1331328	18532672	A	C	0.17	0.70 (0.53,0.87)	4.15e-05
7	rs6973414	153524142	G	A	0.11	1.38 (1.22,1.53)	4.37e-05
3	rs17011371	22473729	G	A	0.05	1.56 (1.35,2.78)	4.38e-05
2	rs1050816	220358198	A	G	0.33	0.76 (0.63,0.89)	4.70e-05
13	rs13378379	88710893	C	A	0.34	1.25 (1.14,1.35)	5.24e-05
4	rs12645519	15758931	A	G	0.10	1.39 (1.23,1.54)	5.30e-05

Table 3.3.3.4. Top 20 SNPs from GWAS of Major Depressive Disorder in males only, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), odds ratio with 95% confidence intervals (OR (CI)), and association p-value (P).

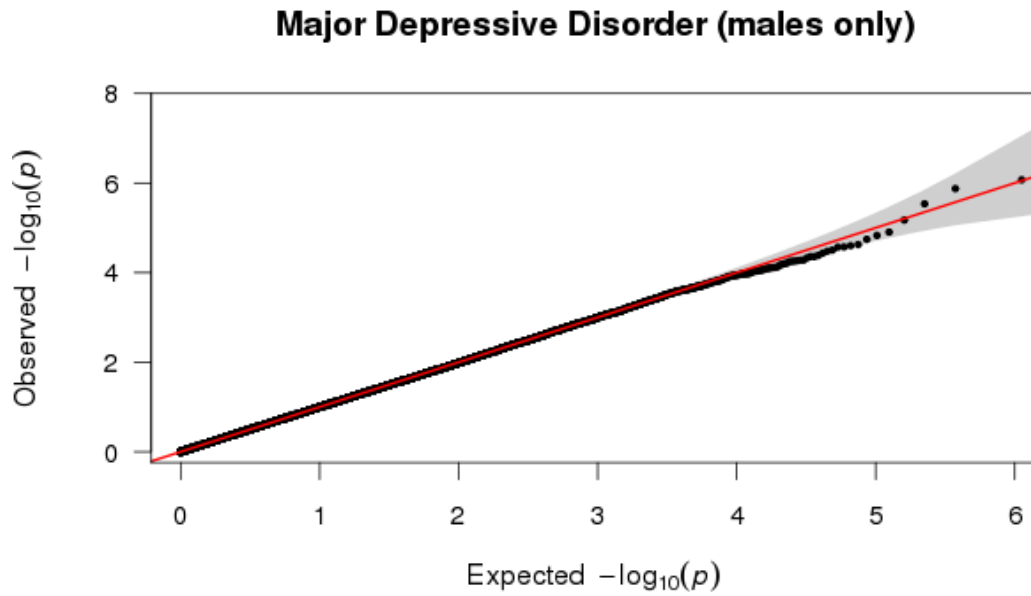


Figure 3.3.3.4.A. Quantile-quantile plot for Major Depressive Disorder in males only, showing expected distribution of GWAS test statistics, $-\log_{10}(p)$, versus the observed distribution. $\lambda(\text{SE}) = 0.995$ ($7.72\text{e-}06$).

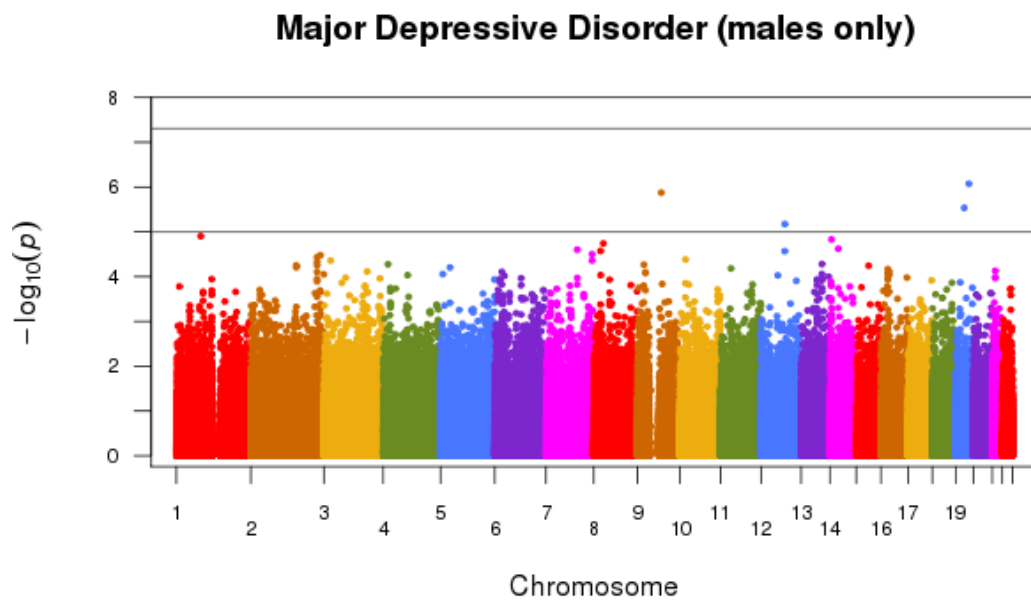


Figure 3.3.3.4.B. Manhattan plot of genomic coordinate against test statistic for association from GWAS for Major Depressive Disorder in males only, showing no genome-wide significant hits.

3.3.3.5. Biological annotation

The top 20 SNPs for each MDD definition were subjected to further bioinformatic analysis. Proxy SNPs were identified using a LD r^2 cut-off of 0.5, over a maximum distance of +/- 500kb from the query SNP using SNAP (SNP Annotation and Proxy Search v2.2.)²⁰⁰, available via the Broad Institute. Pairwise LD estimates were derived from the CEU subpopulation of the 1000 Genomes Project¹⁶¹. Distances are based on the HapMap annotations for each SNP. Where a SNP or proxy SNP ($LD\ r^2 \geq 0.5$) overlapped with a known gene, the gene(s) function was further explored using UCSC genome browser²⁰¹, the GeneCards database and PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>). SNPs or proxy SNPs within known genes of biological interest are shown for MDD, rMDD, fMDD, and mMDD in Tables 3.3.3.5.1.-3.3.3.5.4. Where genes are tagged by proxy SNPs the query SNP and LD r^2 are included. If a query SNP is in a gene, proxy SNPs are only additionally listed if they are in a different annotation type. Regional LD plots are provided for regions of interest in this Section (i.e. a putative gene for psychiatric disorders or clinical covariates).

3.3.3.5.1. Biological annotation of top hits from MDD GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs882151	-	1.00	11:75904137	<i>WNT11</i> (intronic)	development of skeleton, kidney and lungs	-
rs1331328	-	1.00	10:18532672	<i>CACNB2</i> (intronic)	voltage-gated calcium channel	cardiovascular disease; psychiatric cross-disorder
rs1757236	-	1.00	10:18527454			
rs39700	-	1.00	5:66162044	<i>MAST4</i> (intronic)	transferase activity; transferring phosphorus-containing groups; protein tyrosine kinase activity	breast cancer
rs10771350	-	1.00	12:27760130	<i>PPFIBP1</i> (intronic)	axon guidance and mammary gland development	tumour invasiveness and metastasis; inflammatory myofibroblastic tumor
rs6741418	rs13010343	0.71	2:191551690	<i>STAT1</i> (intronic)	transcription activator for cell viability in response to stimuli/pathogens	autosomal dominant immunodeficiency 31a, 31c

Table 3.3.3.5.1. Biological annotation of the top 20 SNPs from GWAS of MDD showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs6741418	rs13035504	0.53	2:191723719	<i>GLS</i> (3'UTR)	catalyzes hydrolysis of glutamine to glutamate and ammonia, synthesizes the neurotransmitter glutamate	hepatic encephalopathy and acute endometritis
rs4868136	-	1.00	5:171472575	<i>STK10</i> (intronic)	lymphocyte-oriented kinase	testicular germ cell tumor, seminal vesicle tumor, cholecystolithiasis, male reproductive organ benign neoplasm, anus disease
rs9942465	-	1.00	6:35432191 ^a	<i>PPARD</i> (intronic) ^a (intronic/splice site) ^b	transcription repression, nuclear receptor signaling	diabetic cataract, arteriosclerosis, obesity, metabolic syndrome x, colorectal cancer
rs9942465	rs9658135	1.00	6:25486975 ^b			
rs9942465	rs16874216	1.00	6:35371219 ^a	<i>ZNF76</i> (3'UTR) ^a (intronic) ^b	nucleic acid binding, may be involved in transcriptional regulation	-
rs9942465	rs9658149	1.00	6:35495516 ^b			

Table 3.3.3.5.1. Continued. Biological annotation of the top 20 SNPs from GWAS of MDD.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs9942465	rs2004	1.00	6:35164525 ^a	<i>ANKS1A</i> (3'UTR) ^a (intronic) ^b	epidermal growth factor receptor recycling and stability	-
rs9942465	rs12110520	1.00	6:35156483 ^b			
rs7593100	rs7775	1.00	2:183407829	<i>FRZB</i> (non-syn)	regulation of bone development	female-specific osteoarthritis
rs13172324	rs10512919	0.65	5:7455004	<i>ADCY2</i> (intronic)	brain expressed - catalyzes the formation of cyclic adenosine monophosphate	bipolar disorder
rs6093898	-	1.00	20:41999991	<i>TOX2</i> (intronic)	transcriptional activator involved in the hypothalamo-pituitary-gonadal system.	-
rs17711148	-	1.00	7:14332000	<i>DGKB</i> (intronic)	calcium ion binding, diacylglycerol kinase activity	type 2 diabetes

Table 3.3.3.5.1. Continued. Biological annotation of the top 20 SNPs from GWAS of MDD. non-syn stands for non-synonymous coding variant.

Seven of the 20 highest ranking SNPs from GWAS of MDD were in intergenic regions (see Appendix 3.B.). Of the 13 remaining SNPs, 7 tagged biologically interesting genes and will be discussed here.

The top SNP for MDD, rs882151, is located in an intron of the *WNT11* gene. Whilst *WNT11* has not been previously reported in psychiatric disorders, it is inhibited by the Interleukin 1 beta (*IL1β*) gene which encodes for the IL-1 β cytokine protein²⁰². This protein has been previously implicated in psychiatric disorders including association with MDD symptoms in maltreated children of pre-school age²⁰³, association with case-control status in a small schizophrenia case-control study²⁰⁴, and increased activity in the bilateral frontal region of the brain, such as the pre-frontal cortex, in schizophrenia patients²⁰⁵. The regional LD plot for rs882151 is shown in Figure 3.3.3.5.1.A.

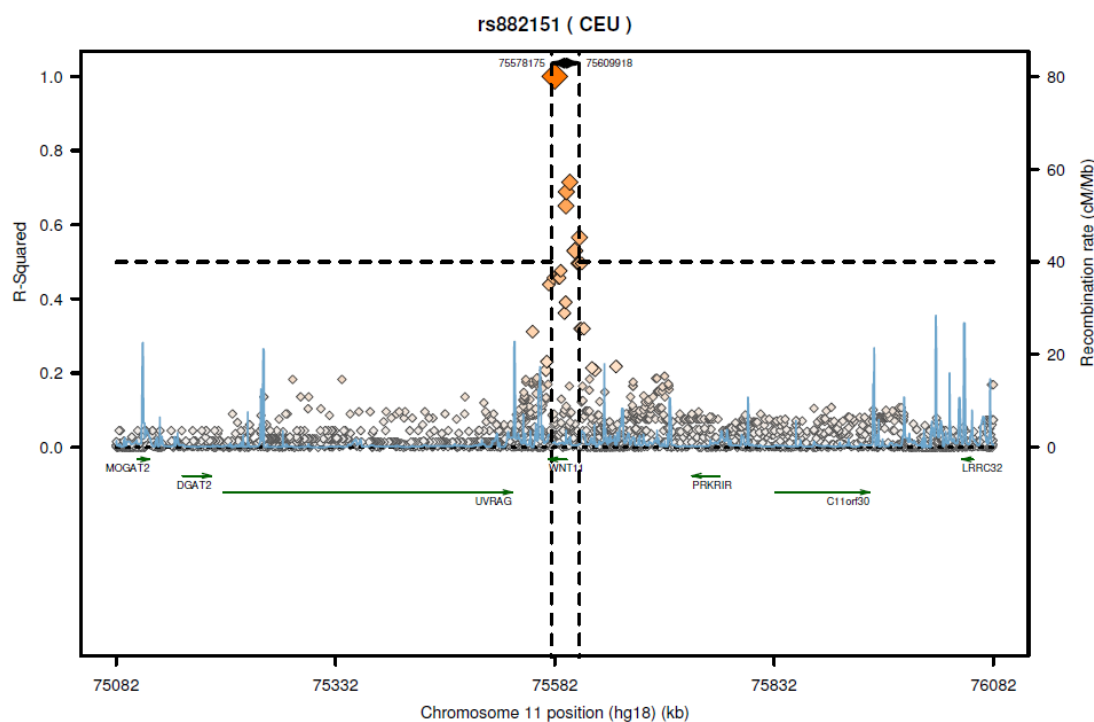


Figure 3.3.3.5.1.A. Regional LD plot for rs882151, an intronic SNP in *WNT11* - ranked 1/20 in MDD GWAS.

Two SNPs in high LD ($r^2=0.93$), rs1331328 and rs1757236 (ranked 2nd and 12th in the MDD GWAS), are intronic variants for *CACNB2*. This gene encodes for the voltage-dependent L-type calcium channel subunit beta-2 protein, and

has been cited numerous times in the literature for its potential involvement in the aetiology of psychiatric disorders. SNPs within *CACNB2* achieved genome-wide significance in the PGC cross-disorder association meta-analysis of autism, ADHD, bipolar disorder, MDD and schizophrenia. Subsequent pathway analysis of the five disorders showed enrichment for *CACNB2*, which remained significant after the exclusion of subjects with bipolar disorder, suggesting a role for L-type calcium channel subunits in psychiatric disorders in general²⁰⁶. *CACNB2* has also been associated with bipolar disorder in GWAS in Han Chinese²⁰⁷ and Taiwanese²⁰⁸ populations. Hypermethylation of *CACNB2* has been observed in haplotype carriers compared to married-in controls in a Scottish family multiply affected by bipolar disorder and MDD, who carry an illness-linked haplotype on chromosome 4p²⁰⁹. *CACNB2* has also been implicated in bipolar disorder via target gene analysis, which predicts the biological targets of microRNAs, identifying *CACNB2* as a target of miR-499 - a microRNA which has shown association with bipolar disorder²¹⁰. MicroRNAs are a class of small noncoding RNAs which contribute to basic mechanisms underlying brain development and synaptic plasticity. Studies of postmortem brain tissue of BD patients have demonstrated altered miRNA expression profiles in the prefrontal cortex²¹¹. Outwith psychiatric phenotypes, *CACNB2* has also shown association with three major cardiovascular diseases - hypertension, heart failure, and arrhythmia²¹². This apparent pleiotropic effect of *CACNB2* on psychiatric disorders and cardiovascular disease is particularly interesting given that the risk of mortality within two years of assessment in patients who are comorbid for coronary heart disease and clinical depression is twice that of patients without depressive symptoms (OR: 2.61; CI: 1.53-4.47)¹¹⁸. A regional LD plot for this region is provided in Figure 3.3.3.5.1.B. *CACNB2* tagged by rs1331328 was also amongst the top associations (ranked 15/20) for MDD in males only.

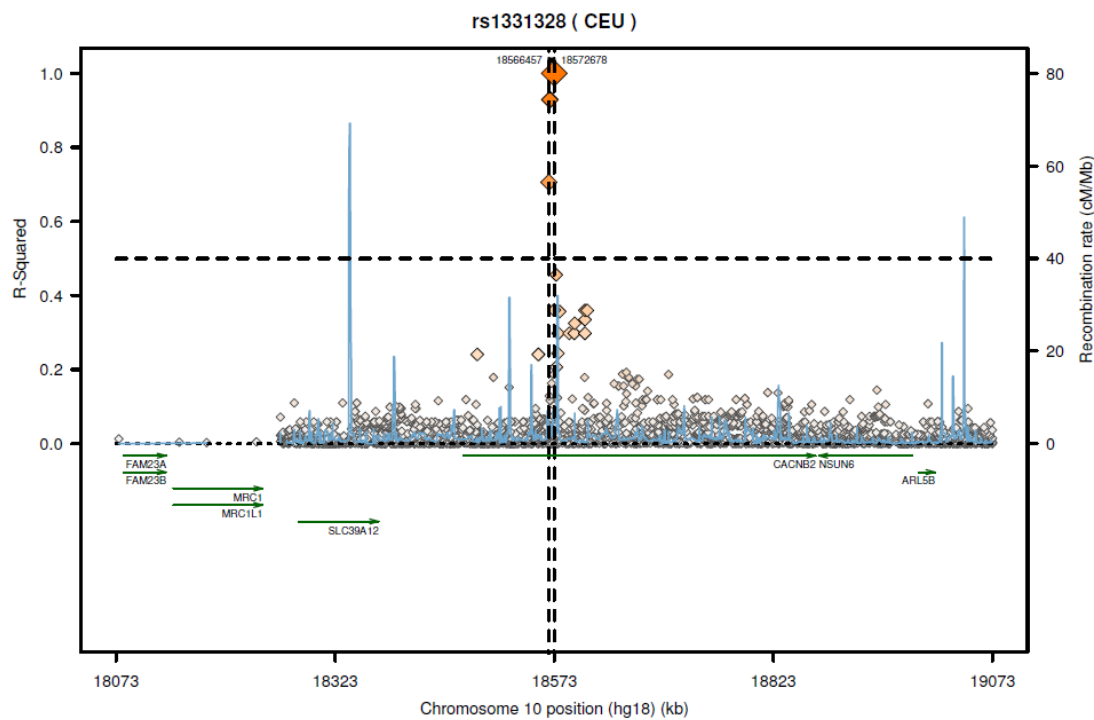


Figure 3.3.3.5.1.B. Regional LD plot for rs1331328 - ranked 2/20 for MDD and 15/20 in MDD in males only GWAS. This SNP is in high LD ($r^2=0.93$) with rs1757236, ranked 12/20 in MDD GWAS. Both SNPs are intronic variants in *CACNB2*.

The next highest ranked SNP of interest, rs9942465, is amongst the top ranked SNPs not only in the GWAS of MDD (13/20), but also in recurrent MDD (2/20) and MDD in females only (4/20). This SNP is located within an intron of the *PPARD* gene, which codes for the delta member of the peroxisome proliferator-activated receptor (PPAR) nuclear hormone receptor subfamily. PPAR δ , in combination with the other PPAR subfamily members, PPAR α and PPAR γ , play an essential role in the regulation of insulin sensitivity, adipogenesis, lipid metabolism and blood pressure. As such, *PPARD* is implicated in Type 2 Diabetes and Metabolic Syndrome²¹³, with a recent GWAS of metabolic traits in a large population of Chinese adults identifying an association with *PPARD*²¹⁴. In terms of psychiatric disorders, a role for *PPARD* has been suggested in bipolar disorder²¹⁵ via the Wnt signalling pathway²¹⁶. Medications used to treat bipolar disorder, such as lithium and valproic acid, are known to affect Wnt signaling pathways, in part, via glycogen synthase kinase 3 β (GSK3 β)^{215,217}. A family-

based study of 554 bipolar offspring and their parents sought to test genes from the Wnt signalling pathway for their association with bipolar disorder. Of the genes tested, *PPARD* was the most significantly associated with bipolar case control status (minimum SNP p-value = 0.0001). *PPARD* also demonstrated a significant association with scores on the Global Assessment Scale, a 100 point questionnaire which indicates the level of functional impairment an individual is experiencing^{217,218}. As shown in Figure 3.3.3.5.1.C, rs9942465 is in perfect LD with several SNPs spanning a region of approximately 300kb and tagging *ZNF76* and *ANKS1A*, the latter of which has some evidence for a potential role in alcohol dependency²¹⁹. It is worth noting that rs9942465 was not available in the 1000 Genomes reference panel, and this plot was therefore generated using the HapMap3 reference panel.

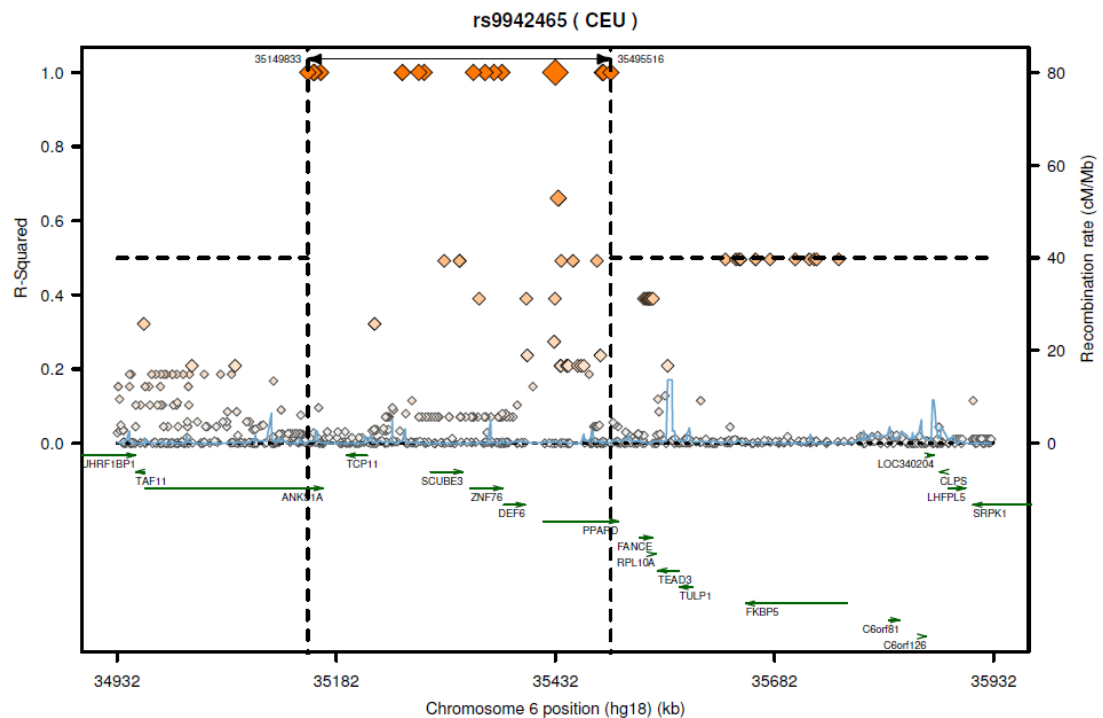


Figure 3.3.3.5.1.C. Regional LD plot for rs4246599, an intronic variant in *PPARD* - ranked 13/20 for MDD, 2/20 for recurrent MDD and 4/20 in MDD in females only GWAS. This SNP is in perfect LD with SNPs in intronic and 3'UTR regions *ZNF76* and *ANKS1A*.

An intronic SNP, rs10512919, in *ADCY2* is in moderately high LD ($r^2=0.65$) with rs13172324, a SNP which ranked 16/20 in the MDD GWAS, and 11/20 in the rMDD GWAS. *ADCY2* has previously exceeded genome-wide significance for association with bipolar disorder in a large GWAS (9,747 patients and 14,278 controls) of European descent²²⁰. A regional LD plot for rs13172324 is shown in Figure 3.3.3.5.1.D.

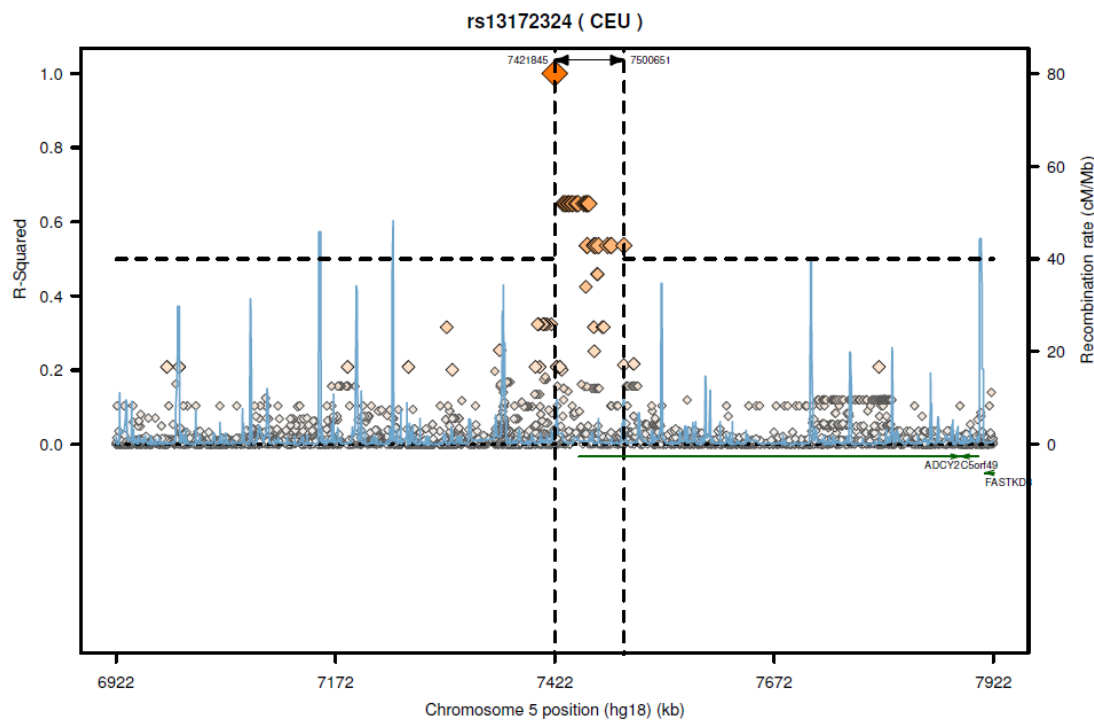


Figure 3.3.3.5.1.D. Regional LD plot for rs13172324, which is in moderately high LD ($r^2=0.65$) with rs10512919, an intronic variant in *ADCY2* - ranked 16/20 in MDD and 11/20 in recurrent MDD GWAS.

The last SNP to be discussed in this Section is the 17th highest ranking SNP, rs6093898, an intronic variant in the *TOX2* gene, a putative transcriptional activator involved in the hypothalamo-pituitary-gonadal system²²¹. This SNP tags an additional 33 intronic SNPs within *TOX2*, including rs6031245 which is ranked 18th. A haplotype block spanning a 24kb region within *TOX2* has demonstrated association achieving genome-wide significance with MDD in GS:SFHS using haplotype-block-based regional heritability mapping²²², a technique which estimates the localized variance explained by common variants within haplotype blocks. This finding was replicated in the UK-Ireland group of the most recent release of the

Psychiatric Genomics consortium (PGC2-MDD). The expression of *TOX2* in the frontal cortex is significantly regulated by MDD-associated SNPs within the same haplotype block, as is the expression of a brain-specific LncRNA, *RP1-269M15.3* in the nucleus accumbens²²³. The Nucleus Accumbens is central to mediating motivation for obtaining rewards and demonstrates increased neuronal activity and dopamine release in response to the expectation and experience of reward. Anhedonia experienced by depressed patients is thought to be under the influence of the dysfunctional activity of the Nucleus Accumbens²²⁴. The regional LD plot for rs6093898 is shown in Figure 3.3.3.5.1.E.

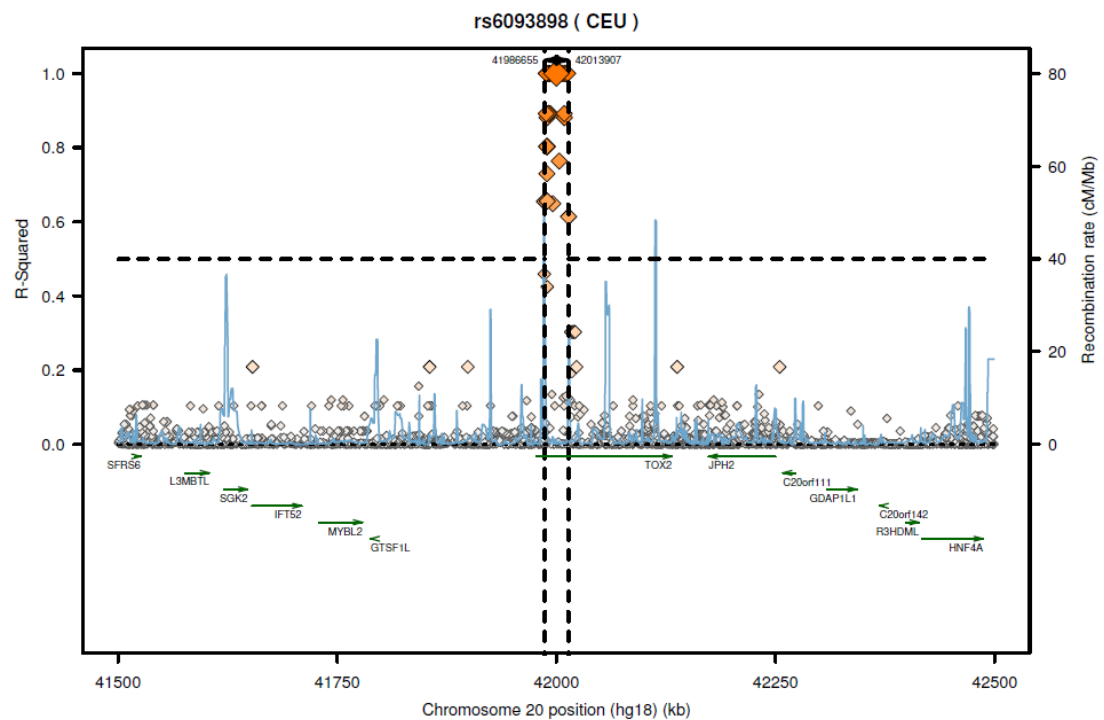


Figure 3.3.3.5.1.E. Regional LD plot for rs6093898, an intronic SNP in the *TOX2* gene - ranked 17/20 in MDD GWAS. This SNP is in high LD ($r^2=0.89$) with rs6031245 - ranked 18/20 in MDD GWAS.

3.3.3.5.2. Biological annotation of top hits from recurrent MDD GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs9942465	-	1.00	6:35432191	<i>PPARD</i> (intronic)	transcription repression, nuclear receptor signaling	diabetic cataract, arteriosclerosis, obesity, metabolic syndrome x, colorectal cancer
rs9942465	rs16874216	1.00	6:35371219 ^a	<i>ZNF76</i> (3'UTR) ^a	nucleic acid binding, may be involved in transcriptional regulation	-
rs9942465	rs9658149	1.00	6:35495516 ^b	(intronic) ^b		
rs9942465	rs2004	1.00	6:35164525 ^a	<i>ANKS1A</i> (3'UTR) ^a	epidermal growth factor receptor recycling and stability	-
rs9942465	rs12110520	1.00	6:35156483 ^b	(intronic) ^b		
rs1331328	-	1.00	10:18532672	<i>CACNB2</i> (intronic)	voltage-gated calcium channel	cardiovascular disease; psychiatric cross-disorder
rs1757236	-	1.00	10:18527454			
rs10946731	-	1.00	6:24952422 ^a	<i>FAM65B</i> (intronic) ^a	myotube formation; T-cell migration.	deafness (autosomal); reading disorder
rs6901565	-	1.00	6:24933490 ^b	(syn) ^b		

Table 3.3.3.5.2. Biological annotation of the top 20 SNPs from GWAS of recurrent MDD showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene, (syn) stands for synonymous coding variant.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs13172324	rs10512	0.65	5:7455004	<i>ADCY2</i> (intronic)	brain expressed - catalyzes the formation of cyclic adenosine monophosphate	bipolar disorder
rs10735017	-	1.00	12:3099234	<i>TSPAN9</i> (intronic)	signal transduction; cell development and motility	gastric cancer
rs4499887	rs11759494	0.74	6:79997836	<i>HMGN3</i> (intronic)	chromatin binding; thyroid receptor binding	-
rs2322217						
rs4047801	rs2702178	0.72	1:18171916 ^a	<i>SMG7</i> (intronic) ^a (3'UTR) ^b	nonsense-mediated mRNA decay; p53 signalling pathway	cancer
rs4047801	rs1044879	0.65	1:181789375 ^b			
rs4047801	rs12753665	0.51	1:181812684	<i>NCF2</i> (intronic)	protein C-terminus binding; Rac GTPase binding	chronic granulomatous disease; Alzheimer's disease

Table 3.3.3.5.2. Continued. Biological annotation of the top 20 SNPs from GWAS of rMDD.

Fourteen of the 20 highest ranking SNPs from the GWAS of rMDD were in intergenic regions (see Appendix 3.C.). Of the 6 query SNPs that were in or near genes, 2 have already been discussed in Section 3.3.3.5.1., due to also being highly ranked in the GWAS of undifferentiated MDD (rs9942465, intronic SNP in *PPARD*, tagging *ZNF76* and *ANKS1A*; rs13172324, intronic SNP in *ADCY2*). Therefore the 4 remaining SNPs in biologically interesting regions shall be discussed here.

The fourth and fifth highest ranking SNPs for rMDD were an intronic (rs10946731) and synonymous coding variant (rs6901565) in the *FAM65B* (Family With Sequence Similarity 65 Member B) gene. *FAM65B* is located on chromosome 6p22, and is part of the *DYX2* locus (Figure 3.3.3.5.2.A.) which is implicated in reading disability, language impairment and cognition²²⁵.



Figure 3.3.3.5.2.A. Schematic of genes in the *DYX2* locus on chromosome 6p22²²⁵, which incorporates *FAM65B*.

In a study seeking to characterize the relationship between *DYX2* and reading-, language- and IQ-related traits in the ALSPAC cohort, *FAM65B* showed nominal association with verbal IQ but not general cognition, with this finding being partially replicated²²⁵. *FAM65B* also yielded a suggestive p-value (6.45e-05) for performance on the Peabody Picture Vocabulary test (a measure of verbal intelligence) in children with autism (n=1,989). Lastly, a genetic neuroimaging study of genes within the *DYX2* locus revealed an association between *FAM65B* and global fractional anisotropy, suggesting that risk alleles negatively impact the white matter integrity of fiber tracts²²⁶. The regional LD plot for rs10946731 is shown in Figure 3.3.3.5.2.B.

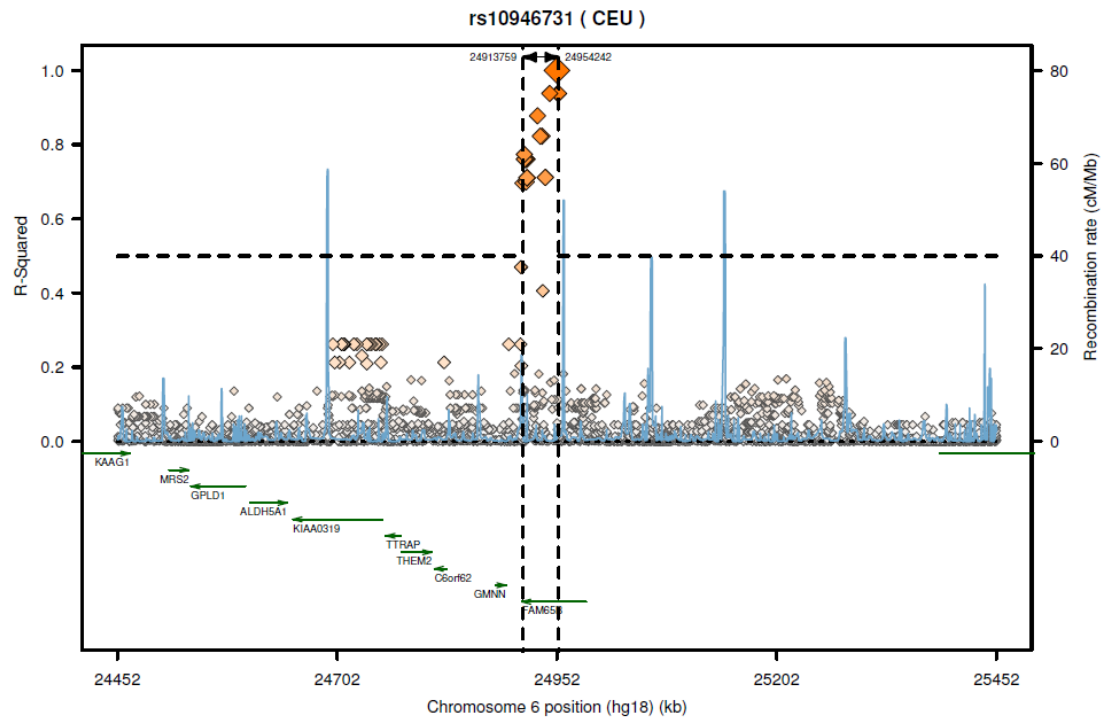


Figure 3.3.3.5.2.B. Regional LD plot for rs10946731, an intronic variant in *FAM65B* - ranked 4/20 in rMDD GWAS, in high LD ($r^2=0.94$) with rs6901565, a synonymous coding variant in *FAM65B* - ranked 5/20 in rMDD GWAS.

The 13th highest SNP in the rMDD GWAS, rs10735017, is an intronic SNP in the *TSPAN9* gene. *TSPAN9* is a member of the transmembrane 4 superfamily of tetraspanins, and has been shown to inhibit the proliferation, migration and invasion of human gastric cancer cells²²⁷. This is an interesting association, as mortality rates in cancer patients with comorbid major or minor depression is up to 39% higher than individuals without depression¹²⁰. However, it is worth noting that this study was looking at cancer in general, and the prevalence of depression in gastric cancer (~11%) is lower than many other cancer types (range: 0 - 58%)²²⁸, therefore gastric cancer may not be driving the mortality finding. The 6th highest ranking SNP for rMDD, rs1005374, is intergenic (see Appendix 3.C.) but ~50kb downstream of *TSPAN9*. However, is only in very weak LD ($r^2=0.18$) with an intronic variant in this gene. A regional LD plot for rs10735017 is shown in Figure 3.3.3.5.2.C.

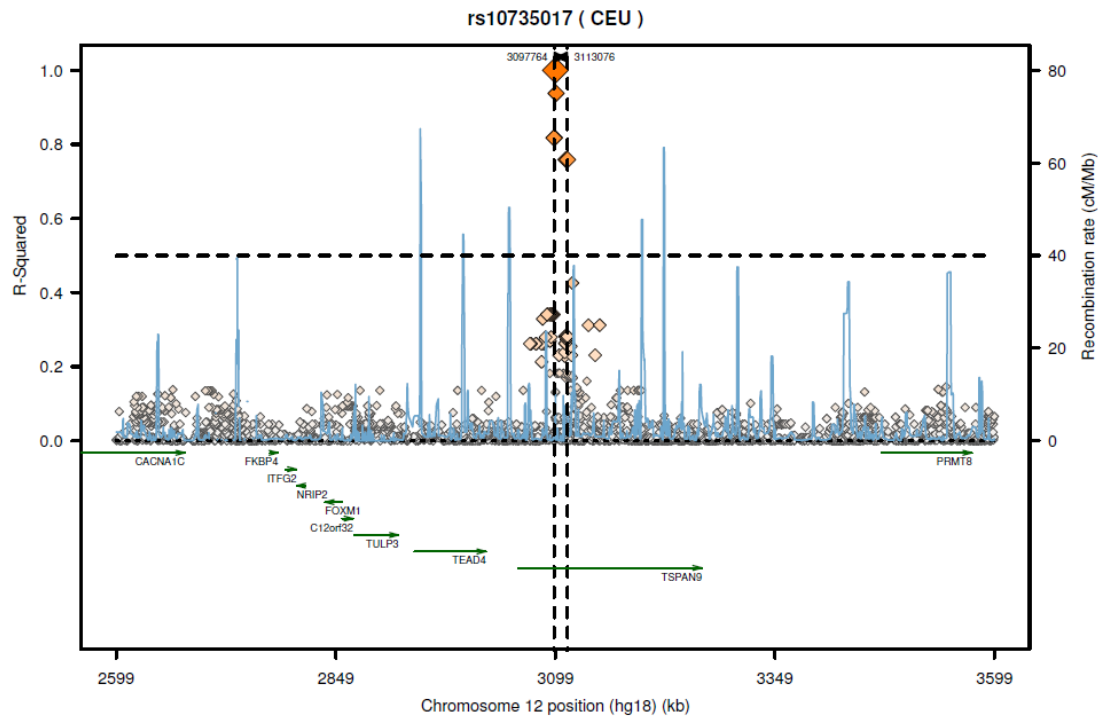


Figure 3.3.3.5.2.C. Regional LD plot for rs10735017, an intronic variant in *TSPAN9* - ranked 13/20 in rMDD GWAS.

The 15th highest ranking SNP for rMDD, rs4047801, is in moderately high LD ($r^2=0.65-0.72$) with 8 SNPs (seven intronic, one within the 3'UTR region) of the *SMG7* (suppressor with morphological defects in genitalia 7) gene. This gene has been implicated in cancer via its *p53*, a putative tumor suppressor gene that is inactivated by somatic mutations in the majority of human cancers. *SMG7* is a *p53*-binding protein which is thought to regulate *p53* stability by degrading nonsense mutations in *p53* mRNA^{229,230}. Again, this is of biological interest due to the increased mortality of cancer patients with depression¹²⁰. This query SNP (rs4047801) is also in moderate LD ($r^2=0.51$) with rs12753665, an intronic SNP in the *NCF2* gene. This gene has also been shown to play a role in the *p53*-signalling pathway²³¹. *NCF2* codes for the p67phox protein, a subunit of the multi-protein complex NADPH oxidase. This complex is found in neutrophils and has been shown to be activated by amyloid β , resulting in the formation of neurotoxic reactive oxygen species, in the brains of diseased Alzheimer's patients²³². This is of interest given the frequent comorbidity observed between depression and dementias, of which Alzheimer's disease is the most common¹²².

Individuals experiencing both dementia and depression exhibit accelerated cognitive decline, increased requirement for institutional care, increased disability in activities of daily living, a poorer quality of life and higher mortality than their non-depressed counterparts²³³. The regional LD plot for rs4047801 is shown in Figure 3.3.3.5.2.D.

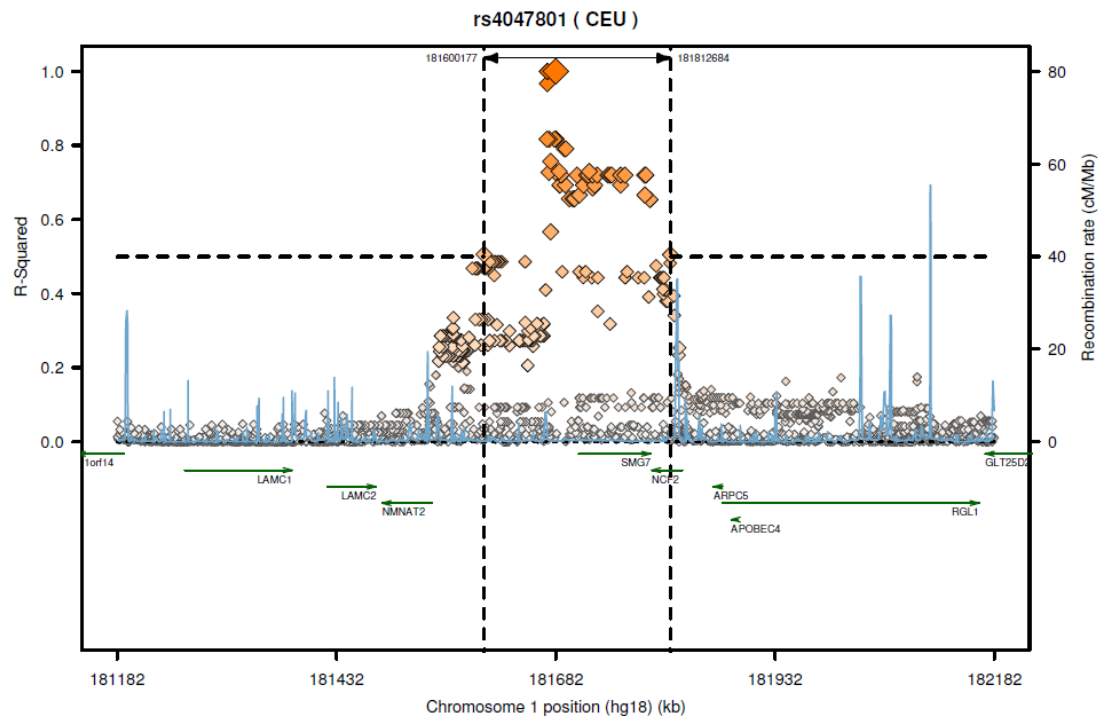


Figure 3.3.3.5.2.D. Regional LD plot for rs4047801 - ranked 15/20 in rMDD GWAS. This SNP is in moderately high LD ($r^2=0.65-0.72$) with SNPs in *SMG7* and in moderate LD ($r^2=0.51$) with an intronic SNP in *NCF2*.

3.3.3.5.3. Biological annotation of top hits from MDD in females only GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs9942465	-	1.00	6:35432191	<i>PPARD</i> (intronic)	transcription repression, nuclear receptor signaling	diabetic cataract, arteriosclerosis, obesity, metabolic syndrome x, colorectal cancer
rs9942465	rs16874216	1.00	6:35371219 ^a	<i>ZNF76</i> (3'UTR) ^a (intronic) ^b	nucleic acid binding, may be involved in transcriptional regulation	-
rs9942465	rs9658149	1.00	6:35495516 ^b			
rs9942465	rs2004	1.00	6:35164525 ^a	<i>ANKS1A</i> (3'UTR) ^a (intronic) ^b	epidermal growth factor receptor recycling and stability	-
rs9942465	rs12110520	1.00	6:35156483 ^b			
rs4868136	-	1.00	5:171472575	<i>STK10</i> (intronic)	lymphocyte-oriented kinase	testicular germ cell tumor, seminal vesicle tumor, cholecystolithiasis, male reproductive organ benign neoplasm, anus disease
rs17764251	-	1.00	17:4160754	<i>UBE2G1</i> (intronic)	ligase activity; ubiquitin protein ligase binding	Parkinson's disease

Table 3.3.3.5.3. Biological annotation of the top 20 SNPs from GWAS of MDD in females only showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs17764251	rs9900637	0.53	17:3898724	<i>ZZEF1</i> (intronic)	calcium ion binding	-
rs39700	-	1.00	5: 66162044	<i>MAST4</i> (intronic)	transferase activity; transferring phosphorus-containing groups; protein tyrosine kinase activity	breast cancer
rs7309194	-	1.00	12:37484213	<i>CPNE8</i> (intronic)	Calcium-dependent membrane- binding proteins	prion diseases
rs1630030	-	1.00	11: 124806797 ^a	<i>PKNOX2</i> (3'UTR) ^{a,b} (intronic) ^c	sequence-specific DNA binding; actin monomer binding	substance abuse in European women
rs1630030	rs645172	1.00	11:124807932 ^b			
rs1630030	rs720658	0.93	11:1247994 ^c			
rs1630030	rs550276	0.93	11: 124951280	<i>EI24</i> (intronic)	p53 signaling pathway	frequently altered in cancers
rs1630030	rs601616	0.86	11:124810795	<i>FEZ1</i> (intronic)	protein N-terminus binding; gamma-tubulin binding	schizophrenia
rs10771350	-	1.00	12:27760130	<i>PPFIBP1</i> (intronic)	axon guidance and mammary gland development	tumour invasiveness and metastasis; inflammatory myofibroblastic tumor
rs1133400	-	1.00	13:53825161	<i>INPP5A</i> (non-syn)	inositol phosphate metabolism	ataxia

Table 3.3.3.5.3. continued. Biological annotation of the top 20 SNPs from GWAS of fMDD.

Eleven of the 20 highest ranking SNPs from GWAS of MDD were in intergenic regions (see Appendix 3.D.). Of the 9 query SNPs that were in or near genes, 4 SNPs have already been discussed in Section 3.3.3.5.1., due to also being highly ranked in the GWAS of undifferentiated MDD (rs9942465, intronic SNP in *PPARD*, tagging *ZNF76* and *ANKS1A*; rs4868136, intronic SNP in *STK10*; rs39700, intronic SNP in *MAST4*; rs10771350, intronic SNP in *PPFIBP1*). Therefore the 5 remaining SNPs in biologically interesting regions shall be discussed here.

The 8th highest ranking SNP, rs17764251, is an intronic variant in *UBE2G1* (Ubiquitin Conjugating Enzyme E2 G1). This gene is directly involved in the functioning of the ubiquitin-proteasome protein degradation system (UPPDS)²³⁴, one of the major mechanisms of pathogenesis in Parkinson's Disease²³⁵. Reduced expression of *UBE2G1* has been observed in the striatum and substantia nigra in murine pre-symptomatic MPTP-induced Parkinson's Disease²³⁴. Parkinson's Disease is characterized by the loss of dopaminergic neurons in the substantia nigra and nerve terminals in the striatum, with this loss causing dysfunction of the nigrostriatal system. The risk of being given a diagnosis of Parkinson's disease is significantly increased for patients with an affective disorder²³⁶. Depression is more prevalent in Parkinson's disease patients relative to the general population²³⁷, with this association not being driven by gender, age, severity of disease, age of onset, duration of Parkinson's disease, treatment with levodopa or dopamine agonists²³⁸. In addition, patients with primary depression have demonstrated reduced volume and decreased metabolism in the basal ganglia, as visualized by MRI and PET studies²³⁶. *UBE2G1* has also been associated with Type 1 Diabetes (T1D) using gene-based pathway analysis from GWAS summary statistics, with this result successfully replicating in an independent sample²³⁹. The prevalence of depression in T1D is approximately four times that of controls²⁴⁰. The regional LD plot for rs17764251 is shown in Figure 3.3.3.5.3.A.

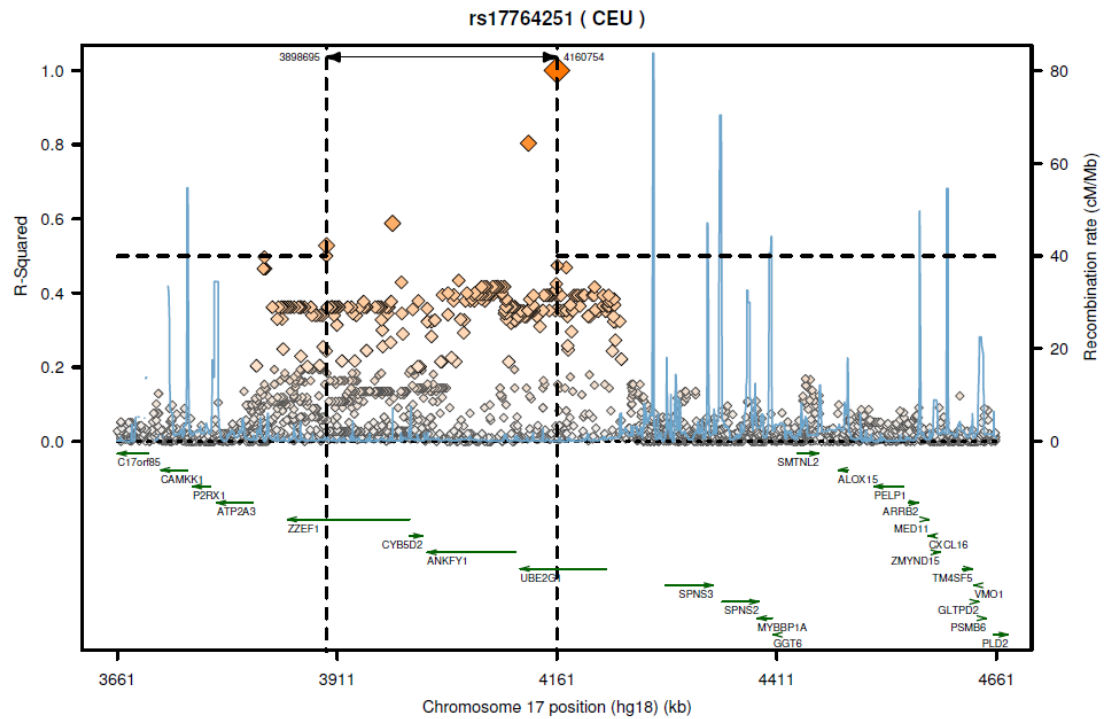


Figure 3.3.3.5.3.A. Regional LD plot for rs17764251 - ranked 8/20 in females only MDD GWAS. This SNP is an intronic variant in *UBE2G1*.

The 13th ranked SNP, rs7309194, is an intronic variant of *CPNE8*. This gene is situated in a previously identified QTL for prion disease incubation time, with evidence of *Cpne8* mRNA upregulation at the terminal stage of disease²⁴¹. Prion diseases are rare neurodegenerative diseases which can occur in sporadic, inherited and acquired forms. Prominent neuropsychiatric symptoms are a known clinical symptom of some forms of prion disease and can be used to differentiate variant/new variant Creutzfeldt-Jakob Disease (CJD) where psychiatric symptoms are a consistent early clinical feature, from sporadic CJD where psychiatric symptoms are absent^{242,243}. Dysregulation of *CPNE8* caused by the *LRRK2* mutation G2019S, the most common genetic cause of Parkinson's Disease, has also been reported. Furthermore, knocking down *CPNE8* resulted in dopaminergic neurodegeneration²⁴⁴. The regional LD plot for rs7309194 is shown in Figure 3.3.3.5.3.B.

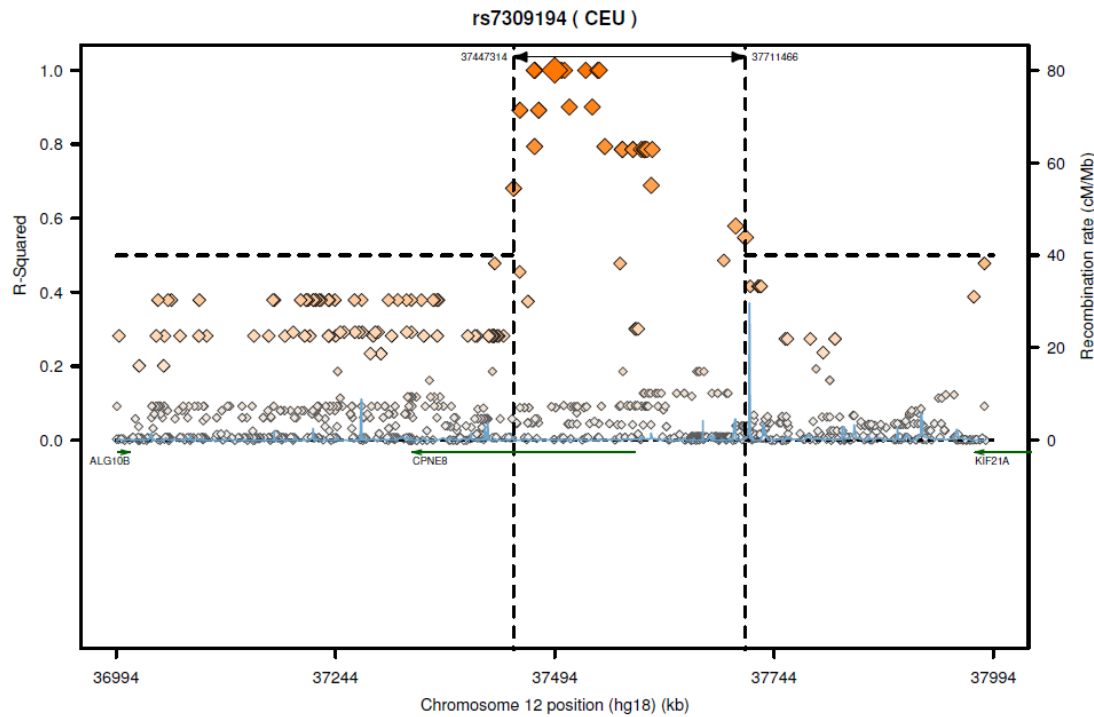


Figure 3.3.3.5.3.B. Regional LD plot for rs7309194 - ranked 13/20 in females only MDD GWAS. This SNP is an intronic variant CPNE8.

The 14th highest ranking SNP, rs1630030, and 17th highest ranking SNP, rs645172, are in perfect LD and located in the 3'UTR of *PKNOX2* (PBX/Knotted 1 Homeobox 2), and in high LD ($r^2=0.93$) with an intronic variant (rs720658) for this gene. *PKNOX2* has been associated with alcohol dependence^{245,246}, and with a composite substance abuse phenotype derived from six individual diagnoses: addiction to nicotine, alcohol, marijuana, cocaine, opiates, or other drugs as a whole in women of European descent²⁴⁷. Furthermore, *PKNOX2* has been implicated in formal thought disorder (otherwise known as disorganized speech), one of the central characteristics of schizophrenia²⁴⁸. This is particularly interesting, given that SNPs in *PKNOX2* identified in the fMDD GWAS are in moderate-high LD ($r^2=0.53-0.86$) with four intronic variants in *FEZ1* (Fasciculation And Elongation Protein Zeta 1), a gene implicated in schizophrenia via its interaction with *DISC1* (Disrupted In Schizophrenia 1)²⁴⁹. Significant reductions of *FEZ1* mRNA have been observed in the hippocampus and dorsolateral pre-frontal cortex of schizophrenia patients, with *FEZ1* mRNA levels dependent on *DISC1* genotype²⁵⁰. The query SNP is also in moderate-high

LD ($r^2=0.61-0.93$) with 5 intronic variants located in *EI24* (Etoposide-Induced Protein 2.4 Homolog), a putative p53-regulated tumor suppressor gene. Overexpression of *EI24* negatively controls cell growth and induces apoptotic cell death²⁵¹. This gene has been implicated in cancer generally, but also more specifically in breast cancer^{252,253} and the progression of cervical carcinoma²⁵⁴. This is interesting, given the differential likelihood of developing breast cancer between sexes (1 in 8 women and 1 in 870 men during their lifetime)²⁵⁵ and the sex specificity of cervical cancer, that the SNP tagging this gene is only high ranking in the GWAS of MDD in females. The homolog of *EI24* in *drosophila melanogaster*, *tank*, highlights an additional role for this gene, demonstrating that tank regulates sensitivity to ethanol. Manipulating the activity of all *tank*-expressing neurons resulted in sexually dimorphic alteration of ethanol sensitivity, with neuronal activation enhancing ethanol sensitivity in males but not females²⁵⁶. It has been suggested that ethanol sensitivity has an inverse correlation with alcohol use disorders^{257,258}, which is interesting given the high linkage disequilibrium between the mammalian homolog of *tank*, *EI24*, and *PKNOX2* - a putative sex-specific alcohol dependence associated gene. The regional LD plot for rs1630030 is shown in Figure 3.3.3.5.3.C.

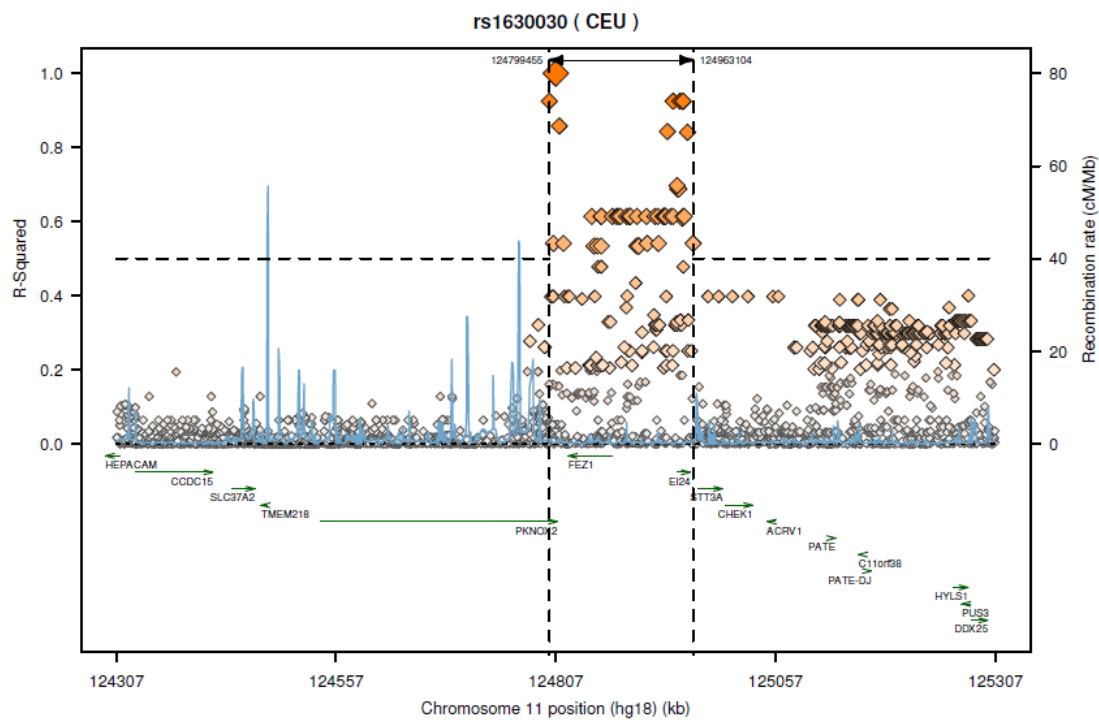


Figure 3.3.3.5.3.C. Regional LD plot for rs1630030 - ranked 14/20 in the females only MDD GWAS. This SNP is in perfect LD with rs645172 - ranked 17/20. Both SNPs are located within the 3'UTR of *PKNOX2*, and are in high LD ($r^2=0.93$) with an intronic SNP in *EI24*, and in high LD ($r^2=0.86$) with an intronic variant in *FEZ1*.

3.3.3.5.4. Biological annotation of top hits from MDD in males only GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs2307187	-	1.00	19:48771451	<i>XRCC1</i> (5'UTR)	base excision repair pathway	non-small-cell lung carcinoma, breast cancer
rs7035578	-	1.00	9:77934997	<i>PCSK5</i> (intronic)	endopeptidase activity	vacterl association; plasma and B cell neoplasms; HDL cholesterol
rs2295306	-	1.00	14:23835524 ^a	<i>DHRS1</i> (intronic) ^a (5'UTR) ^b	oxidoreductase activity	-
rs2295306	rs12892419	0.97	14:23838374 ^b			
rs2295306	rs12895335	1.00	14:23840209	<i>C14orf21/NOP9</i> (intronic)	poly(A) RNA binding	specific language impairment
rs2295306	rs1046587	0.90	14:23855900	<i>LTB4R</i> (3'UTR)	GPCR activity; leukotriene receptor activity	insulin sensitivity
rs2295306	rs14193	0.58	14:23805867	<i>RABGGTA</i> (syn)	Rab GTPase binding; regulate vesicular traffic.	choroideremia
rs2295306	rs6573722	0.56	14:23863487	<i>ADCY4</i> (intronic)	phosphorus-oxygen lyase activity; adenylate cyclase activity	-

Table 3.3.3.5.4. Biological annotation of the top 20 SNPs from GWAS of MDD in males only showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs7822917	-	1.00	8:32310423	<i>NRG1</i> (intronic)	receptor binding; growth factor activity	schizophrenia; bipolar disorder; cancer
rs2416024	-	1.00	14:46655219	<i>MDGA2</i> (intronic)	neuronal migration, axon guidance	neuroticism
rs11761864	-	1.00	7:104005811	<i>LHFPL3</i> (intronic)	tetraspan transmembrane protein	deafness; translocation associated lipoma
rs4284046	-	1.00	8:22768022	<i>PEBP</i> (intronic)	poly(A) RNA binding; enzyme binding	
rs4674406	rs2385539	0.93	2:220075197 ^a	<i>GMPPA</i> (intronic) ^a (5'UTR) ^b	nucleotidyltransferase activity	intellectual disability
rs4674406	rs719337	0.71	2:220072161 ^b			
rs4674406	rs1050816	0.93	2:220066442 ^a	<i>SPEG</i> (3'UTR) ^a (non-syn) ^b	transferase/protein tyrosine kinase activity	centronuclear myopathy; cardiomyopathy
rs4674406	rs12464085	0.93	2:220063773 ^b			
rs4674406	rs1872858	0.90	2:220088369	<i>ACCN4</i> (intronic)	sodium ion channel activity	-
rs1331328	-	1.00	10:18532672	<i>CACNB2</i> (intronic)	voltage-gated calcium channel	cardiovascular disease; psychiatric cross-disorder
rs12645519	rs12644506	0.64	4:15408976	<i>CD38</i> (intronic)	transferase activity; hydrolase activity	-

Table 3.3.3.5.4. continued. Biological annotation of the top 20 SNPs from GWAS of mMDD.

Nine of the 20 highest ranking SNPs from GWAS of MDD were in intergenic regions (see Appendix 3.E.). Of the 11 query SNPs that were in or near genes, only one SNP has been previously discussed, due to also being highly ranked in the GWAS of undifferentiated MDD (rs1331328, an intronic variant in *CACNB2*). Of the 10 remaining SNPs, 7 tagged biologically interesting genes and will be discussed here

The highest ranking SNP, rs2307187, is in the 5'UTR of *XRCC1* (X-Ray Repair Cross-Complementing Protein 1). The *XRCC1* protein encoded by this gene plays an important role in base excision repair pathway, and is responsible for the repair of oxidative DNA damage and single-strand breaks²⁵⁹. Overexpression of *XRCC1* is observed in non-small-cell lung carcinoma²⁶⁰, and breast cancer²⁶¹. Conversely, underexpression of *XRCC1* has been shown to promote tumour suppression²⁶². Some literature supports a role for *XRCC1* in schizophrenia²⁶³ and depression²⁶⁴; it is worth noting that these studies have small sample sizes (n=523, n=102 respectively). *XRCC1* interacts with the major human AP endonuclease (APE1), an essential enzyme in the base excision repair pathway, and stimulates its enzymatic activities²⁶⁵. Activation of *SIRT1* (silent mating type information regulation 2 homolog 1) promotes the binding of APE1 to *XRCC1*, whilst knocking down *SIRT1* suppresses this DNA repair activity²⁶⁶. This interaction is interesting, given that *SIRT1* is one of two genes to be identified in the CONVERGE GWAS of recurrent MDD in female Han Chinese individuals⁴⁹. It is worth noting, however, that the association between *SIRT1* and MDD did not replicate in the European population, where the minor allele for the *SIRT1* SNP, rs12415800, is much rarer (MAF=0.03). The minor allele for the *XRCC1* SNP, rs2307187, is similarly rare (MAF = 0.01) in this population. This SNP was not available on the 1000 Genomes reference panel. The regional LD plot for rs2307187, shown in Figure 3.3.3.5.4.A., therefore uses data from the HapMap3 reference panel, which has a sparser coverage, giving the appearance that rs2307187 is the only SNP within the 500kb window. Further investigation using UCSC genome browser revealed that there are other SNPs within this window, but the LD structure cannot be ascertained from that database.

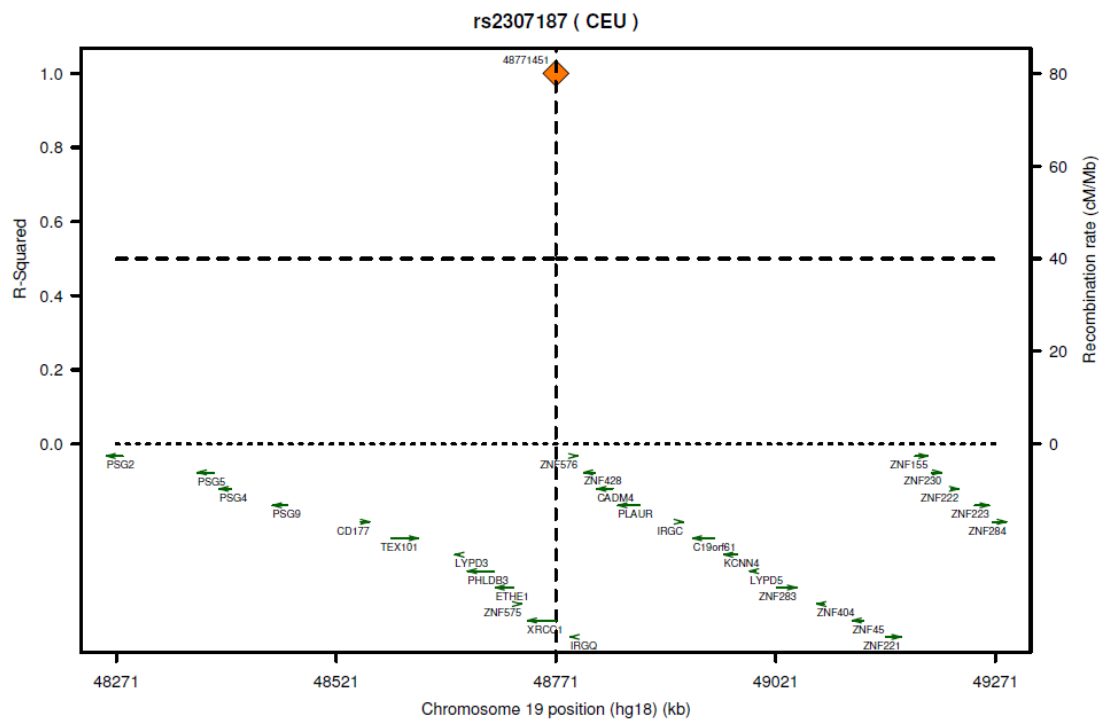


Figure 3.3.3.5.4.A. Regional LD plot, generated using the HapMap3 reference panel, for rs2307187 - ranked 1/20 in the GWAS of MDD in males only. This SNP is located in the 5'UTR of *XRCC1*.

The second highest ranking SNP, rs7035578, is an intronic variant in *PCSK5* (Proprotein Convertase Subtilisin/Kexin Type 5). This gene has been implicated in cardiovascular disease²⁶⁷, and shown sex-specific association with high-density lipoprotein cholesterol (HDL-C) in males from the Korean population²⁶⁸. Low levels of HDL-C are a putative major risk factor for coronary artery disease. As stated in Section 3.3.3.5.1., MDD and elevated depressive symptoms are associated with worse prognosis in patients with cardiovascular disease. The regional LD plot for rs7035578 is shown in Figure 3.3.3.5.4.B.

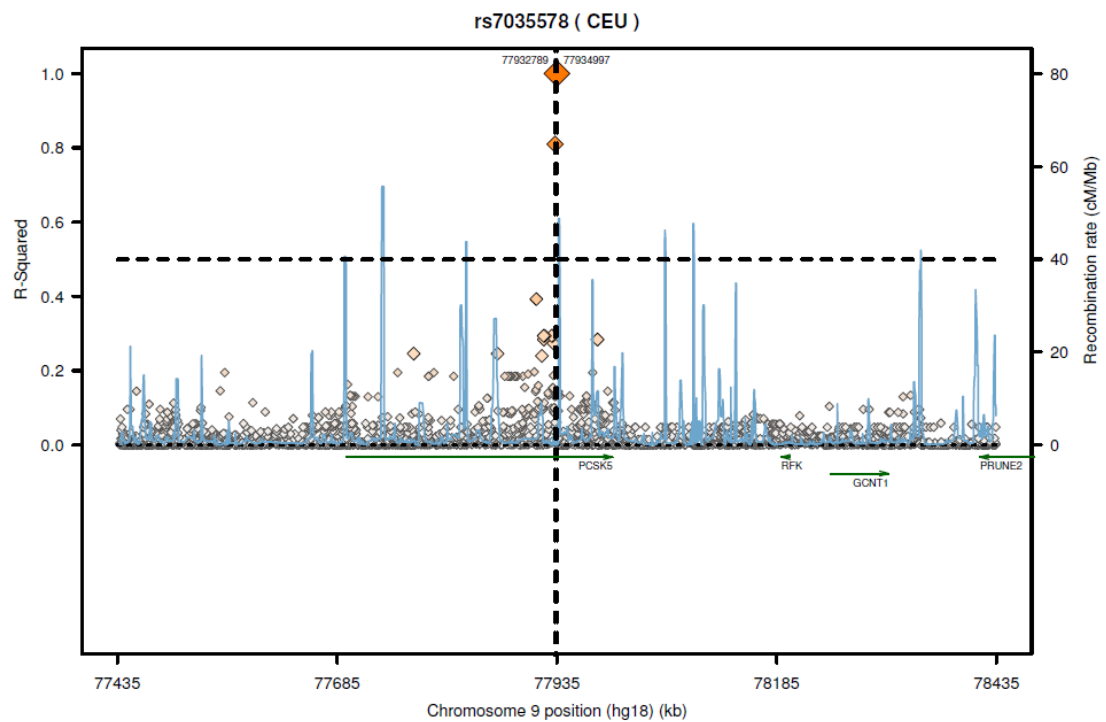


Figure 3.3.3.5.4.B. Regional LD plot for rs7035578, ranked 2/20 in the GWAS of MDD in males only. This SNP is an intronic variant in *PCSK5*.

The sixth highest ranking SNP, rs2295306, is in perfect LD with rs12895335, an intronic variant in *C14orf21*, otherwise known as *NOP9*. This gene has been associated under a parent-of-origin model with specific language impairment²⁶⁹, a neurodevelopmental disorder that is characterized by significant deficits in linguistic ability that cannot be attributed to neurological damage, hearing impairment, or intellectual disability²⁷⁰. The study observed that the major allele increased risk when paternally inherited²⁶⁹. When tested in an independent cohort against a range of language related phenotypes, *NOP9* associated with age at first word. However, parent-of-origin effects in this cohort demonstrated a maternal effect for the major allele across language measures²⁷¹. Therefore, although this gene appears to be associated with challenges in oral language acquisition, parent-of-origin effects, whilst present, are currently inconsistent. The query SNP is also in high LD ($r^2=0.90$) with rs1046587, a variant located in the 3'UTR region of *LTB4R*, which encodes for Ltb4r, a high affinity receptor potent inflammatory mediator LTB4. Pharmacological inhibition of Ltb4r has been shown to increase insulin

sensitivity in liver and muscle tissue, protecting mice from high fat diet induced insulin resistance²⁷². Insulin resistance is central to obesity and type 2 diabetes, both of which show phenotypic correlation with depression. The regional LD plot for rs2295306 is shown in Figure 3.3.3.5.4.C.

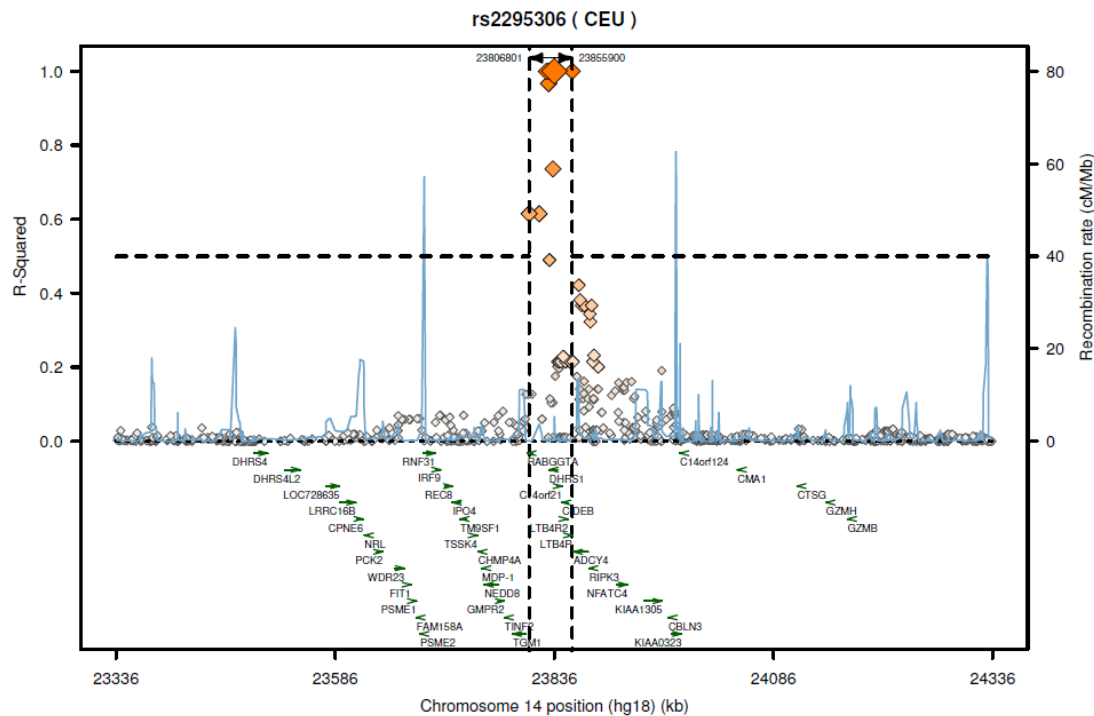


Figure 3.3.3.5.4.C. Regional LD plot for rs2295306 - ranked 6/20 in the GWAS of MDD in males only.

The seventh highest ranking SNP, rs7822917, is located within an intron of *NRG1* (Neuregulin 1). This gene codes for the NRG1 protein which acts on the EGFR family of receptors. Experiments in mice have shown that the loss or overexpression of the major brain isoform of NRG1 results in disrupted excitatory-inhibitory connections, reduced synaptic plasticity, and abnormal dendritic spine growth²⁷³. NRG1 has been repeatedly implicated in schizophrenia²⁷⁴, and to a lesser extent bipolar disorder²⁷⁵. A meta-analysis of associations between *NRG1* and schizophrenia demonstrated significant between-study heterogeneity and no evidence for an association between the most commonly reported *NRG1* marker, although evidence for a haplotype based association remained and showed no between-study heterogeneity²⁷⁶.

This is congruous with the fact that SNPs at the *NRG1* locus are not significantly associated with schizophrenia in the largest available case control study²³, despite consistent haplotype-based associations - reinforcing the point made in Section 1.3.2., that complex disorders are likely caused by rare combinations of common variants, rather than an aggregation of yet undiscovered individual common and/or rare variants¹⁰⁰. The regional LD plot for rs7822917 is shown in Figure 3.3.3.5.4.D.

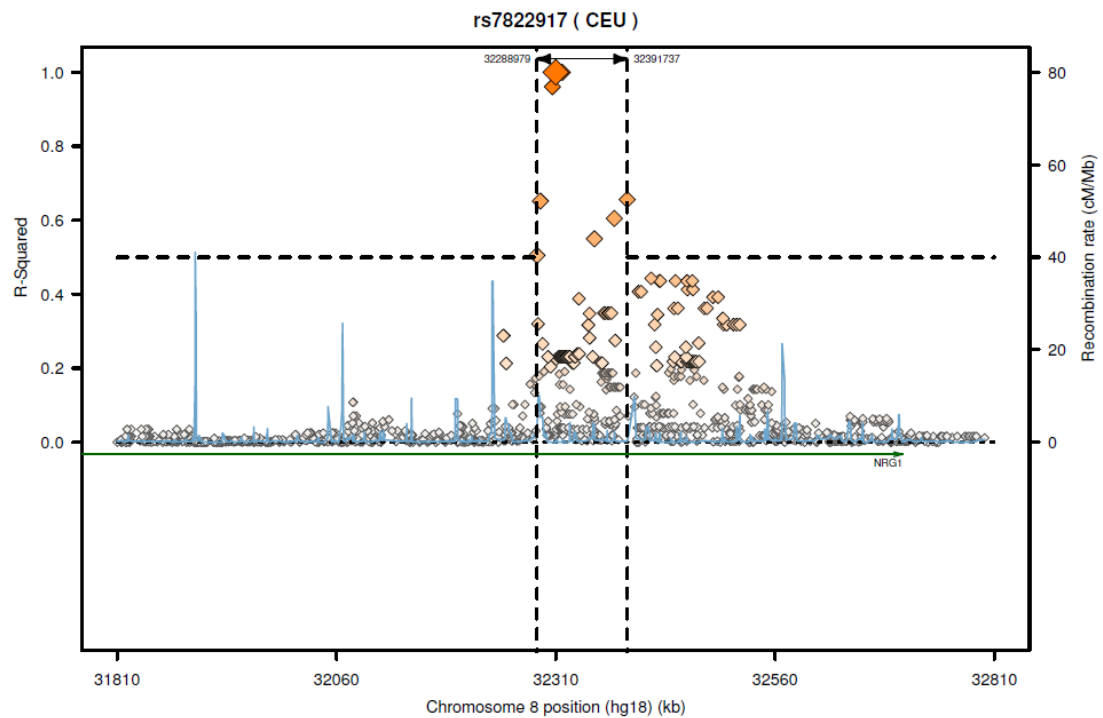


Figure 3.3.3.5.4.D. Regional LD plot for rs7822917 - ranked 7/20 in the GWAS of MDD in males only. This SNP is an intronic variant of *NRG1*.

The eighth highest ranking SNP, rs2416024, is an intronic variant of *MDGA2* (MAM domain containing glycosylphosphatidylinositol anchor 2), also known as *MAMDC1*. *MDGA2* is thought to play a role in regulating neuronal migration and other elements of neural development, such as axon guidance²⁷⁷. Four SNPs in *MDGA2*, which tag 94% of the haplotype variation, have been associated with neuroticism²⁷⁸. Replication attempts of this finding have had mixed success. The replication attempt as part of the original study was successful, as was an independent study which replicated the association between *MDGA2* and

neuroticism and further identified an association between MDGA2 and harm avoidance - a phenotype closely related to neuroticism²⁷⁹. However, the association was not replicated in a follow-up study using the Virginia Adult Twin Registry sample²⁸⁰. Neuroticism is a known correlate of depression, with data from a variety of studies suggesting that neuroticism is indicative of premorbid vulnerability in MDD^{281,282}. The relationship between neuroticism and MDD will be discussed in greater detail in Chapter 4. The regional LD plot for rs2416024 is shown in Figure 3.3.3.5.4.E.

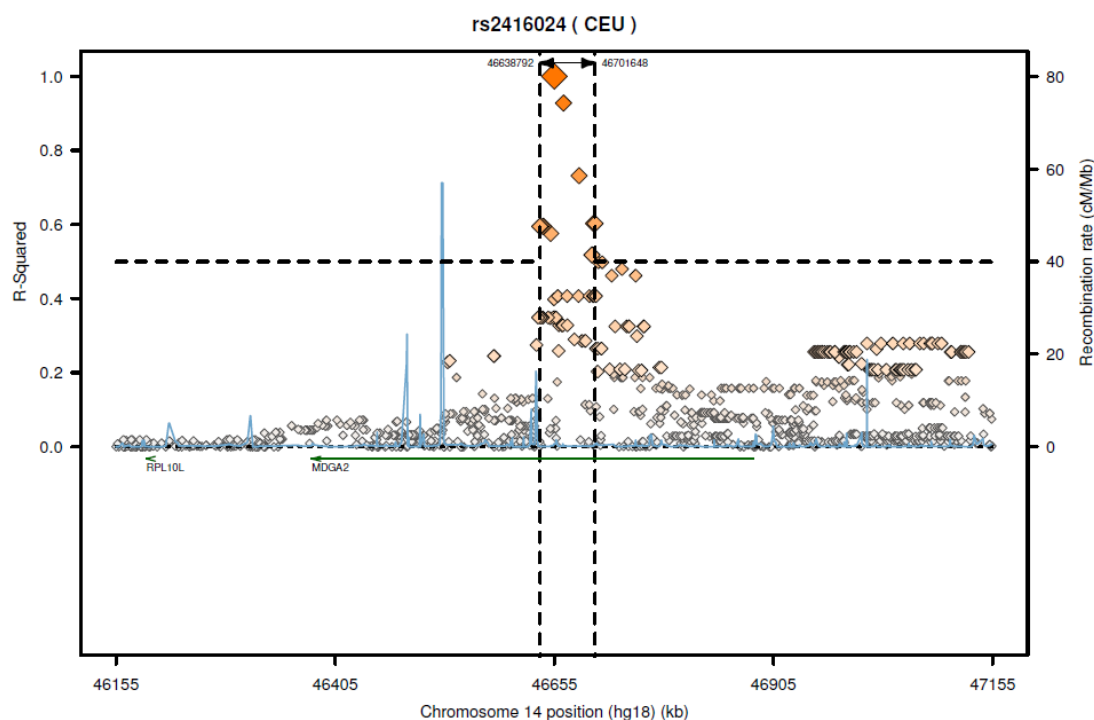


Figure 3.3.3.5.4.E. Regional LD plot for rs2416024 - ranked 8/20 in the GWAS of MDD in males only. This SNP is an intronic variant of *MDGA2*, also known as *MAMDC1*.

The SNP rs4674406, ranked 14th, is in high LD ($r^2=0.71-0.93$) with two variants located within the *GMPPA* (GDP-Mannose Pyrophosphorylase A) gene, rs2385539 (intronic) and rs719337 (5'UTR). A homozygous nonsense mutation in *GMPPA* has been identified in a consanguineous Pakistani family as associating with intellectual disability, in addition to several autonomic dysfunctions²⁸³. The query SNP is also in high LD ($r^2=0.93$) with two variants in *SPEG* (Striated muscle preferentially expressed protein kinase), rs1050816 (3'UTR, and ranked 18th) and rs12464085 (non-synonymous coding variant).

This gene expresses a protein by the same name in cardiac tissue, and has been associated with dilated cardiomyopathy, where the heart becomes enlarged and cannot pump blood efficiently^{284,285}. Patients with dilated cardiomyopathy have been shown to have a higher prevalence of anxiety and depression relative to the general population²⁸⁶. The regional association plot for rs4674406 is shown in Figure 3.3.3.5.4.F. Additional LD plots for the top hits from GWAS of mMDD can be found in Appendix 3.E.

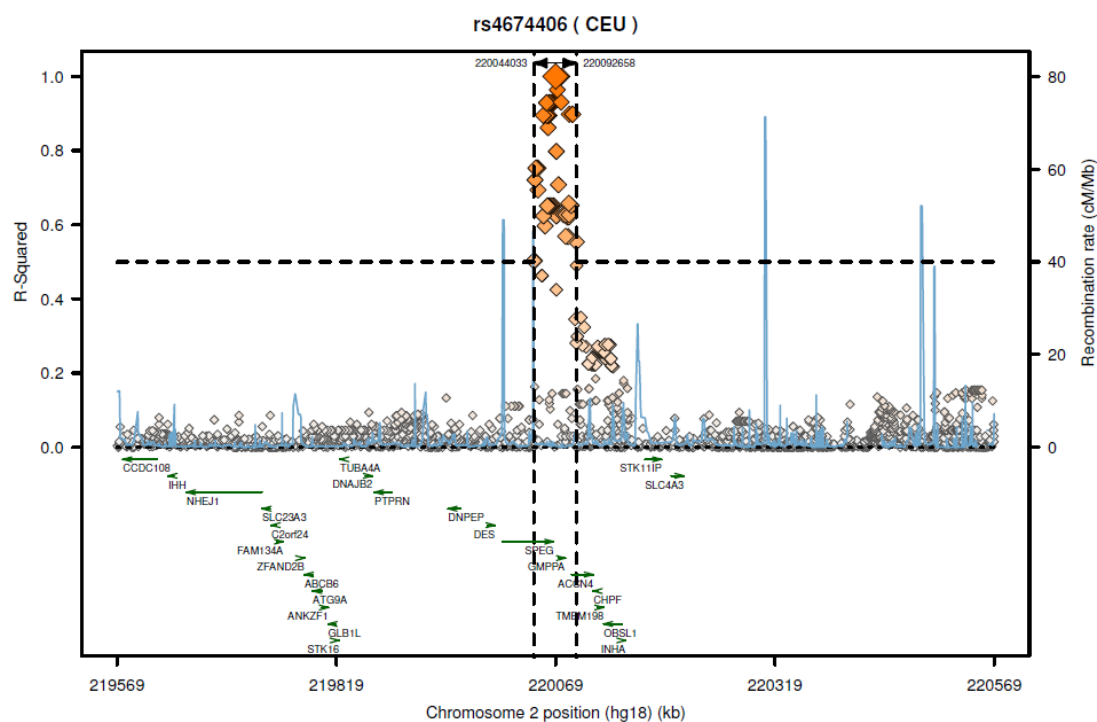


Figure 3.3.3.5.4.F. Regional LD plot for rs4674406 - ranked 14/20 in GWAS of MDD in males only. This variant is upstream of SPEG, in high LD ($r^2=0.93$) with rs1050816 (ranked 18/20) in the 3'UTR of this gene.

3.3.4. Polygenic profiling analysis

Results from GS:SFHS are compared to polygenic risk score (PGRS) analysis from the PGC MDD mega-analysis⁴⁷ for MDD PGRS, and from the PGC cross-disorder analysis²⁰⁶ for other psychiatric PGRS as, with the exception of schizophrenia PGRS, these analyses used effect size weights from the same GWAS summary statistics as have been used here. All p-values reported are uncorrected for multiple testing, since all tests concern overlapping individuals and thus have a complex dependence structure. However, as PGRS analysis was performed using five p-value thresholds (P_T) for four psychiatric disorders (MDD, autism, bipolar disorder and schizophrenia) across 17 traits (including work from Chapters 4 and 5), a conservative Bonferroni correction is $P < 0.05/340 = 1.47E-04$.

3.3.4.1. Association with MDD PGRS

PGRS for MDD were positively associated with all MDD definitions in GS:SFHS. $P_T \leq 1$ and ≤ 0.5 had the strongest associations and explained the highest percentage of variance across MDD definitions. This is to be expected, as the optimum P_T for a given trait dependent on the proportion of null markers which, given the literature, is likely to be low in MDD and therefore a more lenient P_T is optimal⁷⁸. As $P_T \leq 0.5$ explained the most variance in the discovery dataset, this is the P_T which will be summarized here. At this P_T , the variance explained by PGRS for MDD = 0.08%, $rMDD$ = 0.09%, $fMDD$ = 0.08% and $mMDD$ = 0.11%. However, only associations with MDD and $rMDD$ survived multiple testing correction ($p = 1.26E-04$ and $p = 8.26E-05$, respectively), with the strongest association being between MDD PGRS and $rMDD$. All effect sizes were small in magnitude ($OR\ 1.04 \leq 1.12$), with the exception of $mMDD$ at $P_T \leq 0.5$, with an $OR = 2.37$ ($CI: 2.32, 2.42$). However, this is likely due to an anomaly in the Taylor transformation where if the beta on the linear scale (here, $z\beta = 0.10$) is greater than the disease prevalence (here, $P = 0.09$), the odds ratio can become larger than would be expected given the other test statistics. How to resolve this issue is currently under discussion (Ella Wigmore and Jian Yang, personal

communication). The variance in MDD definition explained by PGRS was very small ($\leq 0.11\%$), however this is consistent with the results from the PGC MDD mega-analysis (from which the effect size weights for MDD PGRS were derived) where the percentage of variance explained was small, 0.6%, but highly significant ($P < 10E-6$)⁴⁷. Results from all P_T are presented in Table 3.3.4.1., with variance explained by PGRS visualized in Figure 3.3.4.1.

Trait	PGRS P_T	OR(CI)	Z- ratio	F statistic (df, dendf)	P-value	% variance explained
MDD	1	1.09 (1.04,1.14)	4.09	16.7 (1, 14797)	4.33E-05	0.09
	0.5	1.09 (1.04,1.14)	3.84	14.7 (1, 14764)	1.26E-04	0.08
	0.1	1.08 (1.03,1.13)	3.46	12.0 (1, 14941)	5.32E-04	0.07
	0.05	1.07 (1.02,1.12)	3.05	9.3 (1, 15143)	0.002	0.05
	0.01	1.04 (0.98,1.09)	1.66	2.8 (1, 15047)	0.10	0.02
rMDD	1	1.12 (1.07,1.18)	4.12	17.0 (1, 14037)	3.77E-05	0.10
	0.5	1.12 (1.07,1.18)	3.94	15.5 (1, 14010)	8.26E-05	0.09
	0.1	1.11 (1.05,1.17)	3.31	10.9 (1, 14093)	9.46E-04	0.07
	0.05	1.11 (1.05,1.17)	3.35	11.2 (1, 14278)	8.01E-04	0.07
	0.01	1.06 (1.00,1.12)	1.81	3.3 (1, 14173)	0.07	0.02
fMDD	1	1.09 (1.03,1.15)	3.28	10.7 (1, 9846)	0.001	0.10
	0.5	1.08 (1.03,1.14)	2.91	8.5 (1, 9816)	0.004	0.08
	0.1	1.07 (1.01,1.12)	2.42	5.8 (1, 98520)	0.02	0.06
	0.05	1.05 (1.00,1.11)	2.06	4.2 (1, 9887)	0.04	0.04
	0.01	1.04 (0.99,1.10)	1.63	2.7 (1, 9808)	0.10	0.02

mMDD	1	1.11 (1.04,1.18)	2.83	8.0 (1, 6992)	0.005	0.10
	0.5	2.37 (2.32,2.42)	2.90	8.4 (1, 6986)	0.004	0.11
	0.1	1.11 (1.04,1.18)	2.79	7.8 (1, 7032)	0.005	0.10
	0.05	1.11 (1.04,1.18)	2.61	6.8 (1, 7147)	0.009	0.009
	0.01	1.04 (0.97,1.11)	0.84	0.7 (1, 7159)	0.40	0.009

Table 3.3.4.1. Association with and variance in trait score explained by polygenic risk score for major depressive disorder. P-values which survived multiple testing correction are highlighted in bold.

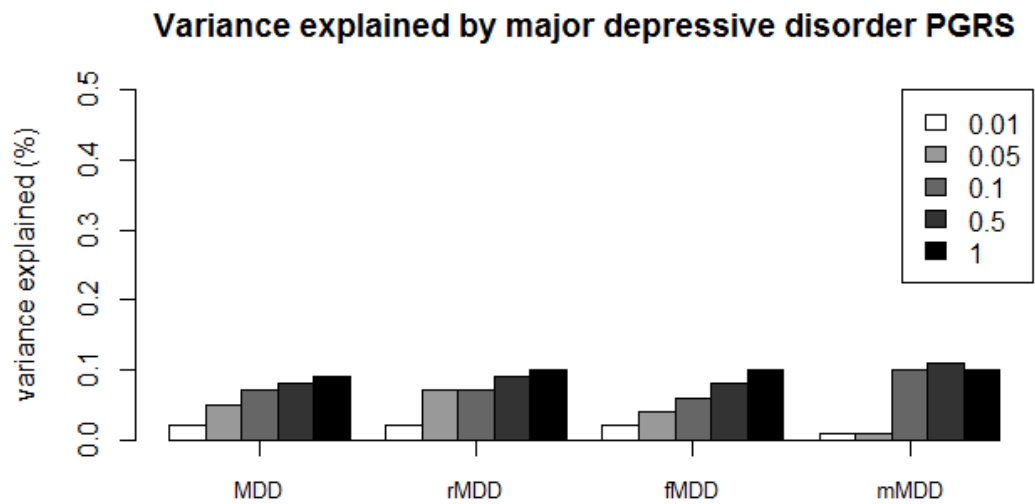


Figure 3.3.4.1. Percentage variance explained in MDD definitions in GS:SFHS by MDD PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

3.3.4.2. Association with autism PGRS

PGRS for autism were not associated with MDD definitions in GS:SFHS at any P_T , with all confidence intervals spanning 1. Variance explained by autism PGRS was negligible ($0.0002 \leq 0.02\%$). This is consistent with previous work from the PGC cross-disorder group where autism PGRS explained a maximum of 0.05% of trait variance for MDD²⁰⁶. This is also consistent with analyses (using the same samples as Smoller *et al*) which estimated the genetic correlation (r_g) between autism and MDD to be not significantly different from zero ($r_g=0.05$; $SE=0.09$)¹⁹. Results from all P_T are presented in Table 3.3.4.2., with variance explained by PGRS visualized in Figure 3.3.4.2.

Trait	PGRS PT	OR(CI)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
MDD	1	0.97 (0.92,1.03)	-1.04	1.1 (1, 14961)	0.3	0.006
	0.5	0.97 (0.92,1.03)	-1.13	1.3 (1, 14963)	0.26	0.007
	0.1	0.98 (0.93,1.03)	-0.77	5.9 (1, 15016)	0.44	0.003
	0.05	0.99 (0.94,1.04)	-0.57	0.33 (1, 14940)	0.57	0.002
	0.01	0.97 (0.92,1.03)	-1.14	1.3 (1, 15135)	0.25	0.007
rMDD	1	0.94 (0.88,1.00)	-1.89	3.6 (1, 14106)	0.06	0.02
	0.5	0.94 (0.88,1.00)	-1.87	3.5 (1, 14102)	0.06	0.02
	0.1	0.95 (0.89,1.02)	-1.53	2.3 (1, 14128)	0.13	0.01
	0.05	0.97 (0.91,1.03)	-1.2	1.4 (1, 14104)	0.23	0.009
	0.01	0.99 (0.93,1.05)	-0.18	0.03 (1, 14234)	0.85	2.08E-04
fMDD	1	0.98 (0.92,1.04)	-0.91	8.4 (1, 9909)	0.36	0.008
	0.5	0.97 (0.91,1.03)	-1.08	1.2 (1, 9894)	0.28	0.01

	0.1	0.99 (0.93,1.04)	-0.65	0.4 (1, 9894)	0.51	0.004
	0.05	0.98 (0.92,1.04)	-0.77	0.6 (1, 9887)	0.44	0.006
	0.01	0.96 (0.90,1.01)	-1.61	2.6 (1,9877)	0.11	0.02
mMDD	1	0.98 (0.90,1.05)	-0.58	0.3 (1, 1729)	0.56	0.004
	0.5	0.98 (0.90,1.05)	-0.52	0.3 (1, 7131)	0.6	0.004
	0.1	0.99 (0.92,1.06)	-0.44	0.02 (1, 7095)	0.66	0.003
	0.05	1.01 (0.93,1.08)	0.16	0.02 (1, 7026)	0.88	3.22E-04
	0.01	1.02 (0.95,1.10)	0.58	0.3 (1, 7076)	0.56	0.005

Table 3.3.4.2. Association with and variance in trait score explained by polygenic risk score for autism. P-values which survived multiple testing correction are highlighted in bold.

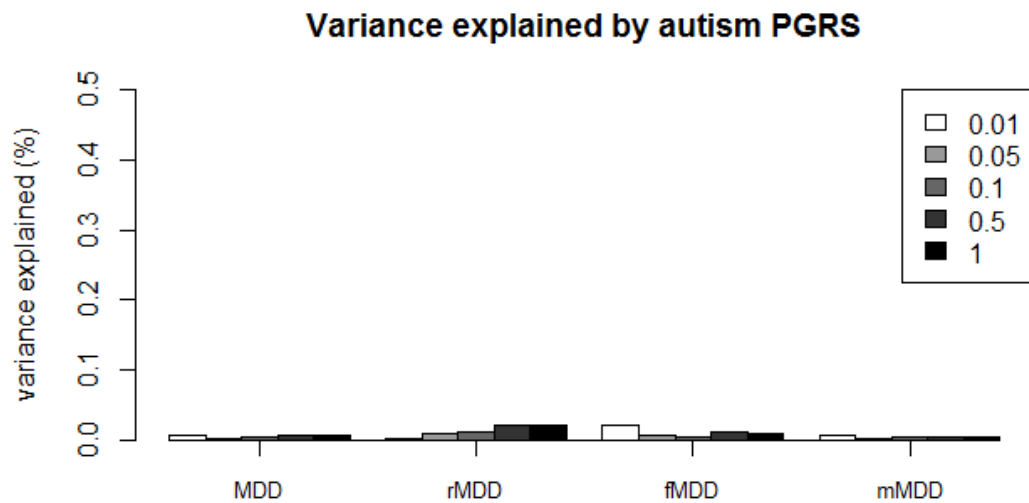


Figure 3.3.4.2. Percentage variance explained in MDD definitions in GS:SFHS by autism PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

3.3.4.3. Association with bipolar disorder PGRS

PGRS for bipolar disorder were positively associated with all MDD definitions in GS:SFHS, although no P_T survived multiple correction testing. As $P_T \leq 0.5$ explained the most variance in the discovery dataset⁵¹, this is the P_T which will be summarized here. At this P_T , the variance explained by PGRS for MDD = 0.04%, rMDD = 0.05%, fMDD = 0.03% and mMDD = 0.10%. This is slightly lower than in previous work from the PGC cross-disorder group where bipolar disorder PGRS explained a maximum of 0.5% of trait²⁰⁶. PGRS for bipolar disorder explained the most variance for mMDD (0.1%-0.14%), and in fact explained more variance for mMDD than the MDD PGRS (0.09%-0.11%). This could represent a greater shared architecture between MDD and bipolar disorder in males. MDD and bipolar disorder are thought to have moderate shared architecture, with estimates of $r_g(\text{SE}) = 0.47 (0.06)$ from genomic data¹⁹ and $r_g(\text{CI}) = 0.65 (0.58, 0.75)$ from twin data. Another possibility could be the misdiagnosis of bipolar cases as unipolar depression cases. It has previously been estimated that up to 69% of bipolar cases are initially misdiagnosed, predominantly as unipolar depression, with individuals consulting a mean of 4 psychiatrists before being diagnosed correctly²⁸⁷. This possibility could not be known without longitudinal data in GS:SFHS. However it is worth noting that most studies report equal prevalence of illness in males and females²⁸⁸, so one would expect to see greater variance explained across MDD definitions if misdiagnosis was a significant factor. All effect sizes were small in magnitude ($\text{OR } 1.06 \leq 1.11$). The OR for mMDD at $P_T \leq 0.1$ suffered from the same transformation problem as discussed in Section 3.3.3.1., due to the beta from the linear scale ($z\beta = 0.10$) being greater than the disease prevalence ($P = 0.09$). Results from all P_T are presented in Table 3.3.4.3., with variance explained by PGRS visualized in Figure 3.3.4.3.

Trait	PGRS P _T	OR(CI)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
MDD	1	1.06 (1.01,1.11)	2.79	7.8 (1, 15172)	0.005	0.04
	0.5	1.06 (1.01,1.11)	2.80	7.9 (1, 15140)	0.005	0.04
	0.1	1.06 (1.01,1.11)	2.78	7.7 (1, 15244)	0.006	0.04
	0.05	1.05 (1.00,1.10)	2.36	5.6 (1, 15145)	0.02	0.03
	0.01	1.04 (0.98,1.09)	2.02	4.1 (1, 14939)	0.04	0.02
rMDD	1	1.09 (1.03,1.15)	2.95	8.7 (1, 14293)	0.003	0.05
	0.5	1.09 (1.03,1.15)	2.98	8.9 (1, 14241)	0.003	0.05
	0.1	1.09 (1.03,1.15)	2.87	8.2 (1, 14314)	0.004	0.05
	0.05	1.08 (1.02,1.14)	2.64	7.0 (1, 14209)	0.008	0.04
	0.01	1.08 (1.02,1.14)	2.58	6.6 (1, 14065)	0.01	0.04
fMDD	1	1.04 (0.99,1.10)	1.67	2.8 (1, 9888)	0.10	0.03
	0.5	1.04 (0.99,1.10)	1.65	2.72 (1, 9905)	0.10	0.03
	0.1	1.04 (0.98,1.10)	1.38	1.9 (1, 9990)	0.17	0.02
	0.05	1.02 (0.96,1.08)	0.78	0.6 (1, 10013)	0.44	0.006
	0.01	1.01 (0.96,1.07)	0.61	3.7 (1, 9976)	0.54	0.003
mMDD	1	1.11 (1.04,1.18)	2.71	7.4 (1, 7097)	0.007	0.10
	0.5	1.11 (1.04,1.18)	2.77	7.7 (1, 7065)	0.006	0.10
	0.1	2.37 (2.32,2.42)	2.95	8.7 (1, 7105)	0.003	0.12
	0.05	1.14 (1.07,1.20)	3.25	10.5 (1, 7069)	0.001	0.14
	0.01	1.12 (1.06,1.19)	2.92	8.6 (1, 6967)	0.003	0.11

Table 3.3.4.3. Association with and variance in trait score explained by polygenic risk score for bipolar disorder. P-values which survived multiple testing correction are highlighted in bold.

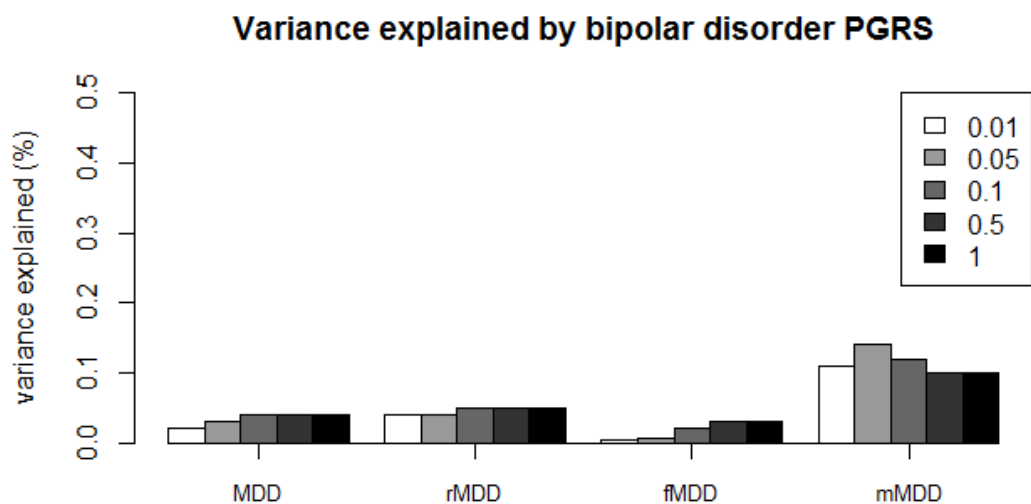


Figure 3.3.4.3. Percentage variance explained in MDD definitions in GS:SFHS by bipolar disorder PGRS at five p -value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

3.3.4.4. Association with schizophrenia PGRS

PGRS for schizophrenia were positively associated with all MDD definitions in GS:SFHS, with several P_T surviving multiple correction testing. As $P_T \leq 0.1$ explained the most variance in the discovery dataset²³, this is the P_T which will be summarized here. At this P_T , the variance explained by PGRS for MDD = 0.15%, rMDD = 0.10%, fMDD = 0.15% and mMDD = 0.16%. Schizophrenia PGRS explained more variance in MDD, fMDD and mMDD than MDD PGRS, and was more significantly associated with undifferentiated MDD. This could be simply related to the increased power of the discovery sample for schizophrenia PGRS (32,838 cases and 44,357 controls) relative to MDD PGRS (4,984 cases and 6,641 controls). Despite this analysis using weights from the largest available schizophrenia GWAS, in which the schizophrenia PGRS explained 18% of the variance in the target dataset²³, less variance is explained in GS:SFHS MDD by schizophrenia PGRS (0.15%) than at the same P_T in the cross disorder analysis (0.7%)²⁰⁶ where marker weights were derived from a GWAS explaining ~6% of

variance in the target dataset⁵⁰. Results from all P_T are presented in Table 3.3.4.4., with variance explained by PGRS visualized in Figure 3.3.4.4.

Trait	PGRS P_T	OR(CI)	Z-ratio	F statistic (df, dendf)	P-value	% variance explained
MDD	1	1.12 (1.07,1.17)	4.92	24.2 (1, 14727)	8.65E-07	0.15
	0.5	1.12 (1.07,1.17)	4.84	23.4 (1, 14705)	1.30E-06	0.14
	0.1	1.12 (1.07,1.17)	4.91	24.1 (1, 14703)	9.19E-07	0.15
	0.05	1.10 (1.05,1.15)	4.26	18.2 (1, 14714)	2.05E-05	0.11
	0.01	1.11 (1.06,1.16)	4.46	19.9 (1, 14772)	8.12E-06	0.13
rMDD	1	1.12 (1.07,1.18)	3.6	12.9 (1, 13852)	3.22E-04	0.09
	0.5	1.11 (1.05,1.17)	3.49	12.2 (1, 13828)	4.79E-04	0.08
	0.1	1.12 (1.07,1.18)	3.83	14.7 (1, 13852)	1.29E-04	0.10
	0.05	1.11 (1.05,1.17)	3.41	11.6 (1, 13817)	6.63E-04	0.08
	0.01	1.12 (1.07,1.18)	3.61	13.1 (1, 13838)	3.01E-04	0.09
fMDD	1	1.11 (1.06,1.17)	3.98	15.9 (1, 9755)	6.82E-05	0.16
	0.5	1.11 (1.05,1.16)	3.85	14.8 (1, 9747)	1.19E-04	0.15
	0.1	1.11 (1.05,1.16)	3.74	14.0 (1, 9748)	1.89E-04	0.15
	0.05	1.11 (1.05,1.16)	3.59	12.9 (1, 9774)	3.31E-04	0.14
	0.01	1.11 (1.05,1.16)	3.68	13.6 (1, 9817)	2.30E-04	0.15
mMDD	1	1.12 (1.06,1.19)	3.01	9.1 (1, 6885)	0.003	0.13
	0.5	1.11 (1.02,1.20)	3.07	9.4 (1, 6859)	0.002	0.14
	0.1	1.15 (1.06,1.24)	3.34	11.2 (1, 6844)	8.38E-04	0.16
	0.05	1.10 (1.01,1.19)	2.40	5.7 (1, 6853)	0.02	0.09
	0.01	1.11 (1.02,1.20)	2.48	6.1 (1, 6839)	0.01	0.10

Table 3.3.4.4. Association with and variance in trait score explained by polygenic risk score for schizophrenia. P-values which survived multiple testing correction are highlighted in bold.

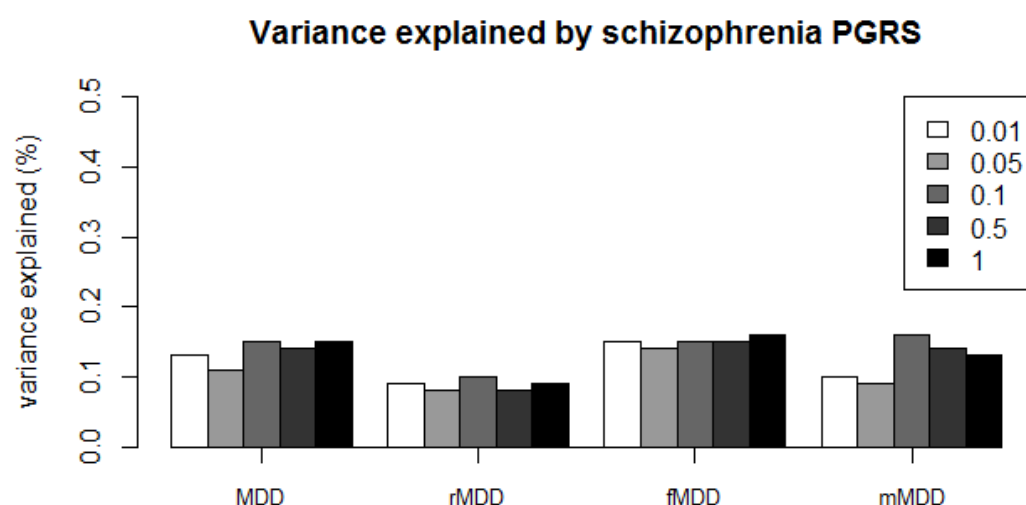


Figure 3.3.4.4. Percentage variance explained in MDD definitions in GS:SFHS by schizophrenia PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

3.3.5. Replication in UKBiobank

3.3.5.1. Paired samples sign test

The ORs of SNPs in GS:SFHS and UKB were tested for consistent directionality using SNPs with a p-value of $<1E-04$ from GWAS of MDD definitions in GS:SFHS to perform a paired samples sign test. This resulted in testing 50 SNPs for MDD, 79 SNPs for rMDD, 60 SNPs for fMDD and 44 SNPs for mMDD. For these SNPs, the replication samples did not produce ORs in the same directions as the discovery analysis more frequently than expected by chance (unadjusted p-value = 0.05). Full results from the paired samples sign test are shown in Table 3.3.5.1.

Trait	UKB definition	nTrials	nSuccesses	Probability of success (CI)	P-value
MDD	broad	50	19	0.38 (0.24, 0.53)	0.12
	narrow		26	0.52 (0.37, 0.66)	0.89
	strict		24	0.48 (0.34, 0.63)	0.89
rMDD	broad	79	48	0.61 (0.49, 0.72)	0.07
	narrow		40	0.51 (0.39, 0.62)	1
	strict		47	0.59 (0.48, 0.70)	0.11
fMDD	broad	60	33	0.55 (0.42, 0.68)	0.52
	narrow		36	0.60 (0.47, 0.72)	0.16
	strict		34	0.57 (0.43, 0.69)	0.37
mMDD	broad	44	24	0.55 (0.39, 0.65)	0.65
	narrow		15	0.34 (0.20, 0.50)	0.05
	strict		26	0.59 (0.43, 0.74)	0.29

Table 3.5.5.1. Paired samples sign test between GS:SFHS and UKB MDD definitions, with nTrials indicating the number of SNPs $p < 1E-04$, and nSuccesses indicating how many times ORs were in the same direction in both samples.

These results do not give any indication as to which UKB MDD definition is most correlated with MDD definitions in GS:SFHS.

3.3.5.2. Bivariate LD Score regression

Results from bivariate LD Score regression of summary statistics from MDD GWAS in GS:SFHS and UKB are shown in Table 3.3.5.2. Broad, narrow and strict definitions of MDD in UKB were positively correlated with MDD, rMDD and fMDD in GS:SFHS. Bivariate LD Score regression between UKB MDD definitions

and mMDD did not return any results. This can occur if the heritability is estimated to be <0 , as the genetic correlation (r_g) is calculated as $r_g = \rho_g / \sqrt{h_i^2 h_j^2}$ where ρ_g is the genetic covariance between traits, h_i^2 is the heritability of trait i and h_j^2 is the heritability of trait j ¹²³. Therefore if one trait has a negative heritability estimate, it results in the square root of a negative number, which is invalid. Although univariate LD Score regression indicated a non-negative h^2 estimate for mMDD ($h^2 = 0.07$; $SE = 0.16$), h^2 estimates generated by the bivariate analysis can differ slightly as this estimate will use SNPs in common across both traits. Indeed, in the univariate analysis, the h^2 estimate for mMDD utilized 464,656 SNPs that were in common between GS:SFHS summary statistics and the pre-computed LD Score SNP set, whereas the h^2 estimates in the bivariate analysis utilized 460,951 SNPs that were in common between GS:SFHS, UKB and the LD Score SNP set. Therefore, a negative h^2 estimate may have been returned using this reduced SNP set. In addition, mMDD has the smallest sample size of the GS:SFHS MDD definitions ($n=8,185$), which can result in near zero true heritabilities being estimated as negative values due to sampling variance¹⁹⁹.

The genetic correlation between undifferentiated MDD in GS:SFHS and all UKB MDD definitions was >1 . This is due to the low h^2 estimate for MDD in GS:SFHS ($h^2 = 0.04$; $SE = 0.07$) as, again due to h^2 being included in the denominator term of the genetic correlation calculation, h^2 being a positive near zero value can cause the genetic correlation to become severely inflated. Genetic correlations between UKB MDD definitions and rMDD and fMDD in GS:SFHS yield more sensible estimates due to these definitions of MDD having higher heritability estimates. The genetic correlation between rMDD and UKB was high for all definitions, with $r_g (SE) = 1.01 (0.46)$ for broad MDD, $r_g (SE) = 0.74 (0.48)$ for narrow MDD and $r_g (SE) = 0.87 (0.45)$ for strict MDD. The most significant genetic correlation for rMDD was with broad MDD ($p=0.01$), although this does not survive multiple testing correction. The genetic correlation between fMDD and UKB was high for all definitions, with $r_g (SE) = 0.89 (0.35)$ for broad MDD, r_g

(SE) = 0.82 (0.43) for narrow MDD and r_g (SE) = 1.00 (0.42) for strict MDD. The most significant genetic correlation for fMDD was with broad MDD ($p=0.01$), although this does not survive multiple testing correction ($p \leq 0.002$). Standard errors are large for all correlations, this could be reduced by using imputed data to increase the number of SNPs used in the analysis. Similarly, the use of imputed data could resolve issues for analyses using MDD and mMDD by potentially increasing the heritability estimates for these traits.

Trait	UKB definition	r_g (SE)	P-value	h^2_{liab} (SE)	h^2_{int} (SE)	$gcov_{int}$ (SE)
MDD	Broad	1.95 (1.68)	0.25	0.04 (0.07)	0.994 (0.007)	-0.007 (0.005)
	Narrow	1.48 (1.38)	0.29	0.03 (0.05)	0.994 (0.007)	-0.007 (0.005)
	Strict	1.91 (1.67)	0.25	0.04 (0.07)	0.994 (0.007)	-0.007 (0.005)
rMDD	Broad	1.01 (0.46)	0.03	0.135 (0.10)	1 (0.007)	-0.003 (0.006)
	Narrow	0.74 (0.48)	0.05	0.135 (0.10)	1 (0.007)	-0.004 (0.005)
	Strict	0.87 (0.45)	0.05	0.135 (0.10)	1 (0.007)	-0.003 (0.005)
fMDD	Broad	0.89 (0.35)	0.01	0.19 (0.09)	0.997 (0.008)	-0.009 (0.005)
	Narrow	0.82 (0.43)	0.06	0.18 (0.09)	0.997 (0.008)	-0.008 (0.005)
	Strict	1 (0.42)	0.02	0.18 (0.09)	0.997 (0.008)	-0.01 (0.005)
mMDD	Broad	NA	NA	NA	NA	NA
	Narrow	NA	NA	NA	NA	NA
	Strict	NA	NA	NA	NA	NA

Table 3.3.5.B. Bivariate LD Score regression between MDD, rMDD, fMDD and mMDD in GS:SFHS and the broad, narrow and strict MDD phenotype definitions in UK Biobank. P-values which survived multiple testing correction are highlighted in bold.

3.4. Discussion

The main purpose of this chapter was to observe how the present sample corresponds with the existing literature for MDD. In addition, analyses which utilized genomic data sought to establish that basic stratification of the phenotype into potentially more familial subtypes was insufficient to yield statistically meaningful results, and thus further justify the primary aim of this thesis - the identification of quantitative measures of depression using genetically correlated traits.

3.4.1. Summary of findings

3.4.1.1. Is MDD in GS:SFHS consistent with the literature?

Looking first at descriptive statistics for MDD in GS:SFHS, there was a significantly ($p < 2.2e-16$) higher proportion of females in the case group (71.2%) compared to the control group cases than controls (60.0%). This is consistent with previously reports which estimate lifetime prevalence being approximately two times greater in females than males¹⁰. A small but significant difference in age distribution between cases and controls was observed ($p = 4.87e-05$), indicating that cases were slightly younger than controls. This may be due to underreporting in depression in older subjects¹⁹⁷. Although the age distribution appears to indicate that MDD is most prevalent in mid-life, as previously described in the literature, it is important to remember that MDD as described here, and throughout this thesis, includes both individuals who are currently depressed and those who have previously suffered one or more episodes of MDD. Therefore, without looking at these subgroups in more detail, it is not possible to say whether MDD in GS:SFHS is consistent with previous observations regarding increased incidence in middle age, or the non-linear association between age and MDD^{176,177}.

In terms of prevalence of MDD and its subtypes, GS:SFHS is consistent with estimates from the literature. The sample prevalence of MDD was 13.4%, which is within the range of estimates (8.2%-16.9%) previously inferred from recent

studies in continent Europe, the USA and Canada^{1,194,195}. The sample prevalence for rMDD (7.1%) was consistent with available prevalence estimates for rMDD in the population (6.9%). Sex-specific sample prevalence estimates (16.2% in females, 9.3% in males) were also in line with estimates derived from averaging over three European populations (17.8% in females, 7.0% in males)¹⁰.

Heritability estimates for MDD and its subtypes using pedigree data in GS:SFHS was in line with other estimates from the literature. The heritability of undifferentiated MDD when unadjusted, adjusted for sociodemographic correlates and adjusted for a broad common family environment effect all had confidence intervals overlapping with the Sullivan *et al*¹⁶ meta-analysis estimate. Sex specific heritability estimates were similar to those observed in Kendler *et al*^{20,21}, although genetic correlation between sexes higher in GS:SFHS, suggesting perhaps fewer sex specific mechanisms.

Heritability estimates for MDD and its subtypes using GWAS summary statistics yielded lower point estimates than pedigree based estimates, as would be expected. However, estimates were ranked in the same order as in the pedigree based analysis, with fMDD having the highest estimate, followed by rMDD, mMDD and MDD. Estimates were in the range of previously reported SNP based heritability estimates^{47,198}.

3.4.1.2. Does basic stratification improve etiological insight?

3.4.1.2.1. GWAS

No definition of MDD yielded any genome-wide significant hits in the genome-wide association analysis. Looking at the distribution of test statistics, recurrent MDD and MDD in females had a higher number of SNPs with p-values <1E-04 (79 and 60, respectively) compared to undifferentiated MDD and MDD in males (50 and 44, respectively). Whilst rMDD and fMDD showed some inflation in test statistics, LD Score analysis ratified that this was likely due to polygenicity rather than population and/or family structure. This suggests that more familial subtypes are yielding a higher proportion of extreme observations given the

data, unfortunately just not to an extent where the null hypothesis can be rejected for any given variant! This is still interesting, given the drop in sample size when using only recurrent cases (1,327 cases; 17,237 controls), or only females (1,894 cases; 9,817 controls). These results are also in keeping with results from the majority of GWAS conducted for MDD to date. From a biological standpoint, many biologically interesting loci previously reported for psychiatric disorders or known clinical covariates of depression were amongst the highest ranked findings. However, it is important to remember that given the wealth of published information available via online databases, it is possible to construct a plausible narrative by simply performing a sufficiently comprehensive literature search. Especially as the examples cited from the literature in this chapter encompassed a whole range of study designs, sample sizes and replication success. It is important to be mindful of publication bias, and the subsequent failure of many studies to attribute results to chance²⁸⁹.

3.4.1.2.2. PGRS

The association of MDD PGRS with MDD definitions is in line with previous work⁴⁷, with the association being significant, but the variance in trait explained by PGRS remaining very small ($\leq 0.1\%$). At $P_T \leq 0.5$, only associations for MDD and rMDD survived multiple testing correction.

Regarding the enrichment of associated alleles for other psychiatric disorders, these are pretty consistent with work from the PGC cross disorder group²⁰⁶. Autism PGRS showed no enrichment in MDD or its subtypes in GS:SFHS. Bipolar disorder PGRS explained less variance than MDD PGRS, with the exception of mMDD where both PGRS explained a similar proportion of trait variance, however no P_T survived multiple testing correction. Schizophrenia PGRS explained the most trait variance for MDD, fMDD and mMDD, but not rMDD (where MDD PGRS explained the most variance). This is, in part, likely due to the improved precision of the schizophrenia PGRS due to the substantial sample size in the training set²³ from which the marker weights were obtained.

3.4.1.3. Does MDD in GS:SFHS correlate with MDD in UKB?

From the results, it cannot be definitively ascertained whether MDD in GS:SFHS correlates with MDD in UKB. The sign test showed no consistency in direction of effects for SNPs with $P \leq 1E-4$ relative to what would be expected by chance. Regarding results from bivariate LD Score analysis, only rMDD and fMDD had genetic correlations with UKB MDD definitions that were within bounds (0-1). This is due to how the genetic correlation is calculated, as the SNP heritability estimate is in the denominator term. Therefore for mMDD, where the h^2_{SNP} was negative, this caused an impossible value, and for undifferentiated MDD, where the h^2_{SNP} was very low but positive, the genetic correlation became very inflated. No definition reached statistical significance after multiple testing correction, therefore it is not possible to conclusively say which, if any, UKB MDD definition correlates best with MDD as defined in GS:SFHS.

3.4.2. Limitations

3.4.2.1. Model selection

To account for the complex structure in GS:SFHS, mixed linear models were employed to avoid an upward bias of effect sizes. To facilitate the use of all genotyped individuals, both closely and distantly related, two GRMs were fitted as random effects¹⁶⁹. How effective this model was in combating confounding was assessed by looking at the output from univariate LD Score regression, namely: the mean χ^2 statistic, the deviation of the LD Score intercept from 1 divided by its standard error, and the Ratio - which is the proportion of inflation ascribed to causes other than polygenic heritability with its standard error. In all MDD definitions, the mean χ^2 statistic was less than 1.02, which usually indicates genomic control. The deviation of the LD Score intercept from 1 divided by its standard error was non-significant in all cases. The Ratio was also less than zero or NA for all definitions, again usually an indication of genomic control. However, post-hoc genomic control had not been implemented in these analyses. This perhaps suggests that this model is overly conservative, and killing any polygenic signal which does exist. Therefore it would perhaps be

beneficial to the power of the study if further models were tested for optimal fit. Another consideration which could improve results would be to use imputed data rather than raw genotype data. Raw genotype data was used, as imputed data was in the process of being quality controlled towards the end of this project. The use of imputed data would increase the number of SNPs that could be tested for association, and facilitate more detailed fine-mapping. Imputation can lead to a boost in statistical power of up to 10%²⁹⁰. However, imputation does only boost what is already there. So whilst genuine signals would perhaps have more structure, as would signals caused by noise. Using imputed data would be useful for bivariate LD Score regression, as one of the main benefits of using imputed data sets is to facilitate meta-analysis or replication between cohorts genotyped on different panels.

3.4.2.2. Linear vs. Liability scale

Ideally, logistic regression would be used for binary traits. However, given the need for a mixed linear model framework in GS:SFHS, this markedly reduces the number of potential programs suitable to analyze the data. MCMCglmm-R, a package for fitting Generalized Linear Mixed Models using Markov chain Monte Carlo techniques¹⁷⁹, would be an appropriate choice for running logistic regression on the liability scale, therefore providing the correct test statistics without the need for conversion from the linear to the liability scale. The drawback to this program would be the timeframe required to run the analysis. There does not appear to be a GWAS function as such in MCMCglmm-R, therefore the regression would need to be performed for each SNP individually. In addition, each regression would need to be run multiple times to ensure model convergence. Therefore, it is possible to run mixed model analyses for binary traits straight onto the liability scale, but this comes with its own set of limitations.

3.4.2.3. Calculating variance explained by PGRS

Generally, variance explained by PGRS in GS:SFHS was even lower than the already small proportions reported previously. Perhaps this is due to the difference in sample prevalence between this study (13.4%) and MDD in the cross-disorder study (49.3%)²⁰⁶. It could also be due to the different methods used to calculate variance explained. Most studies, including the PGC cross disorder paper, report Nagelkerke's R^2 . This method takes the Cox and Snell R^2 , defined as: $R^2_{C\&S} = 1 - (L_0 / L_M)^{2/n}$ (where L_0 is the value of the likelihood function for a reduced model without fitting the predictor of interest, and L_M is the likelihood for the full model being estimated)²⁹¹, and divides it by its upper bound, $[1 - (L_0)^{2/n}]$ ²⁹². This was in a logistic model setting due to the unrelated study participants. However, in GS:SFHS mixed linear models (MLM) were used (and transformed to the liability scale post-hoc). MLM by definition have more than one variance component, meaning the residual variance cannot be easily defined for non-Gaussian responses. Therefore, in GS:SFHS, variance in trait value explained by polygenic profile score was estimated by multiplying the profile score by its corresponding regression coefficient and estimating its variance. This value was then divided by the variance of the observed phenotype to yield a coefficient of determination between 0 and 1¹⁸⁶. There are other methods of calculating R^2 from generalized linear models, such as the method proposed by Lee *et al*²⁹³, however the debate as to which method is most appropriate for fixed effect, random effect and mixed effect models is ongoing.

3.4.3. Conclusions

MDD in GS:SFHS is consistent with the existing literature for MDD in the European subpopulation in terms of descriptive statistics, prevalence estimates, heritability estimates and a notable lack of genome-wide significant findings. In terms of PGRS analysis, MDD in GS:SFHS seems to perform slightly more poorly than other samples. This could be a product of how variance explained is measured differently between studies. It cannot be ascertained whether MDD in GS:SFHS correlates well with MDD in UKB, therefore it is difficult to say at this stage whether any quantitative measures of MDD will have utility outwith GS:SFHS. This Chapter has provided results against which quantitative traits can be assessed, although the interpretation of this comparison shall be performed tentatively owing the statistical differences between Gaussian and non-Gaussian data.

Chapter 4: Assessment of mood, personality and cognitive phenotypes as candidate endophenotypes for depression

4.1. Introduction

In this Chapter ten measures of mood, personality and cognitive ability will be evaluated and ranked as endophenotypes for major depressive disorder (MDD) using data from Generation Scotland: The Scottish Family Health Study (GS:SFHS). The measures of mood, personality and cognitive ability analyzed here are the 28-item General Health Questionnaire (GHQ) likert scale¹⁵², Mood Disorder Questionnaire (MDQ)^{149,150}, Schizotypal Personality Questionnaire-brief (SPQ)¹⁵¹, neuroticism (EPQN) and extraversion (EPQE) sections of the Eysenck Personality Questionnaire - Short Form Revised^{147,148}, Wechsler digit symbol substitution task (digit symbol coding (DSC))¹⁴⁵, Wechsler logical memory test immediate (LM1) and delayed (LM2D)¹⁴³, Mill Hill Vocabulary Scale (MHV)¹⁴⁶, and phonemic verbal fluency (VF)¹⁴⁴. Further details on how these traits were measured is described in Section 2.1.2.

The endophenotype approach has already been introduced in Section 1.5., but in brief, the initial idea behind this approach is to identify traits which could help to bridge the gap between high level symptom presentation and low level genetic variability, such as single nucleotide polymorphisms. The seminal paper purporting the use of endophenotypes in psychiatric genetics outlined a set of criteria which a candidate endophenotype should fulfill. Namely, that an endophenotype should be heritable, associate with illness in the population, be state independent, and be more extreme in non-affected family members than in the population¹²⁶. Traits should also be genetically and phenotypically correlated with the disease trait of interest. Endophenotypes can be categorized into six groups: anatomical, developmental, electrophysiological, metabolic, sensory or psychological/cognitive. The rationale for studying endophenotypes is the hope that this methodology will provide a genetically tractable target

between gene and disease process that can provide much needed biological insight for psychiatric diseases²⁹⁴.

The use of the endophenotype approach has attracted criticism as, ideally, an endophenotype should have a less complex genetic architecture than the disorder it is aiming to delineate. This might be true of transcriptomic and proteomic phenotypes, but is generally not the case for measures of mood, personality and cognitive ability, such as those being interrogated here^{124,294}. Whilst this is a valid point, it assumes that complexity of architecture is the only metric of tractability. One of the most significant problems for case control studies of complex diseases has been attaining sufficient statistical power to detect small effect sizes, as outlined in Section 2.1.4. Typically, studies of disease phenotypes use a case-control design and studies of quantitative traits are population based. Whether a case-control design is likely to outperform a quantitative trait in terms of statistical power for a given sample size depends on the prevalence of the disease. For diseases with a lower prevalence (<10%), a case control design will have greater power to detect an equivalent effect size relative to an equally sized population design for a quantitative trait (provided there are an equal number of cases and controls). Conversely, when the disease prevalence equals the proportion of cases (as is the case for MDD in GS:SFHS), an equally sized study for a quantitative trait has greater power. This is due to the information which is lost when transforming a quantitative trait (susceptibility to disease, or disease liability) into a binary trait (MDD case/control status)⁹⁶. Therefore, for a trait with moderate prevalence such as MDD identifying quantitative traits which have a high genetic correlation with MDD, regardless of complexity of architecture, could increase statistical power sufficiently to afford improved etiological insight.

Analyses presented in this Chapter seek to answer four questions: do the selected measures of mood, personality and cognitive ability fulfill the endophenotype criteria as outlined by Gottesman and Gould¹²⁶; are estimates of coheritability and variance explained by polygenic risk score (PGRS) convergent ranking methods for candidate endophenotypes; do the highest ranking

endophenotypes demonstrate improved statistical power and etiological insight relative to the binary definition of MDD; and do the highest ranking endophenotypes correlate with MDD in an independent sample?

Firstly, results from three types of analysis assessing adherence of traits to the endophenotype criteria will be presented. A Wald Conditional F-test between trait and group (case, unaffected relative, controls) will be used to establish whether traits are associated with illness in the population, are more extreme in unaffected relatives than in the general population and are state independent. Univariate heritability analysis will be used to estimate the pedigree-based heritability of traits. Heritability will also be calculated using summary statistics from GWAA, as per Chapter 3, to observe whether the ranking of traits by their SNP-based heritability estimates are consistent with pedigree based estimates and to establish whether the mixed model is appropriately accounting for potential confounders. Bivariate heritability analysis will be used to estimate genetic and phenotypic correlations between traits, consequently inferring whether traits cosegregate with MDD within families.

Secondly, results from analyses seeking to rank candidate endophenotypes using measures of genetic covariation will be presented. Endophenotypes will be ranked according to their coheritability, or ERV¹²⁴, with MDD and by variance in trait mean explained by PGRS for MDD. Despite associations between PGRS for psychiatric disorders and various quantitative traits being extensively reported^{85-87,89,90,295}, no studies have reported using PGRS explicitly as a tool to identify and rank potential endophenotypes. As this method utilizes only common trait-associated variants, selecting quantitative traits for further genetic analysis based on this method may prove advantageous for gene discovery in GWAA. However, it is not known if these ranking methods give convergent results when measured in the same sample, therefore this shall be tested here. To assess whether the selected measures mood, personality and cognitive ability are specifically correlated with MDD or more generally with psychiatric disorders, traits are also tested for association with PGRS for autism, bipolar disorder and schizophrenia.

Next, results will be presented which assess whether the highest ranking endophenotypes demonstrate improved power relative to binary MDD. This is evaluated by examining the distribution of summary statistics from GWAA to see whether quantitative traits result in an increased number of SNPs with more extreme p-values ($P < E-04$) relative to binary MDD. Top ranking SNPs will be subjected to biological annotation to determine whether genetically correlated quantitative traits can offer etiological insight beyond that of binary MDD. Variance explained by PGRS for MDD will also be used as a metric of improved power, although this is not strictly a valid comparison given the difference in variance structures between Gaussian and non-Gaussian traits.

Lastly, results will be presented from analyses seeking to assess how well the highest ranking endophenotypes in GS:SFHS correlates with definitions of MDD in UKBiobank^{187,189}. This will be achieved using GWAA summary statistics from both samples to perform a paired-samples sign test and bivariate LD score regression¹²³. These results will also indicate whether quantitative traits genetically correlated with MDD have increased utility for use as in populations other than GS:SFHS, relative to binary MDD.

The results from this Chapter will be used to identify which traits to use in Chapter 5, which explores the ability of multivariate methods to identify genetic variants for MDD.

4.2. Methods

4.2.1. Testing endophenotype criteria

Only phenotype data from genotyped individuals (n=19,994) was used in these analyses. Phenotypes were transformed and standardized to improve normality, as described in Section 2.1.3.1. To reduce confounding caused by familial relationships in some analyses (Section 4.2.1.1.-4.2.1.3), the sample was stratified by affection status into cases (n=2,659), unaffected first degree relatives (n=2,106) and unrelated controls (n=11,898).

4.2.1.1. Associated with illness in the population

The effect of affection status on mean trait scores was calculated using the Wald Conditional F-test, implemented in ASReml-R¹⁷³, as described in Section 2.2.4. In the mixed linear model, affection status was fitted as a fixed effect factor with unrelated controls as the reference group. F-statistics, degrees of freedom and p-values were derived, and are presented here. A conservative Bonferroni corrected p-value threshold of 3.85E-03 (incorporating the number of tests in Chapters 3 and 5 for this analysis type) was used to indicate statistical significance. Traits were deemed significantly associated with MDD in the population if the mean trait difference between cases and unrelated controls survived multiple testing correction.

4.2.1.2. More extreme in unaffected relatives than in the general population

This endophenotype criteria is simply an extension of that described in Section 4.2.1.1., with the additional requirement that differences in trait mean between unaffected first degree relatives and unrelated controls also had to survive multiple testing correction and be in the same direction as cases.

4.2.1.3. State independence

It was not possible to explicitly test state independence explicitly, as repeated measures are not yet available in GS:SFHS, and the majority of cases (~80%) were currently depressed at the time of assessment. This introduces collinearity between case/control status and current depression. Therefore, whether the endophenotype was more extreme in unaffected relatives than the general population was used as a proxy measure of state independence²⁹⁶.

4.2.1.4. Heritability of endophenotypes

4.2.1.4.1. Using pedigree data

Heritability estimates for candidate endophenotypes were estimated using restricted maximum likelihood methods, implemented in ASReml-R¹⁷³, as described in Section 2.2.1.3. The heritability of traits fitting common spouse environment as an additional random effect are also estimated and discussed. Although common spouse environment is not fitted in subsequent analyses, this analysis was performed to interrogate how common spouse environment altered heritability estimates, as this random effect substantially improved model fit for some traits (see Section 2.2.1.1.4.).

4.2.1.4.2. Using GWAA summary statistics

Heritability estimates for candidate endophenotypes on the observed scale using GWAA summary statistics were determined by regressing the χ^2 value for each variant onto its LD Score, using LD Score (v1.0.0.). This analysis used the same pre-computed LD Scores and data harmonizing methods as described in Section 3.2.2.2. The deviation of the LD score intercept from 1 against its standard error, and the LD Score ratio were interrogated to assess the contributions of confounding versus polygenicity¹⁶⁵.

4.2.1.3. Genetic and phenotypic correlation

Bivariate genetic analyses of endophenotypes and MDD were conducted in ASReml-R¹⁷³, as described in Section 2.2.3. The full model (which assumes additive covariance between traits) did not converge for three traits - DSC, EPQE and VF. Looking at the variance components from these analyses (Appendix 4.B.1, 4.B.2., and 4.B.10.), it appears to be fitting common sibling environment that is the culprit. This is consistent with results from previous versions of this work which only fitted an additive effect and a common household effect, and did not demonstrate convergence issues for any trait pair. As fitting common sibling environment improved model fit for several endophenotypes, but not for MDD (see Section 2.2.1.1.4. and Appendix 2), a conditional model was used for problematic trait pairs. The conditional model only fits common sibling environment for the quantitative trait, therefore assuming no sibling covariance between traits or sibling variance for MDD, whilst still fitting additive genetic component and common household effect for both traits. Although this approach resolved convergence issues, it is worth noting that this modeling method could alter genetic and phenotypic correlation estimates in the presence unmodelled sibling effects, as the additive or household variance components could become inflated due to absorbing genuine sibling effects. However, looking at estimates of sibling covariance and variance in MDD in bivariate models which did converge, any discrepancy is likely to be negligible.

4.2.2. Ranking candidate endophenotypes

4.2.2.1. Using coheritability

The coheritability between candidate endophenotypes and MDD assesses the covariation between two traits, weighted by their heritability, and evaluates the ability of quantitative traits to predict disease liability. Coheritability (h_{ij}) between traits and MDD was calculated using $h_{ij} = |\sqrt{h_i^2}\sqrt{h_j^2}r_g|$ where h_i^2 is the estimated heritability of the quantitative trait, h_j^2 is the estimated heritability for MDD and r_g is the genetic correlation between trait and MDD.

Endophenotypes are estimated to be a better predictor of disease liability than the observed trait itself if the trait has a higher coheritability value than the heritability of the disease trait as, for MDD, $h_{ij} = |\sqrt{0.44} * \sqrt{0.44} * 1| = 0.44 = h^2$. Coheritability values were calculated using the heritability for MDD estimated on the liability scale, as per Section 3.3.2.1., consistent with work previously presented by Glahn *et al*¹²⁴ which introduces the Endophenotype Ranking Value (ERV). As the absolute value of the coheritability is being taken here, this is analogous to the ERV. Results from GWAS suggest a large number of variants with small effects underlie most common diseases. Therefore, the liability-threshold model is likely to be a reasonable description of the relationship between the unobserved disease liability and the observed disease status⁹⁶. However, arguably the observed heritability estimate for MDD is the relevant comparison, as disease genotype is being predicted from case control status and the observed heritability takes account of the loss of information. Therefore, coheritability estimates are also provided which incorporate the observed heritability estimate for MDD (h_{ij0}). The observed heritability for MDD was calculated in ASReml-R, fitting MDD as the dependent variable; sex, age and age² as fixed effects; and an additive genetic component, common household environment and common sibling environment as random effects.

4.2.2.2. Using variance explained by MDD PGRS

Polygenic risk scores for MDD were created as described in Section 2.2.6. Traits which had a significant association with MDD PGRS surviving a conservative Bonferroni correction ($P < 1.47E-04$) were proposed to be candidate endophenotypes for MDD. Variance explained by PGRS was used to rank traits. As previously described, variance in trait value explained by polygenic profile score was estimated by multiplying the profile score by its corresponding regression coefficient and estimating its variance. This value was then divided by the variance of the observed phenotype to yield a coefficient of determination between 0 and 1¹⁸⁶, and multiplied by 100 to convert R^2 to percentage variance explained.

4.2.2.3. Convergence of ranking methods

The convergence of endophenotype ranking methods was assessed by the regressing variance explained by PGRS for MDD onto coheritability estimates on the observed scale for each candidate endophenotype, and establishing whether the correlation between ranking methods was statistically significant.

4.2.3. Performance of endophenotypes relative to binary MDD

4.2.3.1. Using variance explained by PGRS

The proportion of variance explained by MDD PGRS in the highest ranking endophenotypes were compared to the variance explained by MDD PGRS for binary depression as a method of determining whether quantitative traits had greater statistical power. This was repeated for other psychiatric disorders to explore differential enrichment of associated alleles, and to establish whether candidate endophenotypes are specific to depression, or more representative of psychiatric disorders in general.

4.2.3.2. Genome-wide association analysis

Genome-wide association analysis was conducted as described in Sections and 2.2.5. and 3.2.3. However, unlike the analysis described for binary traits in Section 3.2.3., test statistics from GWAA of quantitative traits were not transformed to the liability scale post-hoc. This is because the mixed linear model analysis (MLMA), implemented in GCTA (v1.25.)^{166,167}, employs restricted maximum likelihood methods and the regression is conducted on the linear scale. The highest ranking endophenotypes were considered to have greater power to detect genetic effects relative to the binary definition of MDD if they yielded an increased number of test statistics with more extreme associated p-values ($P < E-04$). The top 20 SNPs for each endophenotype were subjected to further bioinformatic analysis, as described in Section 3.3.3.5. utilizing SNAP (SNP Annotation and Proxy Search v2.2.)²⁰⁰, UCSC genome browser²⁰¹, the GeneCards database and PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>). Only putative genes tagged by query or proxy SNPs which explicitly relate to

psychiatric phenotypes (and therefore not clinical covariates) are discussed in this Chapter.

4.2.4. Replication in UKBiobank

How well the highest ranking endophenotypes in GS:SFHS correlated with the MDD phenotype in UKBiobank (UKB) was determined by a paired-samples sign test and bivariate LD score regression¹²³.

ORs of SNPs with a p-value $<1e-04$ from GWAS of EPQN, GHQ, MDQ, and SPQ in GS:SFHS were extracted from GWAS summary statistics of broad, narrow, and strict MDD definitions in UKB. These SNPs were subjected to a paired-samples sign test using the `binom.test` function in R (v3.2.3)¹⁵⁵ to assess whether the ORs from UKB were in the same direction as in GS:SFHS more or less frequently than expected by chance ($p=0.05$).

The genetic correlation between pairs of GWAS summary statistics was estimated using the `rg` function in LD Score (v1.0.0.)¹²³. This was performed on the observed scale as this analysis is estimating the correlation between pairs of quantitative and binary traits, and the genetic correlation on the observed scale is equivalent to that on the liability scale due to an approximate linear relationship existing between the random additive genetic effects on different scales¹⁸². Therefore population prevalence estimates were not supplied for this analysis.

4.3. Results

Summary statistics for mood, personality and cognitive variables are presented in Table 4.3. These statistics reflect the full genotyped sample, without considering the effect of depression status. This information is provided as, for analyses which are not utilizing the subset stratified by affection status such as GWAS and PGRS, this is the phenotypic data being used.

Phenotype	N	Max. Score	Range	Mean (sd)
DSC	19,765	133	0-133	72.2 (17.2)
EPQE	19,866	12	0-12	7.8 (3.5)
EPQN	19,874	12	0-12	3.9 (3.2)
GHQ	19,693	84	0-84	16.0 (8.9)
LM1	19,855	25	0-25	16.1 (3.9)
LM2D	19,751	25	0-25	14.9 (4.3)
MDQ	10,357	13	0-13	2.5 (3.2)
MHV	19,646	44	0-44	30.1 (4.7)
SPQ	10,845	22	0-22	3.9 (3.7)
VF	19,760	-	0-97	39.7 (11.7)

Table 4.3. Summary statistics for mood, personality and cognitive traits. Phenotype: DSC - digit symbol coding, EPQE - extraversion, EPQN - neuroticism, GHQ - general health questionnaire likert scale, LM1 - logical memory immediate, LM2D - logical memory delayed, MDQ - mood disorder questionnaire, MHV - mill hill vocabulary, SPQ - schizotypal personality questionnaire, VF - verbal fluency; N: number of individuals with data for each phenotype; Max. Score: maximum theoretical score for each phenotype¹⁴⁰; Range: range of scores obtained for each phenotype; Mean (sd): mean and standard deviation for each phenotype.

4.3.1. Testing endophenotype criteria

4.3.1.1. Effect of affection status on trait mean

Six out of ten traits demonstrated a significant effect of affection status on mean trait score: EPQN ($F_{(2,11170)}=694.8$, $p=3.85e-285$), GHQ ($F_{(2,10771)}=619.4$, $p=2.47e-255$), MDQ ($F_{(2,6305)}=183.9$, $p=2.33e-78$), SPQ ($F_{(2,6705)}=172.6$, $p=7.69e-74$), EPQE ($F_{(2,11355)}=55.5$, $p=1.01e-24$), and DSC ($F_{(2,11680)}=52.9$, $p=1.38e-23$). There was not a significant effect of affection status on mean trait scores for LM1

($F_{(2,11409)}=1.7$, $p=0.18$), LM2D ($F_{(2,11311)}=2.4$, $p=0.09$), MHV ($F_{(2,12053)}=4.6$, $p=0.01$) or VF ($F_{(2,11772)}=1.4$, $p=0.25$). Regression coefficients for fixed effects were used to assess whether the group effect was being driven by both unaffected relatives and cases. EPQN, GHQ, SPQ, DSC and MDQ were significantly more extreme in both unaffected relatives and cases, although arguably, MDQ showed only borderline significance in unaffected relatives. EPQE was being driven only by cases, and MHV only by unaffected controls. Differences in group mean and coefficients for fixed effects are presented in Table 4.3.1.1. Z-scores whose corresponding p-values survived multiple testing correction are highlighted in bold.

Phenotype	Summary Statistic	Group		
		Unrelated Controls	Unaffected Relatives	Cases
DSC	N	11757	2079	2638
	mean (sd)	72.63 (17.18)	72.0 (17.53)	71.56 (16.2)
	$z\beta$	-	-0.07 (0.02)	-0.20 (0.02)
	Z-score	-	-3.39	-10.28
EPQE	N	11818	2089	2653
	mean (sd)	7.92 (3.47)	7.91 (3.43)	7.18 (3.43)
	$z\beta$	-	-0.03 (0.03)	-0.23 (0.02)
	Z-score	-	-1.08	-10.32
EPQN	N	11822	2092	2650
	mean (sd)	3.39 (2.92)	3.71 (3.05)	6.45 (3.32)
	$z\beta$	-	0.09 (0.02)	0.78 (0.02)
	Z-score	-	3.99	36.74
GHQ	N	11717	2081	2617
	mean (sd)	14.81 (7.53)	15.56 (8.08)	22.71 (12.68)
	$z\beta$	-	0.09 (0.02)	0.76 (0.02)
	Z-score	-	3.91	34.85
LM1	N	11806	2085	2655
	mean (sd)	16.08 (3.92)	16.11 (3.93)	16.20 (3.45)
	$z\beta$	-	-8.30e-3 (0.02)	-0.04 (0.02)
	Z-score	-	-0.34	-1.85
LM2D	N	11745	2069	2644
	mean (sd)	14.9 (4.35)	15.04 (4.3)	15.05 (4.15)
	$z\beta$	-	0.01 (0.02)	-0.04 (0.02)
	Z-score	-	0.46	-1.87

MDQ	N	6515	955	1305
	mean (sd)	2.23 (3.07)	2.53 (3.22)	3.9 (3.56)
	$z\beta$	-	0.10 (0.03)	0.57 (0.03)
	Z-score	-	2.83	19.08
MHV	N	11671	2066	2635
	mean (sd)	30.21 (4.7)	29.86 (4.82)	30.17 (4.8)
	$z\beta$	-	-0.05 (0.02)	0.02 (0.02)
	Z-score	-	-2.11	0.83
SPQ	N	6787	1009	1401
	mean (sd)	3.49 (3.46)	3.88 (3.55)	5.98 (4.46)
	$z\beta$	-	0.12 (0.04)	0.56 (0.03)
	Z-score	-	3.51	18.47
VF	N	11750	2073	2647
	mean (sd)	39.79 (11.68)	39.4 (11.85)	40.4 (11.9)
	$z\beta$	-	-0.01 (0.03)	0.03 (0.02)
	Z-score	-	-0.54	1.26

Table 4.3.1.1. Effect of group (case, unaffected relative, unrelated control) on trait mean. N: number of individuals in each subgroup with phenotypic information for that trait; mean (sd): trait mean and standard deviation. $z\beta$: standardized regression coefficient for each affection status group; Z-score: Z-score for each affection status group. Where the corresponding p-value survived multiple testing correction, the Z-score is highlighted in bold.

These results demonstrate that of the ten traits tested, trait values for DSC, EPQN, GHQ, MDQ and SPQ are associated with illness in the general population, more extreme in unaffected relatives than in the general population and therefore state independent. EPQE was negatively associated with illness in the population but did not fulfill the other two endophenotype criteria tested in this Section.

4.3.1.2. Heritability of endophenotypes

4.3.1.2.1. Using pedigree data

All traits were moderately heritable, with heritability estimates ranging from $h^2(se) = 0.43(0.03)$ for DSC to $h^2(se) = 0.11(0.03)$ for the GHQ. Traits with the highest heritability were measures of cognitive ability, followed by personality traits and measures of mood. The observed heritability of MDD was estimated as $h^2(se)=0.17(0.02)$ which, as would be expected, is markedly lower than the

estimate on the liability scale of $h^2(se) = 0.44(0.04)^{142}$. Heritability estimates using pedigree-data are shown in Table 4.3.1.2.1.A.

Phenotype	h^2 (se)
MDD (liability) ¹⁴²	0.44 (0.04)
DSC	0.43 (0.03)
VF	0.39 (0.03)
MHV	0.37 (0.03)
LM1	0.32 (0.03)
LM2D	0.29 (0.02)
EPQE	0.28 (0.03)
EPQN	0.24 (0.03)
MDQ	0.21 (0.05)
SPQ	0.19 (0.04)
MDD (observed)	0.17 (0.02)
GHQ	0.11 (0.03)

Table 4.3.1.2.1.A. Narrow-sense heritability (h^2) estimates for MDD, mood, personality and cognitive traits ordered from highest h^2 to lowest, with standard errors for point estimates in parentheses.

The heritability of each trait was then estimated fitting common spouse environment as an additional random effect. These estimates, and the change in estimate caused by adding spouse, are shown in Table 4.3.1.2.1.B. Adding a spouse effect did not decrease the estimated heritability for any trait, but actually increased heritability estimates for the majority of traits (with the exception of EPQN). For this to make sense, the components being fitted (additive genetic component, common household environment, common sib environment and common spouse environment), and their potential contribution to parameter estimates must be considered. The effect of common spouse environment only contributes to the correlation between spouses. The only other parameter to do this is the household effect. If a spousal correlation is present but unmodeled, as in the estimates above, the household estimate is likely to increase to explain the spousal correlation. Therefore, if a spouse effect is added to a model already containing a household effect and spouse explains some of the spouse correlation, this can decrease the household effect. This, in turn, can cause the additive genetic component to increase to explain sibling

and parent-offspring correlations that were previously attributable to the household effect. Variance components for both heritability estimates are provided in Appendix 4.A. Using DSC as an example, it can be seen that in the model which does not fit spouse a household effect is present. However, when the spouse effect is added, the household effect becomes essentially zero and the genetic effect increases to explain the correlation between parents and offspring and between siblings. Therefore, modeling common spouse environment does affect heritability estimates, but estimates are not being upwardly biased by not including spouse. As such, not including a common spousal environment in subsequent analyses, given that spousal correlations cannot be teased apart from assortative mating, is satisfactory in this instance.

Phenotype	h^2 (se); increase/decrease
DSC	0.49 (0.03); ↑0.07
VF	0.49 (0.03); ↑0.10
MHV	0.62 (0.03); ↑0.25
LM1	0.34 (0.02); ↑0.02
LM2D	0.30 (0.03); ↑0.01
EPQE	0.31 (0.03); ↑0.03
EPQN	0.24 (0.03); -
MDQ	0.30 (0.05); ↑0.09
SPQ	0.33 (0.04); ↑0.14
GHQ	0.15 (0.04); ↑0.04

Table 4.3.1.2.1.B. Narrow-sense heritability (h^2) estimates for MDD, mood, personality and cognitive traits, including common spouse environment as a random effect. The order of traits is retained from Table 4.3.1.2.1.A., with the increase in heritability estimate indicated after the semi-colon.

4.3.1.2.2. Using GWAA summary statistics

Most traits had SNP-based h^2 estimates significantly different from zero, with the exception of MDQ, $h^2(se) = 0.03(0.05)$. Again, the traits with the highest heritabilities tended to be the cognitive traits, with estimates ranging from $h^2(se) = 0.24(0.04)$ for MHV to $0.10(0.03)$ for LM2D, followed by personality traits and mood traits. The ratio of the LD Score intercept to its standard error indicated that no trait had a significant intercept (<1.96), suggesting that the mixed linear model used for these analyses is appropriately controlling for

confounding due to population structure and the enrichment of families in GS:SFHS. However, the LD Score Ratio, which indicates the proportion of inflation ascribed to causes other than polygenic heritability, yielded a range of non-zero results suggesting potential confounding. In particular, the Ratio (SE) was highest for MDQ, 0.44 (0.81), and GHQ, 0.39 (0.37). The ratio is calculated using, $\text{Ratio} = (\text{Intercept} - 1)/(\mu\chi^2 - 1)$. Results from LD Score regression are shown in Table 4.3.1.2.2. As in Chapter 3, it would be interesting to repeat this analysis with imputed data to see how much these results are being driven by the subset of SNPs which overlaps between GS:SFHS and the pre-computed LD Scores¹⁹⁹, and whether a larger overlapping SNP set would ameliorate this issue. Whilst the mixed linear model will not be altered on the basis of these results, the ratio should be considered when interpreting subsequent results in this Chapter.

Trait	h² (SE)	λ_{GC}	μχ²	Intercept (SE)	Ratio (SE)
MHV	0.24 (0.04)	1.1	1.1	1.018 (0.009)	0.18 (0.09)
DSC	0.21 (0.03)	1.09	1.09	1.013 (0.008)	0.16 (0.10)
VF	0.21 (0.03)	1.06	1.07	0.996 (0.009)	<0
LM1	0.12 (0.03)	1.05	1.05	1.004 (0.007)	0.09 (0.14)
EPQE	0.11 (0.03)	1.04	1.04	1.008 (0.008)	0.17 (0.17)
LM2D	0.10 (0.03)	1.04	1.03	1.002 (0.007)	0.05 (0.21)
EPQN	0.09 (0.03)	1.04	1.04	1.005 (0.008)	0.14 (0.21)
SPQ	0.09 (0.05)	1.01	1.01	0.992 (0.007)	<0
GHQ	0.04 (0.03)	1.02	1.02	1.008 (0.008)	0.39 (0.37)
MDQ	0.03 (0.05)	1.008	1.01	1.004 (0.007)	0.44 (0.81)

Table 4.3.1.2.2. Output from univariate LD Score regression using summary statistics from GWAS of quantitative traits in GS:SFHS. h² (SE): SNP-based heritability estimate on the liability scale with its standard error; λ_{GC}: genomic control, estimated using median(chi²)/0.4549; μχ²: mean chi-squared statistic; Intercept (SE): LD Score regression intercept with its standard error; Ratio (SE): proportion of inflation ascribed to causes other than polygenic heritability with its standard error.

4.3.1.3. Genetic and phenotypic correlation

Four traits had a strong ($r_g > 0.5$) and highly significant ($p < 1E-10$) genetic correlation with MDD - MDQ, GHQ, SPQ, and EPQN. Whilst all of these traits also had a moderate ($r_p > 0.19$) phenotypic correlation with MDD, this was only statistically significant for two of the traits - GHQ and EPQN. Two traits have a moderate genetic correlation with MDD, DSC (-0.24) and EPQE (-0.16). Only the correlation with DSC survived multiple testing correction ($P < 3.85E-03$). Both traits had low, non-significant phenotypic correlations. The remaining four traits, LM1, LM2D, MHV and VF, had low, non-significant genetic and phenotypic correlations with MDD. Results are shown in Table 4.3.1.3.

Phenotype	r_g (se)	χ^2	p-value	r_p (se)	p-value
MDQ	0.92 (0.18)	36.7	<1E-10	0.20 (0.01)	0.06
GHQ	0.84 (0.16)	70.6	<1E-10	0.25 (0.007)	0.0007
SPQ	0.71 (0.17)	38.2	<1E-10	0.19 (0.009)	0.07
EPQN	0.54 (0.10)	53.9	<1E-10	0.26 (0.007)	0.0004
DSC	-0.24 (0.07)	14.8	5.93E-05	-0.07 (0.007)	0.20
EPQE	-0.16 (0.08)	5.7	0.008	-0.08 (0.007)	0.15
LM1	-0.05 (0.10)	0.8	0.1	-0.01 (0.007)	0.86
VF	-0.05 (0.07)	0.4	0.26	0.01 (0.007)	0.79
LM2D	-0.04 (0.10)	0.24	0.25	-0.01 (0.007)	0.84
MHV	-0.03 (0.08)	0.25	0.24	0.01 (0.007)	0.83

Table 4.3.1.3. Genetic and phenotypic correlations of quantitative traits with MDD. r_g (se): genetic correlation with MDD and corresponding standard error, χ^2 : chi-squared statistic (from a mixed $0.5(\chi^2) + 0.5(0)$ distribution) corresponding to the LRT between full and null (assuming no additive covariance) models; p-value: p value for r_g ; r_p (se): phenotypic correlation with MDD and corresponding standard error; p-value: p value for r_p .

These results demonstrate that of the ten traits tested, five traits fulfilled the endophenotype criteria of being significantly genetically correlated with MDD - MDQ, GHQ, SPQ, EPQN, and DSC. Only two traits fulfilled the criteria of being significantly phenotypically correlated with MDD - GHQ and EPQN.

4.3.2. Ranking candidate endophenotypes

4.3.2.1. Using coheritability

Traits with the highest coheritabilities (h_{ij}) are those with the strongest r_g . These were: MDQ, SPQ, EPQN and GHQ. DSC and EPQE had moderate h_{ij} values, although as the p-values for coheritability estimates are calculated in the same manner as for the bivariate analysis, again only DSC survived multiple testing correction. All other traits have low, non-significant h_{ij} values. Using the observed heritability estimate (h^2_{obs}) of MDD in the coheritability calculation (h_{ijo}) did not change the ranking of traits, but the h_{ijo} estimates are much closer in value to the observed heritability estimate for MDD, $h^2_{obs}(se) = 0.17(0.02)$, with statistically significant observed coheritabilities ranging from $h_{ijo} = 0.15$ for MDQ to $h_{ijo} = 0.06$ for DSC. This suggests that top ranking endophenotypes are almost as good a predictor of disease liability as a diagnosis of depression. Coheritability estimates are shown in Table 4.3.2.1.

Phenotype	h_{ij}	h_{ijo}	P-value
MDD	0.44	0.17	-
MDQ	0.25	0.15	<1E-10
SPQ	0.19	0.12	<1E-10
EPQN	0.18	0.11	<1E-10
GHQ	0.15	0.09	<1E-10
DSC	0.10	0.06	5.93E-05
EPQE	0.06	0.03	0.008
LM1	0.03	0.02	0.1
VF	0.02	0.01	0.26
LM2D	0.02	0.01	0.25
MHV	0.02	0.01	0.24

Table 4.3.2.1. Coheritability estimates for quantitative traits and MDD using h^2 estimates for MDD on the liability scale (h_{ij}) and observed scale (h_{ijo}). Traits where p-values survived multiple testing correction ($P \leq 0.05/26 = 1.92E-03$) are highlighted in bold.

4.3.2.2. Using variance explained by MDD PGRS

It was proposed that candidate endophenotypes should be significantly associated with PGRS for MDD, and that variance explained by MDD PGRS could be used as a ranking method for endophenotypes. PGRS for MDD were positively associated with EPQN, GHQ, MDQ and SPQ, and negatively associated with DSC, EPQE, LM1, LM2D, and VF in GS:SFHS. For MHV, both positive and negative associations were observed at different P_T . As $P_T \leq 0.5$ explained the most variance in the discovery dataset⁴⁷, this is the P_T which will be summarized here, and used to rank candidate endophenotypes. At this P_T , the variance explained ranged from 0.19% for GHQ to 0.004% for EPQE. Only associations for EPQN, GHQ and LM2D survived multiple testing correction at $P_T \leq 0.5$, ($p = 3.22E-08$, $p = 9.85E-09$ and $p = 8.26E-05$, respectively), with the strongest association being between MDD PGRS and GHQ. Traits ranked by variance explained at the $P_T \leq 0.5$ are shown in Table 4.3.2.2.A.

Trait	% variance explained	P-value
GHQ	0.19	9.85E-09
EPQN	0.17	3.22E-08
LM2D	0.09	7.03E-05
MDQ	0.07	0.01
LM1	0.06	0.002
DSC	0.04	0.004
SPQ	0.03	0.09
MHV	0.01	0.11
EPQE	0.004	0.4
VF	0.003	0.51

Table 4.3.2.2.A. Traits ranked by variance explained by MDD PGRS at $P_T \leq 0.5$. P-values which survived multiple testing correction and highlighted in bold.

Interestingly, aside from GHQ and EPQN, other previously top ranking traits from pedigree-based analyses (MDQ, SPQ and DSC) are not high ranking when using variance explained by MDD PGRS, nor are they significantly associated. Conversely, LM2D did not fulfill any endophenotype criteria (other than being heritable) and ranked 9th when using estimates of coheritability, yet it is significantly associated with MDD PGRS. One explanation for these outliers

could be the difference in what each method of assessing covariance is measuring. Polygenic risk scores were derived using only common variants that were on the array, as unimputed data was used here. Coheritability, on the other hand, was calculated from variance components generated from pedigree data, and thus incorporates common and rare genetic variants in addition to epistasis and gene by environment (GxE) interactions. Therefore it could be postulated that the covariance of traits which rank highly when using coheritability estimates, but not variance explained by PGRS, is being mediated by rare variants, epistasis or GxE interaction not captured by the PGRS. Conversely, these factors perhaps had an opposing effect on LM2D. Results from all P_T are presented in Table 4.3.2.2.B with variance explained by PGRS in Figure 4.3.2.2.

Trait	PGRS P_T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
DSC	1	-0.018 (0.006)	-2.75	7.6 (1, 17135)	0.006	0.03
	0.5	-0.019 (0.006)	-2.89	8.4 (1, 17103)	0.004	0.04
	0.1	-0.018 (0.006)	-2.83	8.0 (1, 17212)	0.005	0.03
	0.05	-0.021 (0.006)	-3.21	10.3 (1, 17338)	0.001	0.04
	0.01	-0.026 (0.006)	-4.00	16.0 (1, 17222)	6.37E-05	0.07
EPQE	1	-0.007 (0.008)	-0.91	0.8 (1, 15886)	0.37	0.005
	0.5	-0.006 (0.008)	-0.84	0.7 (1, 15852)	0.40	0.004
	0.1	-0.001 (0.008)	-0.18	0.03 (1, 16003)	0.86	1.82E-04
	0.05	-0.008 (0.008)	-1.03	1.1 (1, 16193)	0.3	0.006
	0.01	-0.004 (0.007)	-0.50	0.3 (1, 16061)	0.61	0.001
EPQN	1	0.043 (0.007)	5.78	33.4 (1, 15757)	7.90E-09	0.18
	0.5	0.041 (0.007)	5.53	30.6 (1, 15721)	3.22E-08	0.17

	0.1	0.034 (0.007)	4.58	21.0 (1, 15886)	4.59E-06	0.12
	0.05	0.036 (0.007)	4.82	23.2 (1, 16078)	1.48E-06	0.13
	0.01	0.025 (0.007)	3.37	11.4 (1, 15936)	7.41E-04	0.06
GHQ	1	0.044 (0.008)	5.84	34.1 (1, 14865)	5.23E-09	0.19
	0.5	0.043 (0.008)	5.74	32.9 (1, 14840)	9.85E-09	0.19
	0.1	0.034 (0.008)	4.54	20.6 (1, 15068)	5.77E-06	0.12
	0.05	0.033 (0.008)	4.35	18.9 (1, 15266)	1.37E-05	0.11
	0.01	0.027 (0.007)	3.58	12.8 (1, 15120)	3.49E-04	0.07
LM1	1	-0.024 (0.007)	-3.21	10.3 (1, 16429)	0.001	0.06
	0.5	-0.024 (0.007)	-3.17	10.0 (1, 16393)	0.002	0.06
	0.1	-0.019 (0.007)	-2.58	6.6 (1, 16519)	0.01	0.04
	0.05	-0.022 (0.007)	-2.92	8.5 (1, 16678)	0.003	0.05
	0.01	-0.020 (0.007)	-2.69	7.2 (1, 16566)	0.007	0.04
LM2D	1	-0.030 (0.007)	-3.99	15.9 (1, 16094)	6.61E-05	0.09
	0.5	-0.029 (0.007)	-3.98	15.8 (1, 16061)	7.03E-05	0.09
	0.1	-0.020 (0.007)	-2.67	7.1 (1, 16186)	0.008	0.04
	0.05	-0.020 (0.007)	-2.74	7.5 (1, 16349)	0.006	0.04
	0.01	-0.018 (0.007)	-2.38	5.7 (1, 16238)	0.02	0.03
MDQ	1	0.024 (0.010)	2.38	5.7 (1, 8575)	0.02	0.06
	0.5	0.026 (0.010)	2.56	6.6 (1, 8542)	0.01	0.07
	0.1	0.029 (0.010)	2.79	7.8 (1, 8577)	0.005	0.08
	0.05	0.032 (0.010)	3.14	9.9 (1, 8618)	0.002	0.10
	0.01	0.026 (0.010)	2.57	6.6 (1, 8621)	0.01	0.07
MHV	1	0.011	-1.51	2.3	0.13	0.01

		(0.007		(1, 18110)		
	0.5	-0.011 (0.007)	-1.60	2.6 (1, 18094)	0.11	0.01
	0.1	-0.009 (0.007)	-1.29	1.7 (1, 18182)	0.20	0.008
	0.05	0.010 (0.007)	-1.40	1.9 (1, 18272)	0.16	0.01
	0.01	-0.006 (0.007)	-0.87	0.8 (1, 18196)	0.38	0.004
SPQ	1	0.018 (0.010)	1.81	3.3 (1, 9244)	0.07	0.03
	0.5	0.017 (0.010)	1.70	2.9 (1, 9200)	0.09	0.03
	0.1	0.013 (0.010)	1.31	1.7 (1, 9220)	0.19	0.02
	0.05	0.022 (0.010)	2.13	4.6 (1, 9258)	0.03	0.05
	0.01	0.021 (0.010)	2.01	4.0 (1, 9201)	0.04	0.04
VF	1	-0.002 (0.008)	-0.29	0.09 (1, 17445)	0.77	5.02E-04
	0.5	-0.005 (0.008)	-0.66	0.4 (1, 17421)	0.51	0.003
	0.1	-0.005 (0.008)	-0.72	0.5 (1, 17535)	0.47	0.003
	0.05	-0.006 (0.008)	-0.82	0.7 (1, 17639)	0.41	0.004
	0.01	-0.003 (0.008)	-0.33	0.1 (1, 17560)	0.74	6.35E-04

Table 4.3.2.2.B. Association with and variance in trait score explained by polygenic risk score for major depressive disorder. P-values which survived multiple testing correction are highlighted in bold.

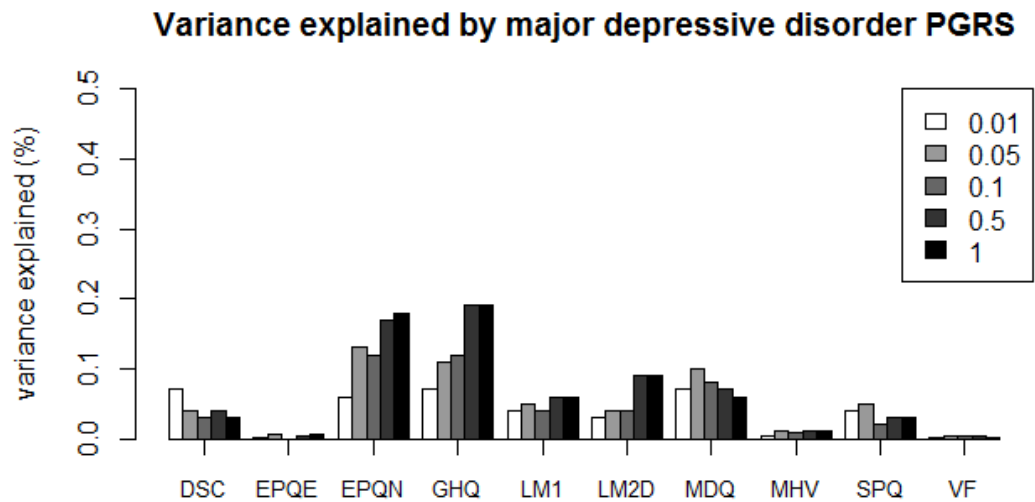


Figure 3.3.4.1. Percentage variance explained in candidate endophenotypes in GS:SFHS by MDD PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

4.3.2.3. Convergence of coheritability measures

The convergence of ranking methods was assessed by regressing variance explained by MDD PGRS onto the observed coheritability with MDD for each quantitative trait. The correlation between these ranking methods was weak (Adjusted $R^2 = 0.09$) and non-significant ($F_{(1,8)} = 1.9$; $p = 0.20$). The relationship is plotted in Figure 4.3.2.3.A.

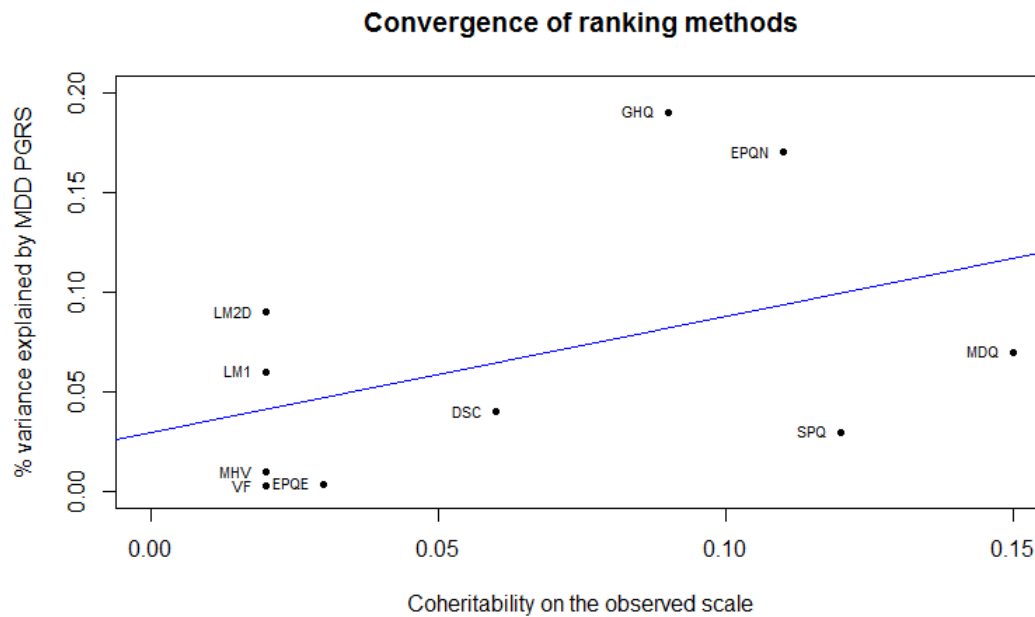


Figure 4.3.2.3.A. Linear relationship between variance explained by MDD PGRS and observed coheritability with MDD for selected mood, personality and cognitive traits, showing a lack of convergence between ranking methods.

To see whether the lack of convergence was being driven by traits with a markedly smaller sample size (MDQ: $n=10,357$; SPQ: $n=10,845$), the linear regression was re-run with these traits removed. This resulted in a significant improvement of model fit, demonstrating a moderately high (Adjusted $R^2 = 0.59$), statistically significant ($F_{(1,6)} = 11.0$; $p = 0.02$), correlation between ranking methods. The relationship is plotted in Figure 4.3.2.3.B.

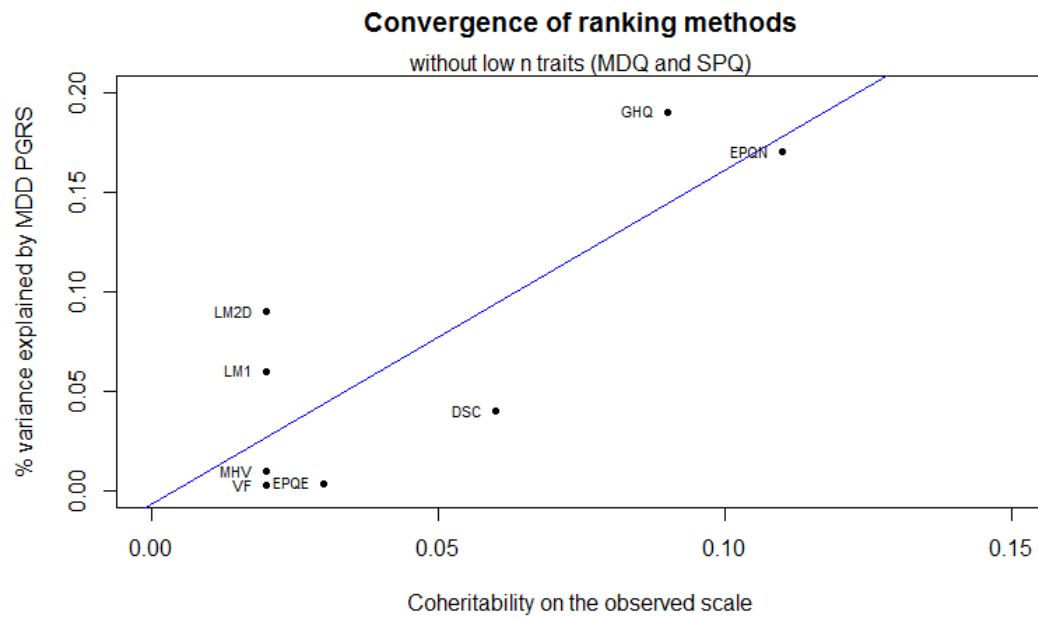


Figure 4.3.2.3.B. Linear relationship between variance explained by MDD PGRS and observed coheritability with MDD for measures of mood, personality and cognitive traits which had the highest number of observations ($n \geq 19,646$), showing improved, statistically significant, convergence between ranking methods.

Therefore, although it cannot be conclusively said that variance explained by disease PGRS is a robust means of ranking candidate endophenotypes, perhaps this method is appropriate in traits measured on more substantial sample sizes ($\sim 20,000$ individuals). It would be useful to repeat this comparison of ranking methodologies in a larger sample, such as UKB, to establish whether or not this approach is valid.

4.3.3. Performance of endophenotypes relative to binary MDD

4.3.3.1. Using variance explained by PGRS

The performance of endophenotypes relative to MDD as a binary trait was determined, in part, by observing whether more trait variance was explained by MDD PGRS in quantitative traits relative to binary MDD. To explore whether endophenotypes were specifically associated with MDD PGRS, or whether these traits were associated with psychiatric disorders in general, analyses were re-

run using PGRS for autism²⁰⁶, bipolar disorder⁵¹ and schizophrenia²³. In Section 3.3.4. and its subsections, it was observed that in undifferentiated MDD and sex-specific MDD, schizophrenia PGRS explained the most trait variance, followed by MDD PGRS, bipolar PGRS and autism PGRS. Only for recurrent MDD did MDD PGRS explain the most variance. Therefore, the ranking of which psychiatric disorders explain the most variance in candidate endophenotypes will be compared to the ranking for MDD to further assess potential genetic similarity.

4.3.3.1.1. Association with MDD PGRS

As stated in Section 4.3.2.2., only MDD PGRS associations for EPQN, GHQ and LM2D survived multiple testing correction. The variance explained by PGRS for these traits (EPQN = 0.17%, GHQ = 0.19% and LM2D = 0.09%) exceeded the variance explained in MDD as a binary trait (0.08%). For traits which did not survive multiple testing correction, variance explained did not exceed that of binary MDD, ranging from 0.004% in EPQE to 0.07% in MDQ.

4.3.3.1.2. Association with autism PGRS

PGRS for autism were positively associated with EPQN, LM1, LM2D, MHV and VF, and negatively associated with SPQ in GS:SFHS. For four traits, DSC, EPQE, GHQ and MDQ, both positive and negative associations were observed at different P_T . The marker weights for the autism PGRS were derived from the PGR Cross-Disorder Group analysis, which do not report which P_T explained the most variance in the discovery dataset. Therefore $P_T \leq 1$, which includes all SNPs, will be summarized here. At this P_T , the variance explained ranged from 0.15% for LM1 to 0.00008% for MDQ. Only associations for LM1, LM2D and MHV survived multiple testing correction at $P_T \leq 1$, ($p = 2.07E-07$, $p = 9.17E-05$ and $p = 8.72E-05$, respectively), with the strongest association being between autism PGRS and LM1. Variance explained for these traits was small (LM1, 0.15%; LM2D, 0.08%; MHV, 0.08%), but considerable higher than observed in binary MDD. These results are consistent with work performed on the original genotyped subset in GS:SFHS ($n=9,863$), replicated in two independent samples, which found that measures of cognitive ability associated positively with autism

PGRS²⁹⁷. Results from all P_T are presented in Table 4.3.3.1.2. with variance explained by PGRS visualized in Figure 4.3.3.1.2.

Trait	PGRS P _T	zβ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
DSC	1	-0.001 (0.006)	-0.20	0.04 (1, 17273)	0.85	1.56E-04
	0.5	-0.001 (0.006)	-0.17	0.03 (1, 17276)	0.87	1.15E-04
	0.1	-0.003 (0.006)	-0.51	0.3 (1, 17318)	0.61	0.001
	0.05	3.37E-04 (0.006)	0.06	0.003 (1, 17289)	0.95	1.39E-05
	0.01	-0.004 (0.006)	-0.64	0.4 (1, 17376)	0.52	0.002
EPQE	1	0.009 (0.007)	1.19	1.4 (1, 16057)	0.24	0.008
	0.5	0.009 (0.007)	1.21	1.5 (1, 16055)	0.23	0.008
	0.1	0.003 (0.007)	0.43	0.2 (1, 16106)	0.67	0.001
	0.05	-0.007 (0.008)	0.97	0.9 (1, 16066)	0.33	0.005
	0.01	-0.0005 (0.008)	-0.07	0.005 (1, 16219)	0.95	2.61E-05
EPQN	1	0.001 (0.007)	0.14	0.02 (1, 15944)	0.89	1.08E-04
	0.5	0.002 (0.007)	0.30	0.09 (1, 15944)	0.76	5.04E-04
	0.1	0.006 (0.007)	0.78	0.6 (1, 15989)	0.44	0.003
	0.05	0.008 (0.007)	1.19	1.4 (1, 15942)	0.23	0.008
	0.01	0.004 (0.007)	0.58	0.3 (1, 16100)	0.56	0.002
GHQ	1	-0.004 (0.007)	-0.48	0.2 (1, 15127)	0.63	0.001
	0.5	0.002 (0.007)	-0.27	0.07 (1, 15130)	0.79	4.12E-04
	0.1	0.005 (0.008)	0.66	0.4 (1, 15212)	0.51	0.002
	0.05	0.004 (0.008)	0.47	0.2 (1, 15120)	0.64	0.001
	0.01	-0.002 (0.008)	-0.25	0.06 (1, 15301)	0.80	3.61E-04

LM1	1	0.038 (0.007)	5.20	27.0 (1, 16577)	2.07E-07	0.15
	0.5	0.037 (0.007)	4.99	24.9 (1, 16578)	6.10E-07	0.14
	0.1	0.036 (0.007)	4.90	24.0 (1, 16619)	9.78E-07	0.13
	0.05	0.036 (0.007)	4.89	23.9 (1, 16614)	1.03E-06	0.13
	0.01	0.025 (0.007)	3.39	11.5 (1, 16725)	6.89E-04	0.06
LM2D	1	0.029 (0.007)	3.91	15.3 (1, 16267)	9.17E-05	0.08
	0.5	0.027 (0.007)	3.68	13.6 (1, 16270)	2.33E-04	0.07
	0.1	0.026 (0.007)	3.48	12.1 (1, 16316)	4.98E-04	0.07
	0.05	0.026 (0.007)	3.52	12.4 (1, 16300)	4.37E-04	0.07
	0.01	0.016 (0.007)	2.16	4.7 (1, 16409)	0.03	0.03
MDQ	1	-0.0009 (0.010)	-0.09	0.008 (1, 8616)	0.93	8.17E-05
	0.5	0.0004 (0.010)	0.04	0.002 (1, 8613)	0.96	2.01E-05
	0.1	0.002 (0.010)	0.22	0.05 (1, 8588)	0.83	4.97E-04
	0.05	0.003 (0.010)	0.32	0.1 (1, 8560)	0.75	0.001
	0.01	0.004 (0.010)	0.36	0.1 (1, 8540)	0.72	0.001
MHV	1	0.027 (0.007)	3.92	15.4 (1, 18199)	8.72E-05	0.08
	0.5	0.026 (0.007)	3.73	1.4 (1, 18202)	1.93E-04	0.07
	0.1	0.020 (0.007)	2.87	8.2 (1, 18270)	0.004	0.04
	0.05	0.016 (0.007)	2.31	5.3 (1, 18183)	0.02	0.03
	0.01	0.018 (0.007)	2.49	6.2 (1, 18278)	0.01	0.03
SPQ	1	-0.015 (0.010)	-1.52	2.3 (1, 9223)	0.13	0.02
	0.5	-0.017 (0.010)	-1.68	2.8 (1, 9226)	0.09	0.03
	0.1	-0.012 (0.010)	-1.16	1.3 (1, 9215)	0.25	0.01
	0.05	-0.013	-1.30	1.7	0.19	0.02

		(0.010)		(1, 9152)		
	0.01	-0.004 (0.010)	-0.40	1.6 (1, 9155)	0.69	0.002
VF	1	0.010 (0.008)	1.27	1.6 (1, 17579)	0.2	0.009
	0.5	0.009 (0.008)	1.15	1.3 (1, 17582)	0.25	0.007
	0.1	0.006 (0.008)	0.82	0.7 (1, 17631)	0.41	0.004
	0.05	0.006 (0.008)	0.83	0.7 (1, 17554)	0.83	0.004
	0.01	0.012 (0.008)	1.54	2.4 (1, 17679)	0.12	0.01

Table 4.3.3.1.2. Association with and variance in trait score explained by polygenic risk score for autism. P-values which survived multiple testing correction are highlighted in bold.

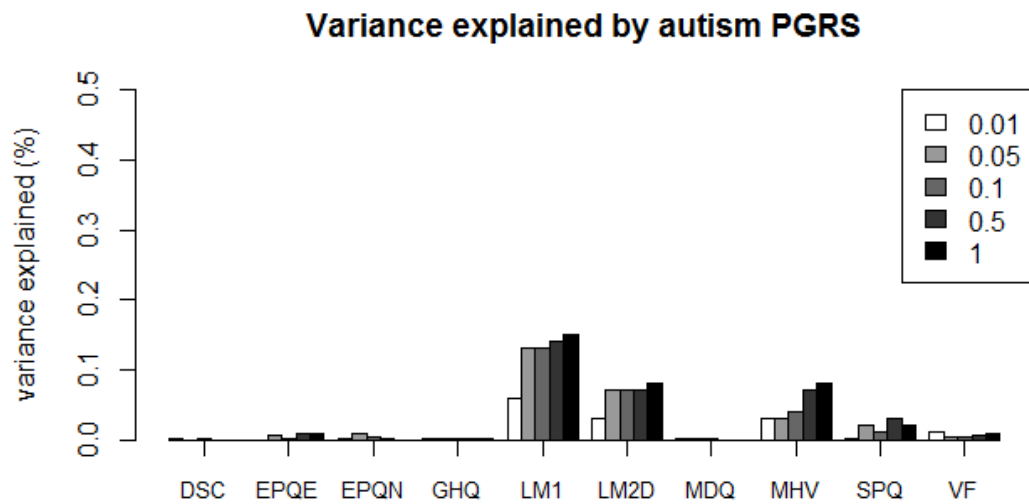


Figure 4.3.3.1.2. Percentage variance explained in candidate endophenotypes in GS:SFHS by autism PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

4.3.3.1.3. Association with bipolar disorder PGRS

PGRS for bipolar disorder were positively associated with EPQE, EPQN, GHQ, MDQ and MHV, and negatively associated with DSC, LM1, and LM2D in GS:SFHS.

For two traits, SPQ and VF, both positive and negative associations were observed at different P_T . As $P_T \leq 0.5$ explained the most variance in the discovery dataset⁵¹, this is the P_T which will be summarized here. At this P_T , the variance explained ranged from 0.08% for GHQ to 0.001% for SPQ. No associations survived multiple testing correction at $P_T \leq 0.5$. The only trait to survive multiple testing correction at any P_T was GHQ, which was significantly associated with bipolar disorder PGRS at $P_T \leq 0.1$, $P_T \leq 0.05$ and $P_T \leq 0.01$ ($p = 1.29E-05$, $p = 1.27E-04$ and $p = 4.79E-05$, respectively). At $P_T \leq 0.5$, variance explained by bipolar disorder PGRS for EPQN (0.05%), GHQ (0.08%) and LM2D (0.07%) exceeded the variance explained in MDD as a binary trait (0.04%). Results from all P_T are presented in Table 4.3.3.1.3. with variance explained by PGRS visualized in Figure 4.3.3.1.3.

Trait	PGRS P_T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
DSC	1	-0.019 (0.006)	-3.00	9.0 (1, 17330)	0.003	0.04
	0.5	-0.017 (0.006)	-2.72	7.4 (1, 17311)	0.007	0.03
	0.1	-0.017 (0.006)	-2.61	6.8 (1, 17413)	0.009	0.03
	0.05	-0.013 (0.006)	-2.01	4.1 (1, 17342)	0.04	0.02
	0.01	-0.009 (0.006)	-1.46	2.1 (1, 17199)	0.15	0.009
EPQE	1	0.019 (0.007)	2.50	6.3 (1, 16197)	0.01	0.03
	0.5	0.019 (0.007)	2.57	6.6 (1, 16174)	0.01	0.04
	0.1	0.014 (0.007)	1.94	3.8 (1, 16271)	0.05	0.02
	0.05	0.015 (0.007)	2.01	4.0 (1, 16197)	0.04	0.02
	0.01	0.006 (0.008)	0.86	0.7 (1, 16011)	0.39	0.004
EPQN	1	0.021 (0.007)	2.90	8.4 (1, 16086)	0.004	0.05
	0.5	0.022 (0.007)	2.96	8.8 (1, 16061)	0.003	0.05
	0.1	0.025	3.33	11.1	8.78E-04	0.06

		(0.007)		(1, 16149)		
	0.05	0.023 (0.007)	3.10	9.6 (1, 16076)	0.002	0.05
	0.01	0.025 (0.007)	3.38	11.5 (1, 15881)	7.16E-04	0.06
GHQ	1	0.027 (0.007)	3.57	12.8 (1, 15272)	3.52E-04	0.07
	0.5	0.028 (0.007)	3.73	13.9 (1, 15248)	1.94E-04	0.08
	0.1	0.033 (0.007)	4.36	19.0 (1, 15333)	1.29E-05	0.11
	0.05	0.029 (0.008)	3.83	14.7 (1, 15265)	1.27E-04	0.08
	0.01	0.031 (0.008)	4.07	16.5 (1, 15120)	4.79E-05	0.09
LM1	1	-0.020 (0.007)	-2.76	7.6 (1, 16708)	0.006	0.04
	0.5	-0.021 (0.007)	-2.79	7.8 (1, 16682)	0.005	0.04
	0.1	-0.020 (0.007)	-2.76	7.6 (1, 16760)	0.006	0.04
	0.05	-0.019 (0.007)	-2.55	6.5 (1, 16684)	0.01	0.04
	0.01	-0.023 (0.007)	-3.10	9.6 (1, 16197)	0.002	0.05
LM2D	1	-0.025 (0.007)	-3.44	11.8 (1, 16384)	5.85E-04	0.06
	0.5	-0.026 (0.007)	-3.57	12.7 (1, 16361)	3.61E-04	0.07
	0.1	-0.022 (0.007)	-3.05	9.3 (1, 16438)	0.002	0.05
	0.05	-0.020 (0.007)	-2.67	7.1 (1, 16359)	0.008	0.04
	0.01	-0.025 (0.007)	-3.13	11.0 (1, 16215)	9.24E-04	0.06
MDQ	1	0.020 (0.010)	1.94	3.8 (1, 8783)	0.05	0.04
	0.5	0.019 (0.010)	1.84	3.4 (1, 8789)	0.07	0.04
	0.1	0.025 (0.010)	2.41	5.8 (1, 8769)	0.02	0.06
	0.05	0.021 (0.010)	2.01	4.1 (1, 8737)	0.04	0.04
	0.01	0.013 (0.010)	1.25	1.6 (1, 8681)	0.21	0.02
MHV	1	0.006 (0.007)	0.82	0.7 (1, 18236)	0.41	0.003

	0.5	0.006 (0.007)	0.87	0.8 (1, 18238)	0.38	0.004
	0.1	0.006 (0.007)	0.87	0.8 (1, 18345)	0.38	0.004
	0.05	0.005 (0.007)	0.67	0.4 (1, 18339)	0.50	0.002
	0.01	0.005 (0.007)	0.66	0.4 (1, 18236)	0.51	0.002
SPQ	1	0.004 (0.010)	0.35	0.1 (1, 9370)	0.72	0.001
	0.5	0.003 (0.010)	0.33	0.1 (1, 9323)	0.74	0.001
	0.1	-0.004 (0.010)	-0.40	0.2 (1, 9398)	0.69	0.002
	0.05	-8.75E-05 (0.010)	-0.009	7.3E-05 (1, 9388)	0.99	7.70E-07
	0.01	-0.005 (0.010)	0.51	0.3 (1, 9314)	0.61	0.003
VF	1	0.006 (0.008)	0.79	0.6 (1, 17607)	0.43	0.004
	0.5	0.006 (0.008)	0.76	0.6 (1, 17604)	0.76	0.003
	0.1	0.007 (0.008)	0.93	0.9 (1, 17731)	0.35	0.005
	0.05	0.002 (0.008)	0.28	0.08 (1, 17680)	0.78	4.64E-04
	0.01	-0.005 (0.008)	-0.7	0.5 (1, 17539)	0.48	0.003

Table 4.3.3.1.3. Association with and variance in trait score explained by polygenic risk score for bipolar disorder. P-values which survived multiple testing correction are highlighted in bold.

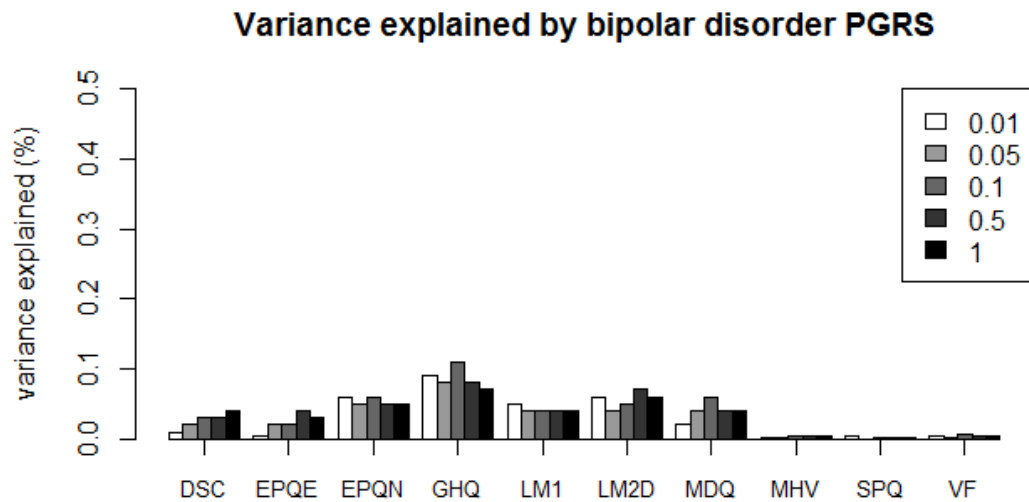


Figure 4.3.3.1.3. Percentage variance explained in candidate endophenotypes in GS:SFHS by bipolar disorder PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

4.3.3.1.4. Association with schizophrenia PGRS

PGRS for schizophrenia were positively associated with EPQN, GHQ, MDQ and SPQ, and negatively associated with DSC, LM1, LM2D, MHV and VF in GS:SFHS. For EPQE, both positive and negative associations were observed at different P_T . Associations at all P_T thresholds survived multiple testing correction for DSC, EPQN, GHQ, LM1 and LM2D. Associations with MDQ survived multiple testing correction at every P_T except $P_T \leq 0.01$. As $P_T \leq 0.1$ explained the most variance in the discovery dataset²³, this is the P_T which will be summarized here. At this P_T , the variance explained ranged from 0.39% for LM2D to 0.0009% for EPQE. Variance explained was highest in traits assessing verbal declarative memory, LM1 (0.38%) and LM2D (0.39%), and processing speed, DSC (0.36%), with these traits also having the most significant associations with schizophrenia PGRS, ($p = 3.37E-15$, $p = 1.40E-15$, and $p = 8.53E-18$, respectively). This is consistent with previous studies which identified that cognitive decline is associated with an increased polygenic load for schizophrenia⁸⁶. GHQ and EPQN, the most consistently high ranking endophenotypes, had variance explained by

schizophrenia PGRS (0.14% and 0.17% respectively), in line with binary depression (0.15%). Results from all P_T are presented in Table 4.3.3.1.4. with variance explained by PGRS visualized in Figure 4.3.3.1.4.

Trait	PGRS P_T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
DSC	1	-0.059 (0.007)	-8.78	77.1 (1, 17073)	1.81E-18	0.35
	0.5	-0.059 (0.007)	-8.75	76.5 (1, 17073)	2.42E-18	0.35
	0.1	-0.060 (0.007)	-8.86	78.6 (1, 17097)	8.53E-19	0.36
	0.05	-0.054 (0.007)	-7.97	63.5 (1, 17107)	1.73E-15	0.30
	0.01	-0.047 (0.007)	-6.68	44.6 (1, 17176)	2.44E-11	0.22
EPQE	1	0.006 (0.008)	0.82	0.7 (1, 15853)	0.41	0.004
	0.5	0.006 (0.008)	0.70	0.5 (1, 15840)	0.48	0.003
	0.1	0.003 (0.008)	0.37	0.1 (1, 15867)	0.71	8.76E-04
	0.05	0.004 (0.008)	0.48	0.2 (1, 15866)	0.63	0.001
	0.01	-0.004 (0.008)	-0.48	0.2 (1, 15936)	0.63	0.002
EPQN	1	0.037 (0.008)	4.83	23.4 (1, 15708)	1.36E-06	0.14
	0.5	0.038 (0.008)	4.84	23.4 (1, 15693)	1.31E-06	0.14
	0.1	0.041 (0.008)	5.24	27.5 (1, 15715)	1.59E-07	0.17
	0.05	0.042 (0.008)	5.30	28.1 (1, 15721)	1.16E-07	0.17
	0.01	0.039 (0.008)	4.79	22.9 (1, 15803)	1.69E-06	0.15
GHQ	1	0.040 (0.008)	5.05	25.5 (1, 14840)	4.43E-07	0.16
	0.5	0.040 (0.008)	5.09	25.9 (1, 14825)	3.61E-07	0.16
	0.1	0.037 (0.008)	4.74	22.4 (1, 14859)	2.19E-06	0.14
	0.05	0.034 (0.008)	4.32	18.7 (1, 14851)	1.57E-05	0.12
	0.01	0.042	5.11	26.2	3.19E-07	0.18

		(0.008)		(1, 14921)		
LM1	1	-0.060 (0.008)	-7.81	61.1 (1, 16367)	5.87E-15	0.37
	0.5	-0.0062 (0.008)	-8.04	64.6 (1, 16356)	9.86E-16	0.39
	0.1	-0.062 (0.008)	-7.88	62.2 (1, 16370)	3.37E-15	0.38
	0.05	-0.065 (0.008)	-8.21	67.4 (1, 16363)	2.39E-16	0.42
	0.01	-0.060 (0.008)	-7.38	54.5 (1, 16445)	1.67E-13	0.36
LM2D	1	-0.062 (0.008)	-8.02	64.3 (1, 16035)	1.16E-15	0.38
	0.5	-0.063 (0.008)	-8.19	67.0 (1, 16024)	2.86E-16	0.40
	0.1	-0.062 (0.008)	-7.99	63.9 (1, 16036)	1.40E-15	0.39
	0.05	-0.066 (0.008)	-8.39	70.4 (1, 16026)	5.28E-17	0.44
	0.01	-0.065 (0.008)	-8.00	64.0 (1, 16100)	1.36E-15	0.42
MDQ	1	0.045 (0.011)	4.25	18.1 (1, 8325)	2.14E-05	0.20
	0.5	0.043 (0.011)	4.07	16.6 (1, 8326)	4.79E-05	0.19
	0.1	0.041 (0.011)	3.85	14.8 (1, 8351)	1.18E-04	0.17
	0.05	0.044 (0.011)	4.04	16.3 (1, 8360)	5.41E-05	0.19
	0.01	0.033 (0.011)	2.94	8.6 (1, 8397)	0.003	0.11
MHV	1	-0.015 (0.007)	-1.99	4.0 (1, 18111)	0.35	0.02
	0.5	-0.015 (0.007)	-2.01	4.0 (1, 18118)	0.04	0.02
	0.1	-0.013 (0.007)	-1.70	2.9 (1, 18164)	0.09	0.02
	0.05	-0.009 (0.007)	-1.17	1.4 (1, 18153)	0.24	0.008
	0.01	-0.007 (0.008)	-0.93	0.9 (1, 18180)	0.35	0.005
SPQ	1	0.012 (0.011)	1.17	1.4 (1, 9010)	0.24	0.02
	0.5	0.012 (0.011)	1.16	1.3 (1, 9011)	0.25	0.02
	0.1	0.016 (0.011)	1.46	2.1 (1, 9034)	0.14	0.02

	0.05	0.013 (0.011)	1.21	1.5 (1, 9029)	0.22	0.02
	0.01	0.010 (0.011)	0.93	8.7 (1, 9094)	0.35	0.01
VF	1	-0.011 (0.008)	-1.34	1.8 (1, 17409)	0.18	0.01
	0.5	-0.012 (0.008)	-1.47	2.2 (1, 17412)	0.14	0.01
	0.1	-0.014 (0.008)	-1.77	3.1 (1, 17451)	0.08	0.02
	0.05	-0.011 (0.008)	-1.37	1.9 (1, 17443)	0.17	0.01
	0.01	-0.007 (0.008)	-0.84	0.7 (1, 17507)	0.40	0.005

Table 4.3.3.1.4. Association with and variance in trait score explained by polygenic risk score for schizophrenia. P-values which survived multiple testing correction are highlighted in bold.

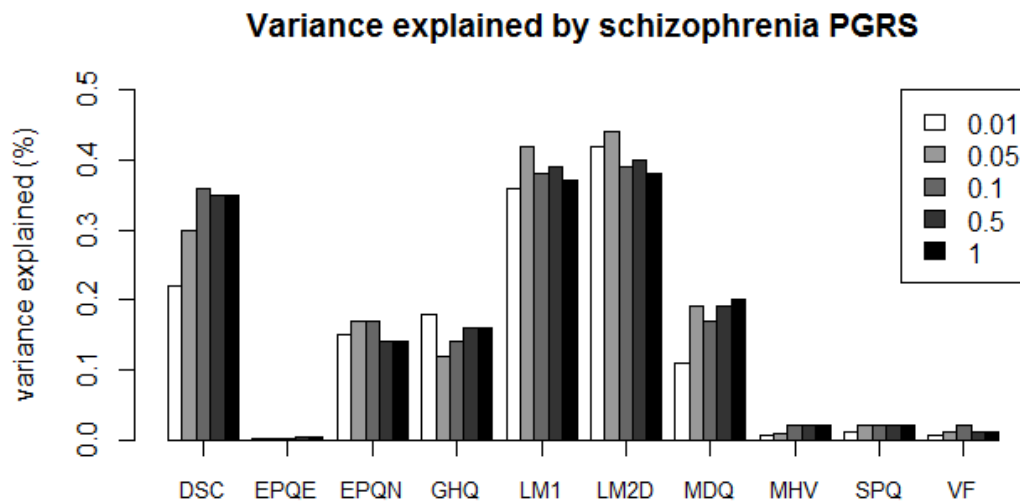


Figure 4.3.3.1.4. Percentage variance explained in candidate endophenotypes in GS:SFHS by schizophrenia PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

4.3.3.2. Genome-wide association analysis

Based on analyses presented in this Chapter thus far, the traits EPQN, GHQ, MDQ and SPQ were deemed the highest ranking endophenotypes and taken forward for genome-wide association analysis. Genome-wide association analysis of EPQN, GHQ, MDQ and SPQ yielded no genome-wide significant hits. The ratio of the observed median χ^2 to that expected by chance, denoted λ , indicated that the observed $-\log(P)$ for these traits were slightly inflated to those expected under the null, with λ (SE) values for EPQN = 1.03 (9.41e-06), GHQ = 1.03 (6.38e-06), MDQ = 1.01 (5.69e-06) and SPQ = 1.01 (8.19e-06). As stated in Section 4.3.1.2.2., the ratio from LD Score regression indicated that this inflation could be due to confounding by population and family structure, rather than polygenicity¹⁶⁵, in GS:SFHS. Although this cannot be definitely stated without repeating the analysis with imputed data, this would explain the inflation of test statistics observed despite no genome-wide significant hits. Quantile-quantile plots of observed $-\log(P)$ against expected $-\log(P)$ and Manhattan plots of genomic coordinates against $-\log(P)$ are shown for EPQN, GHQ, MDQ, and SPQ in Figures 4.3.3.2.1. - 4.3.3.2.4. The 20 most significantly associated SNPs for each endophenotype are shown in Tables 4.3.3.2.1. - 4.3.3.2.4.

4.3.3.2.1. Summary statistics from GWAS of EPQN

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
10	rs10762465	73414206	G	A	0.32	0.051	0.011	3.57e-06
21	rs2284568	35202591	A	G	0.13	-0.07	0.015	3.80e-06
2	rs2005174	41998764	A	G	0.28	0.052	0.012	7.21e-06
1	rs11249039	24552835	G	A	0.33	-0.049	0.011	1.10e-05
19	rs10417579	41465130	A	G	0.40	-0.046	0.011	1.20e-05
19	rs3889806	41459241	A	G	0.39	-0.046	0.011	1.29e-05
20	rs6515031	19670344	A	G	0.23	-0.054	0.012	1.29e-05
20	rs17295618	19652144	C	A	0.17	-0.058	0.014	1.92e-05
22	rs9615065	45611645	A	G	0.28	-0.049	0.012	2.06e-05
10	rs10823805	73386821	A	G	0.28	0.049	0.012	2.12e-05
10	rs1665688	73417236	A	G	0.31	0.048	0.011	2.17e-05
19	rs2014141	41499989	A	G	0.43	-0.044	0.010	2.32e-05
19	rs4803417	41508020	C	A	0.43	-0.044	0.010	2.36e-05
6	rs6916872	37744716	C	A	0.16	-0.060	0.014	2.56e-05
7	rs6959964	68905738	G	A	0.36	0.045	0.011	2.64e-05
19	rs2099361	41498348	C	A	0.38	-0.045	0.011	2.64e-05
10	rs911704	115343700	A	G	0.17	0.057	0.014	2.82e-05
10	rs10902869	124822417	G	A	0.40	0.044	0.011	2.87e-05
8	rs10958280	52493581	G	A	0.11	0.070	0.017	3.09e-05
4	rs381032	86309824	G	A	0.40	-0.044	0.011	3.13e-05

Table 4.3.3.2.1. Top 20 SNPs from GWAS of EPQN, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

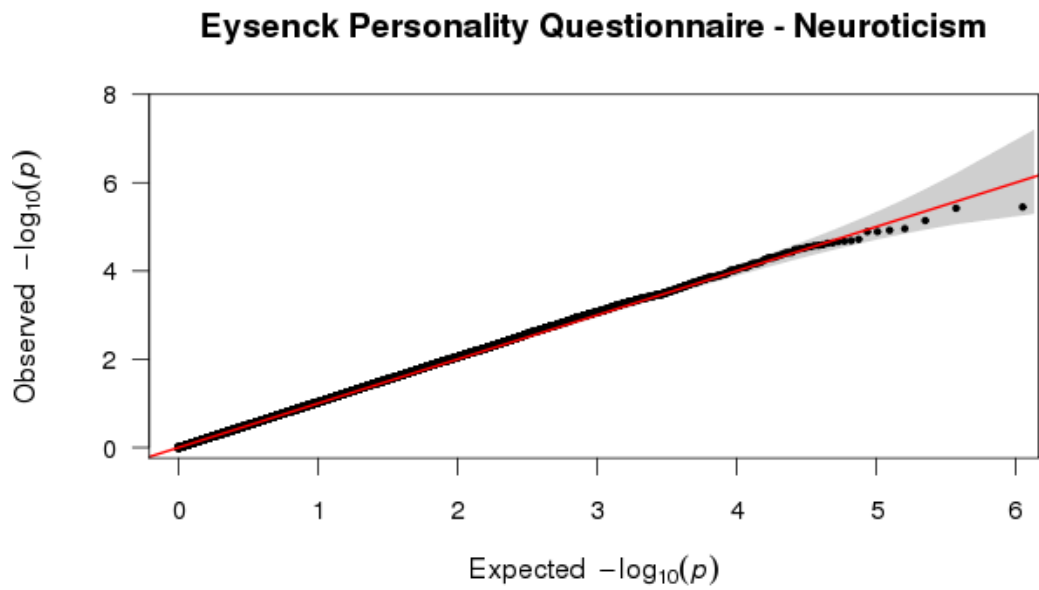


Figure 4.3.3.2.1.A. Quantile-quantile plot for Eysenck Personality Questionnaire for Neuroticism. $\lambda(\text{SE}) = 1.03 (9.41\text{e-}06)$.

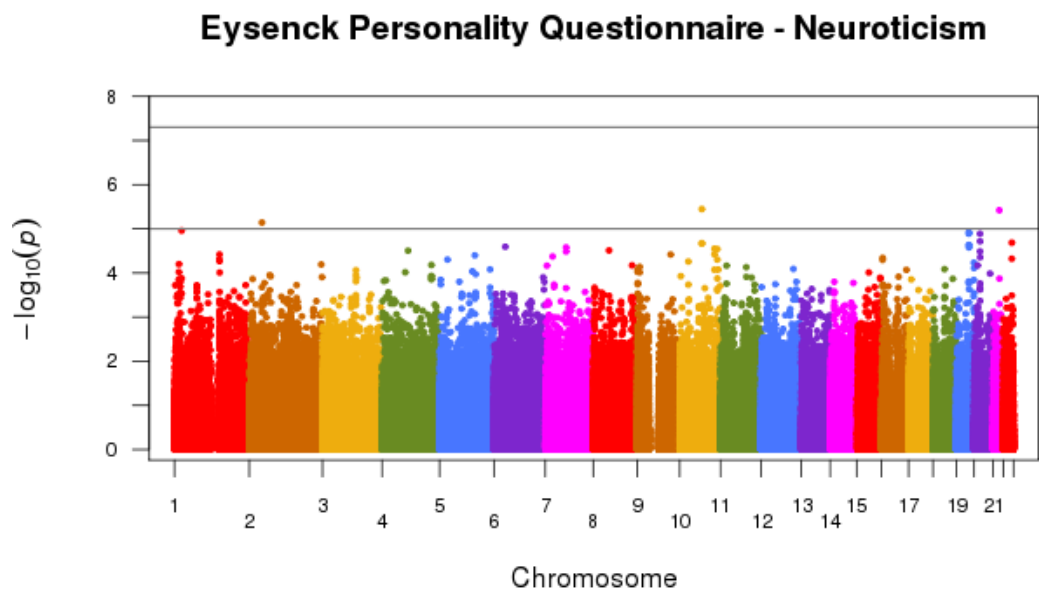


Figure 4.3.3.2.1.B. Manhattan plot for Eysenck Personality Questionnaire for Neuroticism, showing no genome-wide significant hits.

4.3.3.2.2. Summary statistics from GWAS of GHQ

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
5	rs4869652	36346515	A	G	0.46	0.051	0.011	1.66e-06
15	rs4906706	24912194	G	A	0.41	-0.051	0.011	2.17e-06
15	rs6576373	24870925	A	G	0.41	-0.051	0.011	2.39e-06
2	rs16846962	142475665	A	G	0.04	0.124	0.026	2.49e-06
4	rs6843137	129442692	A	C	0.47	0.049	0.011	2.88e-06
6	rs6923772	44533311	A	G	0.08	-0.091	0.020	3.48e-06
1	rs11205387	150482738	A	G	0.27	0.054	0.012	3.74e-06
13	rs4771539	107398943	A	G	0.37	0.050	0.011	4.25e-06
2	rs16846884	142456242	A	C	0.02	0.149	0.034	1.05e-05
6	rs11757353	107216674	G	A	0.11	0.073	0.017	1.23e-05
1	rs2820871	163508344	G	A	0.21	-0.057	0.013	1.27e-05
10	rs10466038	80097952	A	C	0.15	0.06	0.015	1.27e-05
8	rs7014898	73104407	G	A	0.19	0.058	0.013	1.57e-05
4	rs6534689	129441798	G	A	0.47	-0.045	0.011	1.65e-05
13	rs9514583	107451377	A	G	0.29	0.050	0.012	1.82e-05
7	rs1917651	145768267	G	A	0.13	-0.066	0.016	2.10e-05
4	rs7695069	129445759	G	A	0.24	-0.052	0.012	2.41e-05
18	rs1788571	71915587	G	A	0.29	0.049	0.012	2.70e-05
13	rs9558839	107391098	G	A	0.27	0.050	0.012	2.92e-05
10	rs11008099	30834632	A	G	0.09	0.077	0.019	3.04e-05

Table 4.3.3.2.2. Top 20 SNPs from GWAS of GHQ, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

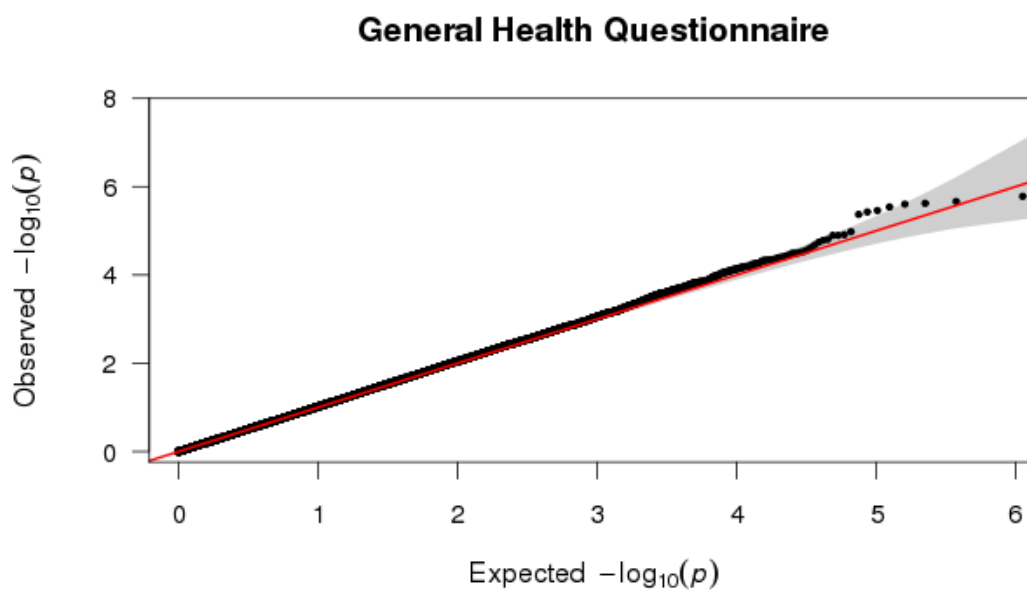


Figure 4.3.3.2.2.A Quantile-quantile plot for General Health Questionnaire.
 $\lambda(\text{SE}) = 1.03 \text{ (} 6.38\text{e-}06\text{)}$.

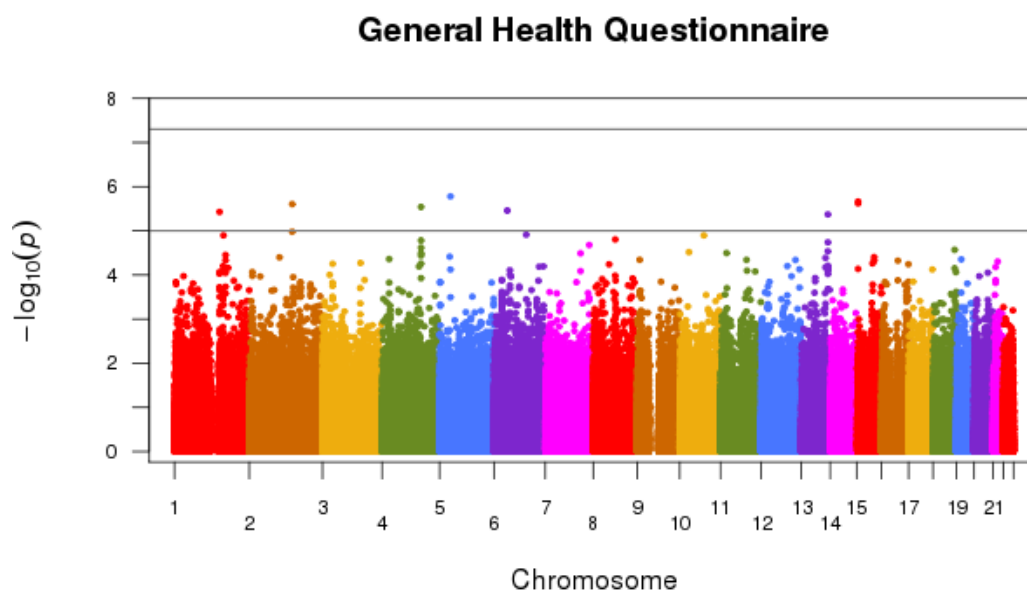


Figure 4.3.3.2.2.B. Manhattan plot for General Health Questionnaire, showing no genome-wide significant hits.

4.3.3.2.3. Summary statistics from GWAS of MDQ

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
21	rs8132771	32999423	A	G	0.06	0.142	0.030	2.77e-06
21	rs2070424	33039320	G	A	0.07	0.133	0.029	4.26e-06
5	rs166354	169330699	G	A	0.34	-0.068	0.015	8.54e-06
21	rs2833479	33051815	G	A	0.07	0.128	0.029	9.23e-06
9	rs11143216	74882572	A	C	0.08	-0.118	0.027	9.46e-06
13	rs1781553	81970842	A	G	0.39	-0.065	0.015	9.78e-06
22	rs2269670	43700013	A	G	0.02	0.238	0.054	1.16e-05
21	rs16988427	33050785	G	A	0.06	0.126	0.029	1.46e-05
7	rs10247014	133431038	A	C	0.44	0.062	0.014	1.61e-05
1	rs10927237	244624337	C	A	0.20	-0.077	0.018	1.68e-05
15	rs9302160	51590011	A	G	0.04	-0.161	0.038	2.02e-05
1	rs7541781	242293918	G	A	0.19	-0.078	0.018	2.16e-05
11	rs949000	130698548	C	A	0.16	-0.082	0.019	2.20e-05
16	rs1011097	80753476	G	A	0.37	0.063	0.015	2.24e-05
1	rs12068598	244636995	A	G	0.20	-0.076	0.018	2.29e-05
1	rs1929219	194729943	A	G	0.09	0.104	0.025	2.48e-05
9	rs11143202	74871708	A	G	0.04	-0.156	0.037	2.70e-05
11	rs7941928	2315404	A	G	0.37	0.062	0.015	2.73e-05
14	rs1956494	26851982	G	A	0.06	-0.124	0.030	2.73e-05
11	rs12421865	18742043	G	A	0.22	-0.071	0.017	3.42e-05

Table 4.3.3.2.3. Top 20 SNPs from GWAS of MDQ, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

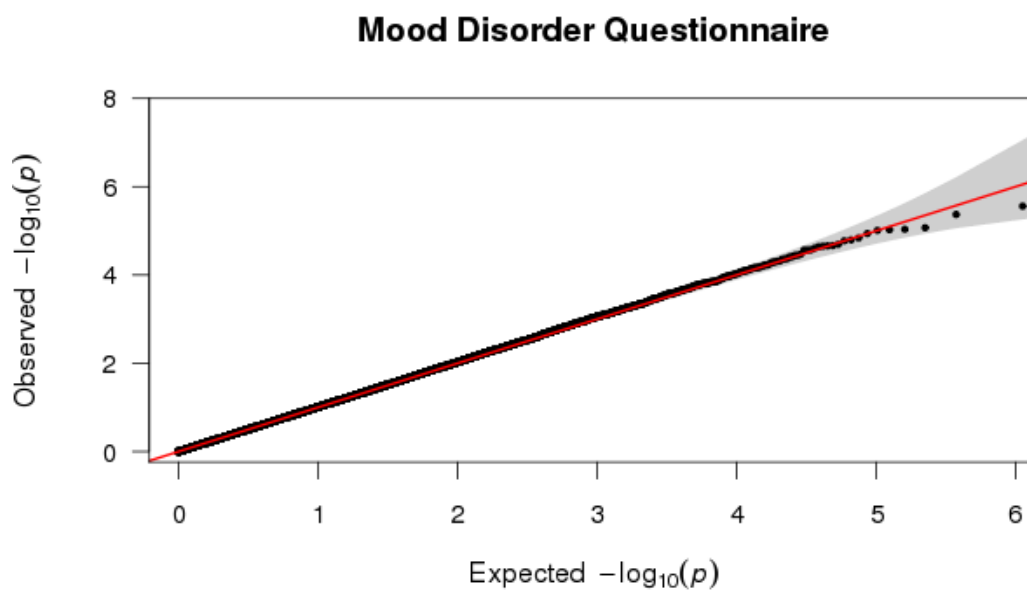


Figure 4.3.3.2.3.A Quantile-quantile plot for Mood Disorder Questionnaire. $\lambda(\text{SE}) = 1.01$ ($5.69\text{e-}06$).

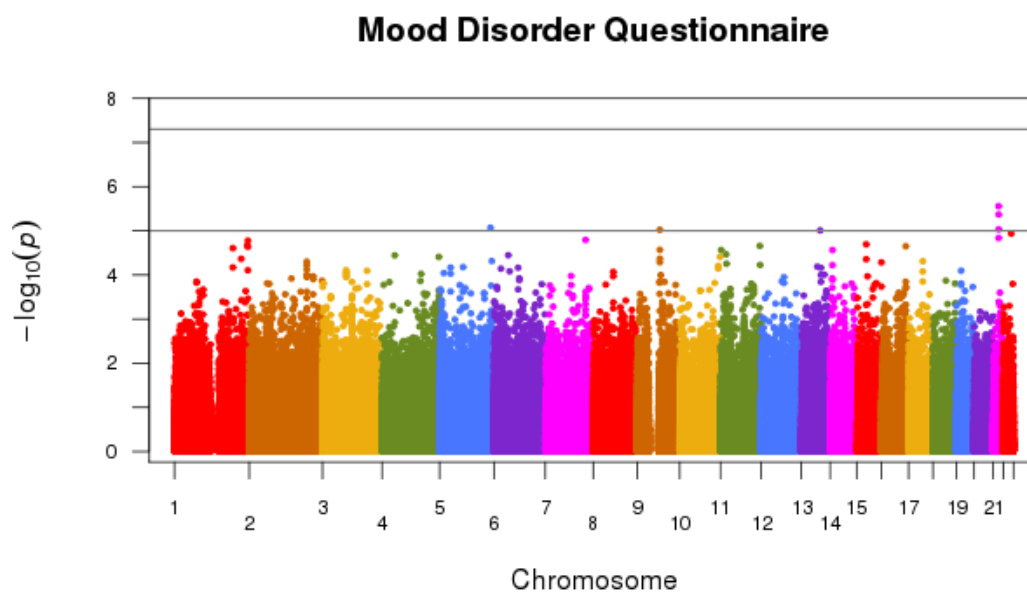


Figure 4.3.3.2.3.B. Manhattan plot for Mood Disorder Questionnaire, showing no genome-wide significant hits.

4.3.3.2.4. Summary statistics from GWAS of SPQ

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
2	rs6433223	171461637	G	A	0.17	-0.098	0.019	3.26e-07
16	rs6564387	77138585	A	G	0.11	-0.107	0.022	1.80e-06
2	rs4500910	171450353	G	A	0.14	-0.093	0.021	5.74e-06
1	rs10465638	244785995	A	G	0.17	-0.086	0.019	6.04e-06
3	rs1379078	147939350	G	A	0.20	0.080	0.018	6.90e-06
1	rs10924694	236358092	A	G	0.26	-0.073	0.016	8.09e-06
3	rs2314612	154716357	A	G	0.41	0.065	0.015	8.62e-06
3	rs6771516	147943603	A	G	0.20	0.079	0.018	9.49e-06
21	rs243268	24223261	A	G	0.22	0.076	0.017	1.04e-05
8	rs17089124	23193309	C	A	0.01	-0.275	0.062	1.07e-05
1	rs4489539	90721832	A	C	0.03	-0.192	0.044	1.23e-05
9	rs1839967	28083672	G	A	0.22	0.075	0.017	1.27e-05
3	rs7628508	147867628	G	A	0.21	0.076	0.018	1.34e-05
3	rs6810197	147911005	A	C	0.20	0.078	0.018	1.38e-05
4	rs10029731	140816848	A	G	0.15	-0.088	0.020	1.45e-05
6	rs17826169	146953914	C	A	0.45	-0.062	0.014	1.59e-05
3	rs1456682	147845656	G	A	0.21	0.076	0.018	1.60e-05
18	rs9952705	31763797	A	G	0.49	0.061	0.014	1.61e-05
20	rs2252115	51620857	G	A	0.47	0.061	0.014	1.63e-05
16	rs6564393	77149834	G	A	0.17	-0.083	0.019	1.64e-05

Table 4.3.3.2.4. Top 20 SNPs from GWAS of SPQ, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

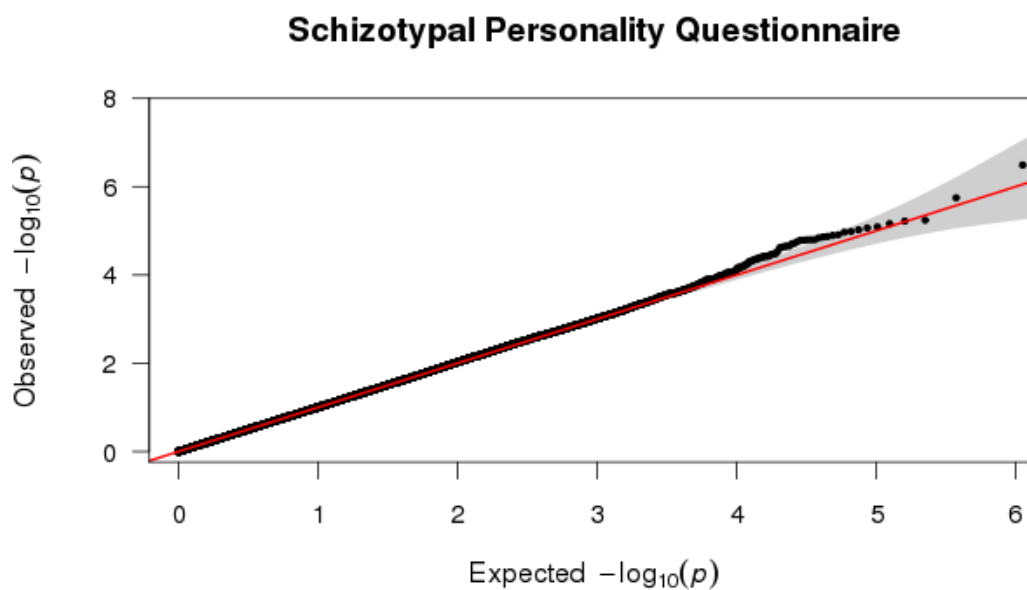


Figure 4.3.3.2.4.A. Quantile-quantile plot for Schizotypal Personality Questionnaire. $\lambda(\text{SE}) = 1.01$ ($8.19\text{e-}06$).

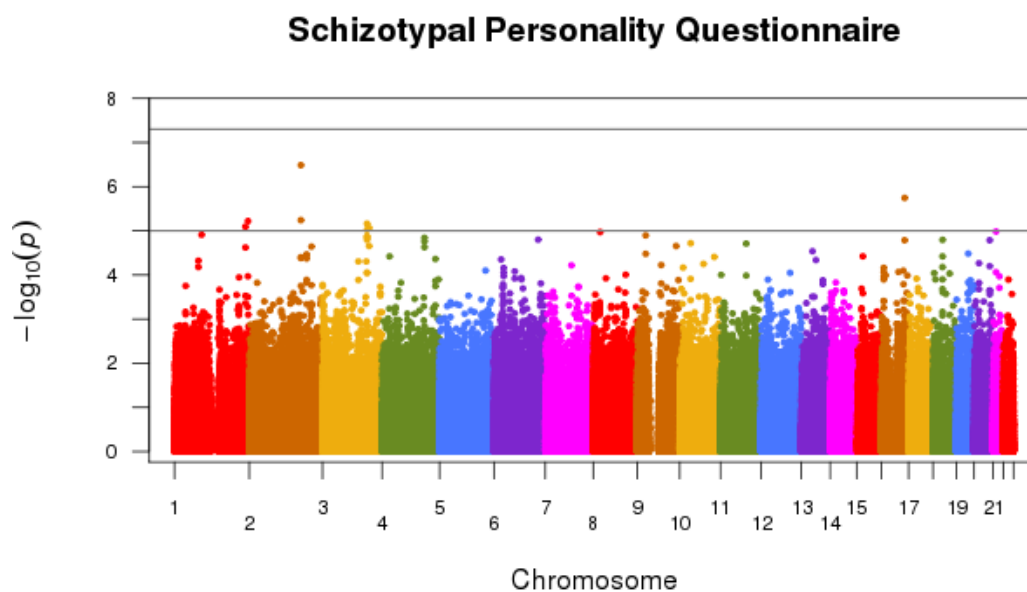


Figure 4.3.3.2.4.B. Manhattan plot for Schizotypal Personality Questionnaire, showing no genome-wide significant hits.

4.3.3.2.5. Biological annotation of top hits from EPQN GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs10762465	-	1	10:73084212 ^a	CDH23 (intronic) ^{a,b,c}	calcium ion binding	hearing loss
rs10762465	rs10823805	0.8	10:73056827 ^b			
rs10762465	rs1665688	0.79	10:73087242 ^c			
rs2284568	-	1	21:34124461 ^a	ITSN1 (intronic) ^a (non-syn) ^b	calcium ion binding; endocytosis	higher order cognition
rs2284568	rs2245099	0.67	21:34123151 ^b			
rs10417579	rs2099361	0.97	19:46190188 ^a	CYP2B6 (intronic) ^{a,d,e} (3'UTR) ^b (downstr) ^c	drug metabolism; synthesis of cholesterol, steroids and other lipids	-
rs10417579	rs707265	0.74	19:46215927 ^b			
rs10417579	rs3889806	0.93	19:46151081 ^c			
rs10417579	rs2014141	0.77	19:46191829 ^d			
rs10417579	rs4803417	0.77	19:46199860 ^e			

Table 4.3.3.2.5. Biological annotation of the top 20 SNPs from GWAS of EPQN showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs6515031	-	1	20:19618344 ^a	SLC24A3 (intronic) ^{a,e} (3'UTR) ^{b,c} (3'UTR/reg) ^d	symporter activity; calcium, potassium:sodium antiporter activity.	posterior polymorphous corneal dystrophy
rs6515031	rs2277866	0.65	20:19649943 ^b			
rs6515031	rs3748484	0.65	20:19650694 ^c			
rs6515031	rs6046271	0.58	20:19650171 ^d			
rs6515031	rs17295618	0.65	20:19600144 ^e			
rs9615065	-	1	22:43990309 ^a	C22orf9 (intronic) ^a (3'UTR) ^{b,c}	-	-
rs9615065	rs2072720	0.59	22:43972063 ^b			
rs9615065	rs1125398	0.59	22:43968154 ^c			
rs911704	-	1	10:115333690	HABP2 (intronic)	calium ion binding; glycosaminoglycan binding	thyroid cancer; carotid artery stenosis
rs10902869	rs6838	0.83	10:124807738	ACADSB (3'UTR) (intronic)	acyl-CoA dehydrogenase activity	2-methylbutyrylglycinuria; 2- methylbutyryl-coa dehydrogenase deficiency
rs10902869	rs1898462	0.51	10:124801456			
rs10958280	-	1	10:52656134	PXDNL (intronic)	heme binding; peroxidase activity	hypertrophy of breast

Table 4.3.3.2.5. Continued. Biological annotation of the top 20 SNPs from GWAS of EPQN.

Four of the 20 highest ranking SNPs from GWAS of EPQN were in intergenic regions. Of the 16 remaining SNPs, four tagged the same biologically relevant area of the genome and will be discussed here.

The fifth highest ranking SNP, rs10417579, was in high LD ($r^2=0.74$) with a 3'UTR variant, rs707265, in the *CYP2B6* (Cytochrome P450 2B6) gene. Three other top ranking variants were intronic, rs2014141, rs4803417, rs2099361 and one other downstream, rs3889806, for this gene. The gene encodes Cytochrome P450 2B6, an enzyme which catalyzes reactions involved in drug metabolism, including some antidepressants such as bupropion²⁹⁸ and mood stabilizers such as valproic acid²⁹⁹. Polymorphisms in *CYP2B6* have been shown to reduce the metabolism of sertraline, a commonly prescribed antidepressant, in patients with MDD³⁰⁰. A study characterizing the interaction of *CYP2B6* genotype with the efficacy of the antidepressant mirtazapine found that the reduction in Hamilton Depression Rating Scale scores in response to mirtazapine were significantly more pronounced in individuals with the *CYP2B6* *6/*6 genotype³⁰¹. The regional association plot for rs10417579 is shown in Figure 4.3.3.2.5.

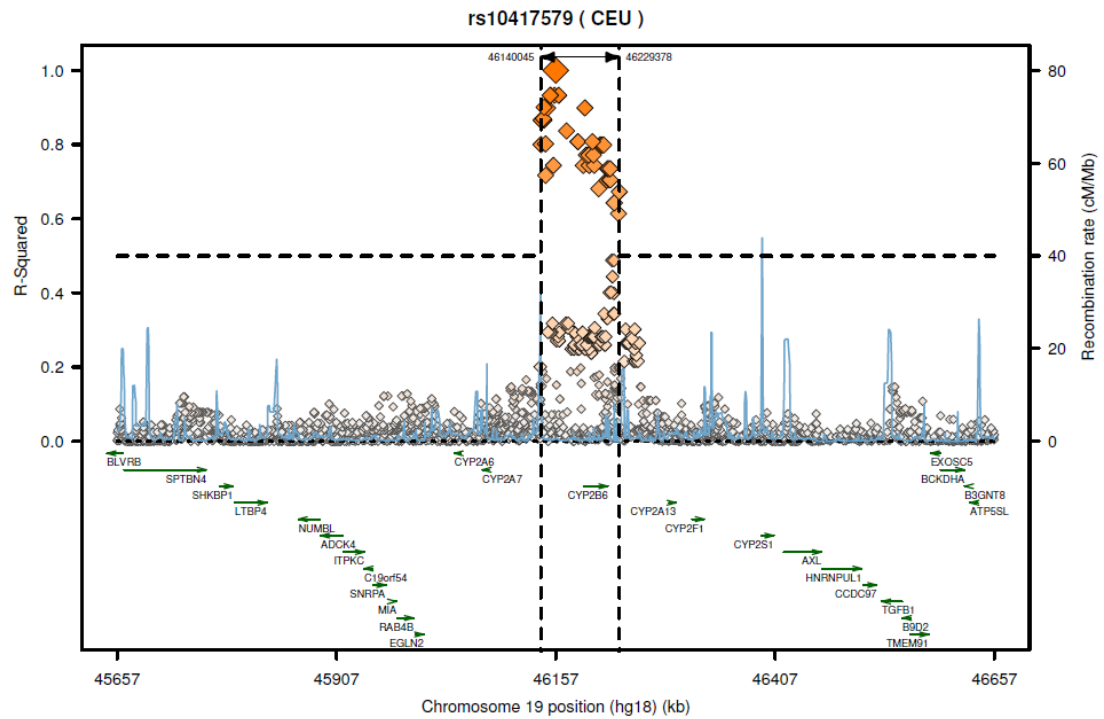


Figure 4.3.3.2.5. Regional LD plot for rs10417579 - ranked 5/20 in GWAS of EPQN. In moderate to high LD ($r^2=0.77 - 0.97$) with four genotyped variants in the *CYP2B6* gene: rs3889806 (downstream), rs2014141 (intronic), rs4803417 (intronic), and rs2099361 (intronic) ranked 6th, 12th, 13th and 16th respectively in GWAS of EPQN. Also in high LD ($r^2=0.74$) with rs707265, a 3'UTR variant in the same gene.

4.3.3.2.6. Biological annotation of top hits from GHQ GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs4869652	rs1392455	0.63	5:36336593	RANBP3L (intronic)	-	-
rs16846962	-	1	2:142192135 ^a	LRP1B (intronic) ^{a,b}	calcium ion binding; low-density lipoprotein receptor activity	endocervical carcinoma
rs16846884	-	1	2:142172712 ^b			
rs11205387	-	1	1:148749362 ^a	ECM1 (intronic) ^a (non-syn) ^b	enzyme binding; protein C-terminus binding	breast, thyroid, and other cancers; ulcerative colitis
rs11205387	rs13294	0.52	1:148751611 ^b			
rs11205387	rs7511649	0.68	1:148537415	C1orf51 (intronic)	represses circadian associated transcription	-
rs11205387	rs12029359	0.65	1:148699426	KIAA0460 (intronic)	-	-

Table 4.3.3.2.6. Biological annotation of the top 20 SNPs from GWAS of GHQ showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r^2), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r²	CHR:POS	Gene	Gene function	Disease association
rs11205387	rs1625468	0.54	1:148588110	PRPF3 (intronic)	poly(A) RNA binding; identical protein binding	retinitis pigmentosa
rs11205387	rs3738487	0.52	1:148743685	TARS2 (intronic)	nucleotide binding; ligase activity; forming aminoacyl- tRNA	Combined oxidative phosphorylation deficiency
rs1788571	rs1790815	0.92	18:70077756a	CYB5A (intronic) ^a (splice/syn) ^b	enzyme binding; cytochrome-c oxidase activity	methemoglobinemia
rs1788571	rs7238987	0.56	18:70079130b			

Table 4.3.3.2.6. Continued. Biological annotation of the top 20 SNPs from GWAS of GHQ.

Fifteen of the 20 highest ranking SNPs from GWAS of GHQ were in intergenic regions. Of the remaining 5 SNPs, no query SNP or proxy SNP tagged genes directly linked to psychiatric phenotypes, and therefore will not be explored in greater depth here.

4.3.3.2.7. Biological annotation of top hits from MDQ GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs8132771	rs2070424	0.89	21:31961191	SOD1 (intronic)	protein homodimerization activity; enzyme binding	amyotrophic lateral sclerosis
rs8132771	rs8132771	0.89	21:31969189	SFRS15 (intronic)	nucleic acid binding; nucleotide binding	Hay-Wells syndrome; Wells syndrome
rs8132771	rs2833479	0.89	21:31921294			
rs8132771	rs16988427	0.89	21:31972656			
rs166354	-	1	5:169263277	DOCK2 ^a , FAM196B ^b (intronic)	binding; guanyl-nucleotide exchange factor activity ^a ; - ^b	Immunodeficiency 40 ^a
rs11143216	rs5025470	1	9:74009090	GDA (intronic)	hydrolase activity; guanine deaminase activity	-
rs11143202	rs11143202	1	9:74052181			
rs2269670	-	1	22:42029957	SCUBE1 (intronic)	calcium ion binding; protein heterodimerization activity	-
rs10247014	-	1	7:133081578	EXOC4 (intronic)	PDZ domain binding; Ral GTPase binding	
rs16988427	rs12064941	1	1:24270358	C1orf101 (intronic)	-	-
rs16988427	rs12068598	1	1:242703618			

Table 4.3.3.2.7. Biological annotation of the top 20 SNPs from GWAS of MDQ showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r^2), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r²	CHR:POS	Gene	Gene function	Disease association
rs9302160	-	1	15:49377303	CYP19A1 (intronic) (non-syn)	estrogen biosynthesis	aromatase imbalance; endometriosis
rs9302160	rs700519	0.83	15:49295260			
rs7541781	-	1	1:240360541	PLD5 (intronic)	-	benign uterine neoplasms
rs1011097	-	1	16:79310977	CDYL2 (intronic) (syn)	methylated histone binding	-
rs1011097	rs8049284	0.59	16:79276380			
rs7941928	rs2074019	0.68	11:2275443	C11orf21 (3'UTR)	-	beckwith-wiedemann syndrome (bws)
rs7941928	rs10831733	0.64	11:2281016	TSPAN32 (intronic)	immune system; NAD metabolism	bws; some cancers
rs12421865	-	1	11:18697741	IGSF22 (intronic)	-	-

Table 4.3.3.2.7. Continued. Biological annotation of the top 20 SNPs from GWAS of MDQ.

Four of the 20 highest ranking SNPs from GWAS of GHQ were in intergenic regions. Of the 16 remaining SNPs, two tagged biologically interesting genes and will be discussed here.

The highest ranking SNP in the MDQ GWAS, rs8132771, is an intronic variant in *SOD1* (Superoxide Dismutase 1). This gene encodes an enzyme which catalyzes the dismutation of superoxide anions, a type of reactive oxygen species (ROS). In mice, transgenic overexpression of *SOD1* resulted in increased resilience to glucocorticoid-induced depressive-like behaviour relative to wild type mice in the social interaction test, sucrose preference test and forced swim test. Overexpression of *SOD1* also resulted in the reduction of a marker of oxidative stress, suggesting that this increased resilience is mediated by decreasing cellular ROS levels³⁰². In humans, significantly reduced expression of *SOD1* has been observed in conjunction with telomere shortening in white matter oligodendrocytes from MDD donors, relative to control donors³⁰³. MDD patients have also been shown to exhibit significantly higher activity levels of *SOD1* during an acute depressive episode relative to healthy controls³⁰⁴. This SNP, rs8312771, is also in high LD ($r^2=0.89$) with rs2070424, rs2833479 and rs16988427, three other high ranking SNPs (2nd, 4th and 8th, respectively) which are intronic variants in *SFRS15* (Splicing factor, arginine/serine rich 15). However, this gene did not appear to be directly relevant to any psychiatric phenotype. The regional LD plot for rs8312771 is shown in Figure 4.3.3.2.7.A.

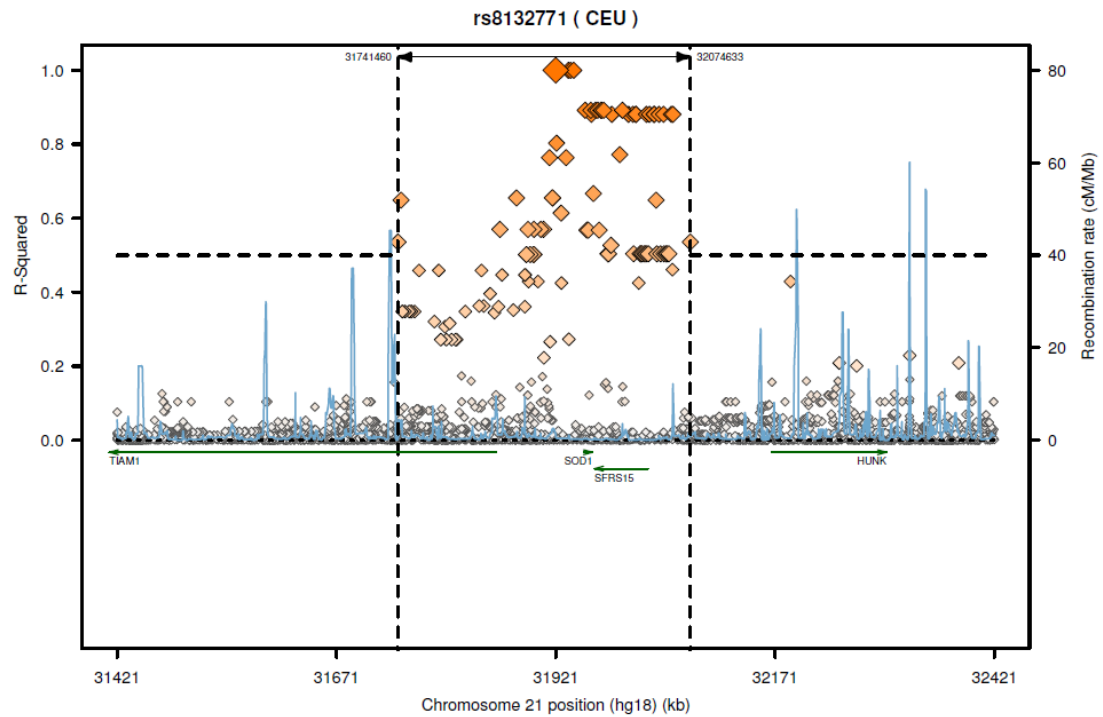


Figure 4.3.3.2.7.A. Regional LD plot for rs8132771, an intronic SNP in *SOD1* - ranked 1/20 in GWAS of MDQ. In high LD ($r^2=0.89$) with rs2070424, rs2833479 and rs16988427, intronic SNPs in *SFRS15* - ranked 2/20, 4/20 and 8/20, respectively, in GWAS of MDQ.

The eleventh highest ranking SNP, rs9302160, is an intronic variant in *CYP19A1* (Cytochrome P450 19A1), and is in high LD with ($r^2=0.83$) with rs700519 - a non-synonymous variant in the same gene. This gene encodes for aromatase, an enzyme which catalyzes the last steps of estrogen biosynthesis. In mice, estrogen deficient (aromatase knockout) female mice exhibit depressive-like behaviour, such as deficits in sexual behaviour and decreased struggling and swimming. Additionally, increased serotonergic activity was observed in the hippocampus of knockout females compared to wild types³⁰⁵. The findings pertaining to deficits in sexual behaviour are of particular interest as sexual dysfunction is frequently comorbid in patients with depression, in particular, reduced sexual desire. Female sexual arousal disorder is treated in females with estrogen replacement therapy³⁰⁶. In male gonadectomized rats both testosterone and estradiol had antidepressant-like effects. Chronic inhibition of aromatase within the dentate gyrus blocked the protective effects of

testosterone, suggesting that these effects are mediated, in part, by the aromatization of testosterone into estrogen³⁰⁷. In humans, aromatase inhibitors are used to improve disease-free survival in breast cancer patients. The literature suggests that this treatment may cause depressive symptoms (Henry et al 2008). Also, breast cancer survivors treated with aromatase inhibitors often report reduced cognitive capacity. This was found to be predominantly caused by depression, rather than treatment with aromatase inhibitors. However, this study did not seek to test whether the treatment actually exacerbated depressive symptoms³⁰⁸. The regional LD plot for rs9302160 is shown in Figure 4.3.3.2.7.

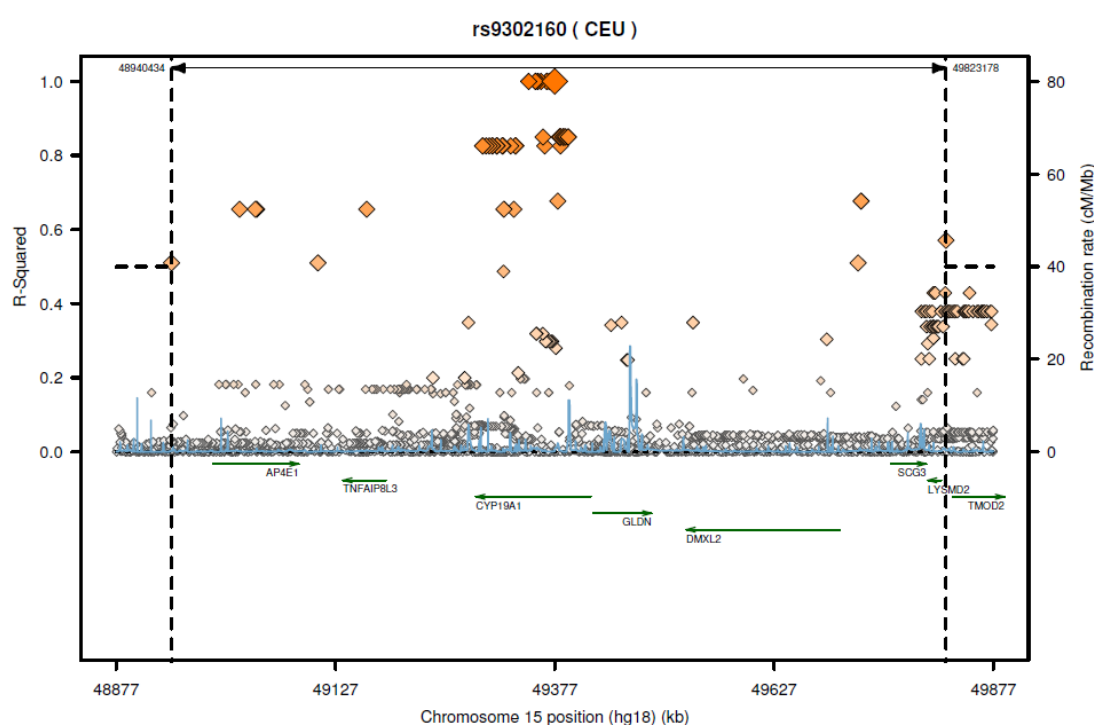


Figure 4.3.3.2.7. Regional LD plot for rs9302160, an intronic SNP in *CYP15A1* - ranked 11/20 in GWAS of MDQ.

4.3.3.2.8. Biological annotation of top hits from SPQ GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs6433223	-	1	2:171169883	MYO3B (intronic)	transferase activity; protein tyrosine kinase activity	Bardet-biedl syndrome
rs6433223	rs4500910	0.5	2:171158599			
rs10465638	-	1	1:242852618	C1orf101 (intronic)	-	-
rs10924694	-	1	1:234424715	GPR137B (intronic)	signal transduction across lysosome membrane	some cancers
rs17089124	-	1	8:23249254	LOXL2 (intronic)	chromatin binding; electron carrier activity	cancer, poorer prognosis
rs1839967	-	1	9:28073672	LINGO2 (intronic)	-	essential tremor/ Parkinson's Disease
rs10029731	-	1	4:141036298	MAML3 (intronic)	transcription coactivator activity	malignant triton tumor
rs17826169	-	1	6:146995607	C6orf103/ADGB (intronic)	iron ion binding; oxygen binding	-
rs9952705	-	1	18:30017795	NOL4 (intronic)	RNA binding	some cancers
rs2252115	-	1	20:51054264	TSHZ2 (intronic)	probable transcriptional regulator	some cancers

Table 4.3.3.2.8. Biological annotation of the top 20 SNPs from GWAS of SPQ showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

Ten of the 20 highest ranking SNPs from GWAS of GHQ were in intergenic regions. Of the 10 remaining SNPs, no query SNP or proxy SNP tagged genes directly linked to psychiatric phenotypes, and therefore will not be explored in greater depth here.

4.3.4. Replication in UKBiobank

4.3.4.1. Paired samples sign test

The ORs of SNPs in GS:SFHS and UKB were tested for consistent directionality using SNPs with a p-value of $<1E-04$ from GWAS of top ranking endophenotypes (EPQN, GHQ, MDQ, and SPQ) in GS:SFHS to perform a paired samples sign test. This resulted in testing 63 SNPs for EPQN, 76 SNPs for GHQ, 59 SNPs for MDQ and 69 SNPs for SPQ. For these SNPs, the replication samples did not produce ORs in the same directions as the discovery analysis more frequently than expected by chance (unadjusted p-value = 0.05). However, in two instances (SPQ and UKB narrow MDD definition, SPQ and UKB strict MDD definition) the SNPs were significantly in different directions more than would be expected by chance ($p=0.03$ and $p=0.02$, respectively). Full results from the paired samples sign test are shown in Table 4.3.4.1.

Trait	UKB definition	nTrials	nSuccesses	Probability of success (CI)	P-value
EPQN	broad	63	32	0.51 (0.38, 0.64)	1
	narrow		25	0.40 (0.28, 0.53)	0.13
	strict		31	0.49 (0.36, 0.62)	1
GHQ	broad	76	40	0.53 (0.41, 0.64)	0.73
	narrow		45	0.59 (0.47, 0.70)	0.14
	strict		44	0.58 (0.46, 0.69)	0.21
MDQ	broad	59	26	0.44 (0.31, 0.58)	0.44
	narrow		30	0.51 (0.37, 0.64)	1
	strict		27	0.46 (0.33, 0.59)	0.60
SPQ	broad	69	32	0.46 (0.34, 0.59)	0.63
	narrow		25	0.36 (0.25, 0.49)	0.03
	strict		24	0.35 (0.24, 0.47)	0.02

Table 4.4.3.1. Paired samples sign test between highest ranking endophenotypes in GS:SFHS and UKB MDD definitions, with nTrials indicating the number of SNPs $p < 1E-04$, and nSuccesses indicating how many times ORs were in the same direction in both samples.

4.3.4.2. Bivariate LD Score regression

Results from bivariate LD Score regression of summary statistics from GWAS of the highest ranking endophenotypes in GS:SFHS and MDD in UKB are shown in Table 4.3.4.2. Broad, narrow and strict definitions of MDD in UKB were positively correlated with EPQN, GHQ, MDQ and SPQ in GS:SFHS. Unlike the binary definition of MDD, the genetic correlation between these traits and MDD

in UKB were constrained between 0 and 1. However, only the correlation between EPQN and the strict definition of MDD in UKB survived multiple testing correction ($p \leq 0.002$). Standard errors are large for all correlations, this could be reduced by using imputed data to increase the number of SNPs used in the analysis.

Trait	UKB definition	rg (SE)	P-value	h2_liab (SE)	h2_int (SE)	gcov_int (SE)
EPQN	Broad	0.55 (0.19)	0.004	0.092 (0.03)	1.00 (0.008)	-1.04e-05 (0.005)
	Narrow	0.63 (0.29)	0.03	0.09 (0.03)	1.00 (0.008)	-0.005 (0.006)
	Strict	0.73 (0.23)	0.002	0.09 (0.03)	1.00 (0.008)	-0.005 (0.005)
GHQ	Broad	0.84 (0.37)	0.02	0.04 (0.03)	1.01 (0.008)	-0.002 (0.005)
	Narrow	0.66 (0.46)	0.148	0.04 (0.03)	1.01 (0.008)	-0.001 (0.006)
	Strict	0.76 (0.38)	0.04	0.04 (0.03)	1.01 (0.008)	-0.003 (0.005)
MDQ	Broad	0.75 (1.00)	0.46	0.03 (0.05)	1.00 (0.008)	0.008 (0.005)
	Narrow	0.60 (0.74)	0.418	0.03 (0.05)	1.00 (0.008)	-1.17e-04 (0.006)
	Strict	0.41 (0.57)	0.47	0.03 (0.05)	1.00 (0.008)	0.005 (0.006)
SPQ	Broad	0.24 (0.25)	0.33	0.09 (0.05)	0.99 (0.007)	0.004 (0.005)
	Narrow	0.10 (0.32)	0.75	0.09 (0.05)	0.99 (0.007)	0.005 (0.005)
	Strict	0.04 (0.27)	0.895	0.091 (0.048)	0.991 (0.007)	0.006 (0.005)

Table 4.3.4.2. Bivariate LD Score regression between the highest ranking candidate endophenotypes and the broad, narrow and strict MDD phenotype definition in UK Biobank.

4.4. Discussion

4.4.1. Summary of findings

4.4.1.1. Did quantitative traits fulfill endophenotype criteria?

As stated in the introduction to this chapter, the selected measures of mood, personality and cognitive ability were assessed on whether they fulfilled the endophenotype criteria, as outlined by Gottesman and Gould. The results of analyses seeking to test this have been summarized in Table 4.4.1.1. Two traits, GHQ and EPQN, fulfilled all endophenotype criteria. A further three traits, DSC, MDQ and SPQ, fulfilled all endophenotype criteria except sharing a statistically significant phenotypic correlation with MDD.

Trait	Assoc with illness in pop.	More extreme in 1st deg. rel.	State indep.	Heritable	r_g	r_p
DSC	Y	Y	Y	Y	Y	N
EPQE	Y	N	N	Y	N	N
EPQN	Y	Y	Y	Y	Y	Y
GHQ	Y	Y	Y	Y	Y	Y
LM1	N	N	N	Y	N	N
LM2D	N	N	N	Y	N	N
MDQ	Y	Y	Y	Y	Y	N
MHV	N	N	N	Y	N	N
SPQ	Y	Y	Y	Y	Y	N
VF	N	N	N	Y	N	N

Table 4.4.1.1. Summary of how well selected quantitative traits fulfilled endophenotype criteria. Assoc with illness in pop.: the trait is associated with MDD in the population; More extreme in 1st deg. rel.: the trait is more extreme in first degree relatives; State indep.: the trait is state independent; Heritable: the trait is heritable; r_g : the trait is genetically correlated with MDD; r_p : the trait is phenotypically correlated with MDD; Y/N: Yes/No. Traits which fulfill most or all of these criteria are highlighted in bold.

4.4.1.2. Which endophenotypes ranked highest using measures covariation?

Endophenotypes ranked by their coheritability with MDD, and variance explained by MDD PGRS are shown in Table 4.4.1.2. EPQN and GHQ ranked highly, and showed statistically significant association, using both methods. MDQ, SPQ, and DSC ranked highly and were significantly associated with MDD using pedigree-based methods, but were not significantly associated with or highly ranked using MDD PGRS. LM2D was ranked highly, and significantly, using MDD for PGRS, but not in pedigree based methods. The convergence of ranking methods was significant when only incorporating traits which had a high number of observations ($n \geq 19,646$), but not when using all traits.

Coheritability with MDD (h_{ij} , h_{ij0})	MDD PGRS variance explained (%)
<i>MDQ (0.25, 0.15)</i>	GHQ (0.19)
<i>SPQ (0.19, 0.12)</i>	EPQN (0.17)
EPQN (0.18, 0.11)	<i>LM2D (0.09)</i>
GHQ (0.15, 0.09)	<i>MDQ (0.07)</i>
<i>DSC (0.10, 0.06)</i>	LM1 (0.06)
EPQE (0.06, 0.03)	<i>DSC (0.04)</i>
LM1 (0.03, 0.02)	<i>SPQ (0.03)</i>
VF (0.02, 0.02)	MHV (0.01)
LM2D (0.02, 0.02)	EPQE (0.004)
MHV (0.02, 0.02)	VF (0.003)

Table 4.4.1.2. Measures of genetic covariance as ranking methods for candidate endophenotypes. Traits highlighted in bold survived multiple testing correction. Traits highlighted in italics are significant for one ranking method, but not the other.

4.4.1.3. Did endophenotypes "outperform" binary MDD?

4.4.1.4.1. Using variance explained by PGRS

Which PGRS explained the most variance in MDD and candidate endophenotypes (at the P_T which explained the most variance in the discovery sample) is summarized in Table 4.3.3.1.5. Interestingly, the PGRS which significantly explained the most trait variance in consistently high ranking endophenotypes, EPQN and GHQ, was MDD PGRS. This is consistent with results from recurrent MDD, which in theory should be a more extreme depressed phenotype. These traits also had a higher proportion of variance explained by MDD PGRS (EPQN = 0.17%, GHQ = 0.19%) than MDD as a binary phenotype (0.08%). MDQ, which had a very high genetic correlation with MDD ($r_g=0.92$), was only significantly associated with schizophrenia PGRS. Whilst MDD is also significantly associated with schizophrenia PGRS, it is surprising that a trait so highly genetically correlated with MDD does not correlate with MDD PGRS. SPQ, also highly genetically correlated with MDD ($r_g=0.71$), was not significantly associated with any PGRS at any threshold.

Phenotype	Variance explained by PGRS (high → low)
MDD	SCZ, MDD , BIP, AUT
DSC	SCZ , MDD, BIP, AUT
EPQE	BIP, MDD, AUT, SCZ
EPQN	MDD, SCZ , BIP, AUT
GHQ	MDD, SCZ , BIP, AUT
LM1	SCZ, AUT , MDD, BIP
LM2D	SCZ, MDD, AUT , BIP
MDQ	SCZ , MDD, BIP, AUT
MHV	AUT , SCZ, MDD, BIP
SPQ	MDD, AUT, SCZ, BIP
VF	SCZ, AUT, MDD, BIP

Table 4.4.1.4.1. Psychiatric PGRS ranked by the trait variance they explain in MDD and candidate endophenotypes, with significant associations highlighted in bold. SCZ: schizophrenia; MDD: major depressive disorder; BIP: bipolar disorder; AUT: autism spectrum disorder.

4.4.1.4.2. In genome-wide association analysis

GWAA of the highest ranking endophenotypes yielded a higher number of SNPs showing more extreme p-values ($P \leq 1E-04$: EPQN = 63 SNPs, GHQ = 76 SNPs, MDQ = 59 SNPs and SPQ = 69 SNPs) than in the GWAA of MDD as a binary trait (50 SNPs). This suggests that quantitative traits that are genetically correlated with MDD have improved power to detect genetic variants in GWAA, albeit there were no genome-wide significant hits. The majority of top ranking SNPs did not tag genes which were explicitly linked to psychiatric phenotypes. However, three genes were identified that were biologically plausible: *CYP2B6*, *SOD1* and *CYP19A1* - none of which were identified in GWAA of MDD as a binary trait, or more familial subtypes of MDD1. This supports the utility of analyzing these traits.

4.4.1.4. Did endophenotypes correlate with MDD in UKB?

From the results, it cannot be definitively ascertained whether the highest ranking endophenotypes (EPQN, GHQ, MDQ and SPQ) in GS:SFHS correlate with MDD in UKB. The sign test showed no consistency in direction of effects for SNPs with $P \leq 1E-4$ relative to what would be expected by chance. Regarding results from bivariate LD Score analysis, although all traits had genetic correlations with UKB MDD definitions that were within bounds (0-1), only the correlation between EPQN and the strict definition of MDD in UKB survived multiple testing correction. Whilst EPQN and GHQ have correlations with UKB MDD definitions which are $p < 0.05$, but do not survive multiple testing correction, correlations between MDQ, SPQ and UKB MDD definitions appear to be completely non-significant. This indicates that although these traits are associated with MDD within GS:SFHS, they possibly do not have utility in other large samples of European descent such as UKB. These results may be a product of the smaller number of observations available for MDQ and SPQ.

4.4.2. Limitations

4.4.2.1. Traits selected for testing endophenotype criteria

Regarding the endophenotype approach the seminal paper, which suggests their use in psychiatric genetics, states that the assumption of variation in an endophenotype depending upon variation in fewer genes relative to the more complex disease phenotype is fundamental to the concept¹²⁶. Although this point was argued against in the introduction, GWAA yielded no genome-wide significant hits, and this will be, in part, a product of the complex architecture of the selected traits. Secondly, it must be considered that even if a candidate endophenotype satisfies all the criteria, this does not exclude the possibility that it is epiphenomenal with respect to the disease, and may not lie on the disease pathway³⁰⁹.

4.4.2.1. Model selection

This limitation is covered in Chapter 3, due to the mixed linear model appearing to be overly conservative in correcting for population and family structure in GS:SFHS. However, for quantitative traits, this model appears to be insufficient in correcting for population and family structure in some traits, notably GHQ and MDQ. This, again, suggests that a more comprehensive model selection process could be beneficial to identify a model which is not overly conservative for binary traits, and insufficient for some quantitative traits.

4.4.3. Conclusions

This Chapter has established that two out of ten available measures of mood, personality and cognitive ability, EPQN and GHQ, fulfill all endophenotype criteria and rank highly using measures of genetic covariation. These traits also performed better than the binary definition of MDD in terms of number of extreme p-values in GWAA, the variance explained by MDD PGRS, and the strength of association with MDD PGRS. Three traits, MDQ, SPQ and DSC, fulfilled all endophenotype criteria, except phenotypic correlation with MDD, and ranked highly when the metric was coheritability, but not when the metric was variance explained by PGRS. Although DSC was a possible candidate and generally survived multiple testing correction, there was a notable drop in strength of association between DSC and MDD relative to other high ranking traits, hence why for GWAA and replication in UKB, only EPQN, GHQ, MDQ and SPQ were considered. Therefore, EPQN, GHQ, MDQ and SPQ will be taken forward for multivariate analysis.

Chapter 5: Using multivariate methods to jointly maximize heritability and genetic correlation with depression

5.1. Introduction

The aim of analyses in this Chapter was to assess whether the highest ranking endophenotypes for MDD, as identified in Chapter 4, could be used to generate a quantitative endophenotype for MDD where genetic correlation and heritability were jointly maximized. The analyses performed in this Chapter, namely heritability-based analysis, GWAA and PGRS were used to test the suitability of derived quantitative traits as proximal measures of MDD and to assess whether composite traits had greater power to identify genetic risk variants for depression than MDD as a binary trait, or single quantitative traits.

Firstly, results from analyses seeking to test whether composite traits fulfill endophenotype criteria to a greater extent than single traits will be presented. Secondly, results from endophenotype ranking methods will determine whether composite traits have increased genetic covariation with MDD relative to single traits, in terms of coheritability and variance explained by polygenic risk score.

Next, results from genome-wide association analysis will be presented to establish whether composite traits have improved statistical power to detect genetic effects relative to single traits and MDD as a binary trait. These results will be biologically annotated to reveal whether there is any increased tractability yielded by creating a trait which aims to encompass the heritable component of depression.

Lastly, results from replication attempts in UKBiobank will be presented to assess whether the derived quantitative trait has utility outwith GS:SFHS.

5.2. Methods

5.2.1. Creating the composite traits

Three composite measures of depression were derived incorporating the two, three or four traits which ranked most highly across endophenotype criteria. The two trait model used information from EPQN and GHQ, the three trait measure from EPQN, GHQ and MDQ, and the four trait measure from EPQN, GHQ, MDQ and SPQ.

5.2.1.1. Using two traits

The two trait measure was constructed by simply summing the two standardized variables ($z_{EPQN} + z_{GHQ}$). This is because the contributing trait weights cannot be estimated via PCA for fewer than three traits, therefore this method assigns equal weight to each trait. This is a reasonable alternative to PCA when only two variables are being used. This is not the first PC *per se*, but an alternative scoring. In PCA, score vectors are created for the number of PCs retained for individuals. These are weighted composites of the input variables and component weights derived through an eigen decomposition. These weights cannot be estimated with only 2 variables, so there is no way to assess differing contribution. Therefore, equal weights are assigned by using a sum scores. Using the standardized phenotypes is the equivalent to the scaling and centring of traits which is performed as part of PCA. This trait is referred to throughout this Chapter as SoS (Sum of Z-Scores).

5.2.1.2. Using more than two traits

When incorporating three or four traits, measures were generated using genetic and phenotypic correlations between traits. Genetic correlations were derived using the bivariate heritability methods, as outlined in Section 2.2.3. Variance component estimates from bivariate analysis of quantitative trait pairs are provided in Appendix 5.A. From this, a genetic correlation matrix between traits was constructed, as shown in Table 5.2.1.2.A.

	GHQ	EPQN	MDQ	SPQ
GHQ	1	0.64	0.74	0.75
EPQN	0.64	1	0.84	0.71
MDQ	0.74	0.84	1	0.76
SPQ	0.75	0.71	0.76	1

Table 5.2.1.2.A. genetic correlation matrix for pairs of quantitative traits, estimated using bivariate heritability methods.

This matrix was then subjected to spectral decomposition using the eigen function in R to compute eigenvalues and eigenvectors. The loading for the first principal component of each trait was then extracted, multiplied by its corresponding standardized trait value, and the resulting values summed together across traits to create the composite trait, as shown in Figure 2.3. As these traits were created using the genetic correlations between three and four trait pairs, these traits are called Gen3 and Gen4 for the remainder of this Chapter. For Gen3, traits loaded quite equally onto the 1st PC, with eigenvectors for EPQN = 0.58, GHQ = 0.55 = and MDQ = 0.60. For Gen4 traits were loaded evenly but negatively onto the 1st PC, with eigenvectors for EPQN = -0.50, GHQ = -0.48, MDQ = -0.52, and SPQ = -0.50.

Alternative composite measures were also constructed by PCA on the traits, and using the first unrotated principal component as the trait. This was to observe the convergence between traits derived using the eigen function versus the prcomp function, as the latter method is preferred for numerical accuracy. As standardized variables were used, traits were already scaled and centred to have a mean of zero, and unit variance.

A summary of results from the PCA of EPQN, GHQ and MDQ are shown in Table 5.2.1.2.B. EPQN and GHQ loaded fairly equally whilst MDQ contributed a lower loading onto the first PC, with eigenvectors for EPQN = 0.62, GHQ = 0.59, and MDQ = 0.52.

	PC1	PC2	PC3
Proportion of variance	0.59	0.24	0.17
Standard deviation	1.33	0.86	0.71

Table 5.2.1.2.B. The proportion of variance explained by each principal component and the standard deviation of the principal components, calculated as the square root of the eigenvalues of the covariance/correlation matrix.

A summary of results from the PCA of EPQN, GHQ, MDQ and SPQ are shown in Table 5.2.1.2.C. Variables loaded more equally onto the first PC across traits, with eigenvalues for EPQN = 0.52, GHQ = 0.47, MDQ = 0.49 and SPQ = 0.52.

	PC1	PC2	PC3	PC4
Proportion of variance	0.55	0.20	0.13	0.12
Standard deviation	1.49	0.90	0.72	0.69

Table 5.2.1.2.B. The proportion of variance explained by each principal component and the standard deviation of the principal components, calculated as the square root of the eigenvalues of the covariance/correlation matrix.

Traits derived from eigendecomposition of the genetic correlation matrix were almost perfectly correlated with traits derived from PCA, although due to the negative loadings of traits onto the 1st PC in Gen4, this correlation was negative between Gen4 and Phen4.

5.2.2. Testing endophenotype criteria

5.2.2.1. Associated with illness in the population

The effect of affection status on mean trait scores was calculated using the Wald Conditional F-test, implemented in ASReml-R¹⁷³, as described in Section 2.2.4. and 4.2.1.1. Composite traits were compared to single traits in terms of effect size and strength of association.

5.2.2.2. More extreme in unaffected relatives than in the general population

Whether unaffected first degree relatives had more extreme trait values relative to the population was calculated as described in Section 4.2.1.2. Again, composite traits were compared to single traits in terms of effect size and strength of association.

5.2.2.3. State independent

State independence was assessed as described in Section 4.2.1.3, by observing whether the endophenotype was more extreme in unaffected relatives than the general population was used as a proxy measure of state independence.

5.2.2.4. Heritability of composite traits

5.2.2.4.1. Using pedigree data

Heritability estimates for composite traits were estimated using restricted maximum likelihood methods, implemented in ASReml-R¹⁷³, as described in Section 2.2.1.3. Variance component estimates for these analyses are provided in Appendix 5.B.

5.2.2.4.2. Using GWAA summary statistics

Heritability estimates for composite traits on the observed scale using GWAA summary statistics were determined using single trait LD Score regression, implemented in LD Score (v1.0.0.), as described in Section 3.2.2.2. The deviation

of the LD score intercept from 1 against its standard error, and the LD Score ratio were interrogated to assess the contributions of confounding versus polygenicity¹⁶⁵.

5.2.2.5. Genetic and phenotypic correlation

Bivariate genetic analyses of composite traits and MDD were conducted in ASReml-R¹⁷³, as described in Section 2.2.3. The full model (which assumes additive covariance between traits) did not converge for Gen3 or Gen4. Unlike models described in Section 4.2.1.3. which ran, but did not converge, these models required termination as they got stuck in an iterative loop. Looking at the bivariate models between constituent single traits (EPQN, GHQ, MDQ, and SPQ) and MDD from Appendix 4, it is apparent that again, fitting shared sibling environment that is proving problematic. This is due to very low sibling variance for MDD and low sibling covariance between these quantitative traits and MDD. Therefore, again, this problem was circumnavigated using the conditional model, as described in Section 4.2.1.3., which only fits common sibling environmental for composite traits. This resolved convergence issues. Variance component estimates for these analyses are provided in Appendix 5.C. Composite traits were compared to single traits in terms of the magnitude and strength of their genetic and phenotypic correlation with MDD.

5.2.3. Performance of composite traits relative to single traits

5.2.3.1. Coheritability

Estimates of coheritability between composite traits and MDD were calculated using heritability estimates for MDD on the liability¹⁴² and observed scale, as described in Section 4.2.2.1. Composite traits were compared to single traits in terms of change in coheritability estimate, indicating whether composite traits were a better or worse predictor of MDD than single, quantitative traits.

5.2.3.2. Polygenic risk score analysis

Polygenic risk scores for MDD were created as described in Section 2.2.6. Composite traits were compared to single traits and binary MDD in terms of

strength of association with and variance explained by MDD PGRS to determine whether composite traits had improved power relative to these traits. Again, analysis was repeated with PGRS for autism, bipolar disorder and schizophrenia to explore differential enrichment of associated alleles, and to establish whether composite traits are specific to depression, or more representative of psychiatric disorders in general.

5.2.3.3. Genome-wide association analysis

Genome-wide association analysis for each composite traits was conducted using restricted maximum likelihood methods, implemented in GCTA (v1.25.)^{166,167}, as described in Sections 2.2.5. and 3.2.3. As composite traits are quantitative, test statistics did not require post hoc transformation, as described in Section 4.2.3.2. Composite traits were compared to single traits and the binary definition of MDD in terms of their power to detect genetic effects, which was assessed by observing the number of test statistics with more extreme associated p-values ($P < E-04$) in each GWAS. An increase in more extreme p-values was assumed to indicate increased power, provided the mixed linear model is adequately correcting for population and family structure. The top 20 SNPs for each endophenotype were subjected to further bioinformatic analysis, as described in Section 3.3.3.5. utilizing SNAP (SNP Annotation and Proxy Search v2.2.)²⁰⁰, UCSC genome browser²⁰¹, the GeneCards database and PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>). Only putative genes tagged by query or proxy SNPs which explicitly relate to psychiatric phenotypes (and therefore not clinical covariates) are discussed in this Chapter.

5.2.4. Replication in UKBiobank

How well the composite traits in GS:SFHS correlated with the MDD phenotype in UKBiobank (UKB) was determined by a paired-samples sign test and bivariate LD score regression¹²³, as described in Section 4.2.4.

5.3. Results

5.3.1. Testing endophenotype criteria

5.3.1.1. Effect of affection status on trait mean

All composite traits demonstrated a significant effect of group on mean trait score: SoS ($F_{(2,11036)}=898.0$, $p<3.85e-285$), Gen3 ($F_{(2,6250)}=503.0$, $p=2.90e-203$), MDQ ($F_{(2,6340)}=467.9$, $p=2.77e-190$). Regression coefficients for fixed effects were used to assess whether the group effect was being driven by both unaffected relatives and cases. All composite traits were significantly more extreme in both unaffected relatives and cases. Differences in group mean and coefficients for fixed effects are presented in Table 5.3.1.1.

Phenotype	Summary Statistic	Group		
		Unrelated Controls	Unaffected Relatives	Cases
SoS	N	11706	2081	2611
	mean (sd)	-0.26 (1.63)	-0.07 (1.64)	1.38 (1.67)
	$z\beta$	-	0.19 (0.04)	1.53 (0.04)
	Z-score	-	4.70	41.72
Gen3	N	6434	945	1277
	mean (sd)	-0.27 (1.25)	-0.12 (1.27)	1.05 (1.24)
	$z\beta$	-	0.15 (0.05)	1.24 (0.04)
	Z-score	-	3.36	31.3
Gen4	N	6434	945	1277
	mean (sd)	0.29 (1.42)	0.10 (1.41)	-1.15 (1.39)
	$z\beta$	-	-0.20 (0.05)	-1.35 (0.04)
	Z-score	-	-3.85	-30.16

Table 5.3.1.1. Effect of group (case, unaffected relative, unrelated control) on trait mean. N: number of individuals in each subgroup with phenotypic information for that trait; mean (sd): trait mean and standard deviation. $z\beta$: standardized regression coefficient for each affection status group; Z-score: Z-score for each affection status group. Where the corresponding p-value survived multiple testing correction, the Z-score is highlighted in bold.

These results demonstrate that SoS, Gen3 and Gen4 are associated with illness in the general population, more extreme in unaffected relatives than in the general population and therefore state independent. Relative to single

quantitative traits, all composite traits have larger effect sizes ($z\beta$) in unaffected relatives and cases compared to single traits, where the largest effect sized were in SPQ for unaffected relatives, $z\beta(se)=0.12$ (0.04), and in EPQN for cases, $z\beta(se)=0.78$ (0.02). Only SoS had a stronger association ($p<3.85E-285$) than all of its constituent traits.

5.3.1.2. Heritability

5.3.1.2.1. Using pedigree data

All composite traits were moderately heritable. Both Gen3 and Gen4 had heritability estimates, $h^2(se) = 0.30(0.05)$, that were higher than the heritability estimates for any of their constituent single traits, EPQN = $0.24(0.03)$, GHQ = $0.11(0.03)$, MDQ = $0.21(0.05)$ and SPQ = $0.19(0.04)$. The heritability of SoS was a lot closer to the heritability of EPQN than GHQ. Heritability estimates using pedigree data are shown in Table 5.3.1.2.1.

Trait	h^2 (SE)
SoS	0.23 (0.03)
Gen3	0.30 (0.05)
Gen4	0.30 (0.05)

Table 5.3.1.2.1. Narrow-sense heritability (h^2) estimates for MDD, mood, personality and cognitive traits ordered from highest h^2 to lowest, with standard errors for point estimates in parentheses.

5.3.1.2.2. Using GWAA summary statistics

Only SoS had a SNP-based heritability which was significantly greater than zero, $h^2_{SNP}(se) = 0.07$ (0.03). Similar to pedigree based analyses, h^2_{SNP} for SoS was in between h^2_{SNP} estimates for EPQN = 0.09 (0.03) and GHQ = 0.04 (0.03). The ratio of the LD Score intercept to its standard error indicated that no composite trait had a significant intercept (<1.96), suggesting that the mixed linear model used for these analyses is appropriately controlling for confounding due to population structure and the enrichment of families in GS:SFHS. However, the LD Score Ratio, which indicates the proportion of inflation ascribed to causes

other than polygenic heritability, yielded a non-zero estimate for SoS, Ratio(SE) = 0.31 (0.25), suggesting potential confounding. Results from LD Score regression are shown in Table 5.3.1.2.2.

Trait	h^2_{SNP} (SE)	λ_{GC}	$\mu\chi^2$	Intercept (SE)	Ratio (SE)
SoS	0.07 (0.03)	1.035	1.032	1.010 (0.008)	0.31 (0.25)
Gen3	0.05 (0.06)	1.017	1.012	1.003 (0.008)	0.22 (0.61)
Gen4	0.05 (0.05)	1.014	1.011	1.001 (0.007)	0.13 (0.68)

Table 5.3.1.2.2. Output from univariate LD Score regression using summary statistics from GWAS of composite traits in GS:SFHS. h^2 (SE): SNP-based heritability estimate on the liability scale with its standard error; λ_{GC} : genomic control, estimated using $\text{median}(\chi^2)/0.4549$; $\mu\chi^2$: mean chi-squared statistic; Intercept (SE): LD Score regression intercept with its standard error; Ratio (SE): proportion of inflation ascribed to causes other than polygenic heritability with its standard error.

5.3.1.3. Genetic and phenotypic correlation

All composite traits had high, statistically significant genetic correlations with MDD, and moderate phenotypic correlations which all survived multiple testing correction. The genetic correlation between SoS and MDD is between values observed for EPQN ($r_g = 0.54$) and GHQ ($r_g = 0.84$). Gen3 and Gen4 both stronger genetic correlations with MDD than all their constituent traits, except MDQ ($r_g = 0.92$). All composite traits have higher and more statistically significant phenotypic associations with MDD than single quantitative traits. Results are shown in Table 5.3.1.3.

Phenotype	r_g (se)	χ^2	p-value	r_p (se)	p-value
SoS	0.73 (0.11)	80.3	<2.22E-16	0.30 (0.007)	6.82E-05
Gen3	0.90 (0.11)	67.7	<2.22E-16	0.32 (0.009)	6.89E-05
Gen4	-0.90 (0.11)	70.5	<2.22E-16	-0.31 (0.009)	1.14E-04

Table 4.3.1.3. Genetic and phenotypic correlations of quantitative traits with MDD. r_g (se): genetic correlation with MDD and corresponding standard error, χ^2 : chi-squared statistic (from a mixed $0.5(\chi^2) + 0.5(0)$ distribution) corresponding to the LRT between full and null (assuming no additive covariance) models; p-value: p value for r_g ; r_p (se): phenotypic correlation with MDD and corresponding standard error; p-value: p value for r_p .

5.3.2. Performance of composite traits relative to single traits

5.3.2.1. Coheritability

All composite traits have higher coheritability estimates of the liability and observed scale relative to their constituent traits. Both Gen3 and Gen4 have a coheritability on the observed scale which is higher than the observed heritability estimate of MDD. This suggests that Gen3 and Gen4 are better predictors of disease liability than the binary definition of MDD. Coheritability estimates are shown in Table 5.3.2.1.

Phenotype	h_{ij}	h_{ijo}	P-value
MDD	0.44	0.17	-
SoS	0.23	0.14	<2.22E-16
Gen3	0.33	0.20	<2.22E-16
Gen4	0.33	0.20	<2.22E-16

Table 5.3.2.1. Coheritability estimates for quantitative traits and MDD using h^2 estimates for MDD on the liability scale (h_{ij}) and observed scale (h_{ijo}). Traits where p-values survived multiple testing correction ($P \leq 0.05/26 = 1.92E-03$) are highlighted in bold.

5.3.2.2. Polygenic risk score analysis

5.3.2.2.1. Association with MDD PGRS

PGRS for MDD were positively associated with SoS and Gen3, and negatively associated with Gen4 at all P_T . $P_T \leq 1$ and ≤ 0.5 had the strongest associations and explained the highest percentage of variance across composite traits, as observed across binary MDD, EPQN and GHQ (but not MDQ and SPQ). As $P_T \leq 0.5$ explained the most variance in the discovery dataset, this is the P_T which will be summarized here. At this P_T , the variance explained by PGRS for SoS = 0.24%, Gen3 = 0.16%, and Gen4 = 0.13%. However, only associations with SoS and Gen3 survived multiple testing correction ($p = 9.00E-11$ and $p = 9.79E-05$, respectively), with the strongest association being between MDD PGRS and SoS. All effect sizes were small in magnitude ($OR\ 1.04 \leq 1.12$) The variance in composite trait explained by PGRS was smaller ($\leq 0.25\%$) than the variance explained in the PGC MDD mega-analysis (0.6%)⁴⁷, but still 2-3 times greater

than the variance explained in GS:SFHS binary MDD definitions. Compared to the analysis of single traits, there is a modest improvement in variance explained for SoS relative to EPQN (41.2% increase) and GHQ (26.3% increase), and effect sizes for SoS ($z\beta=0.084$) are double that of either constituent trait (EPQN $z\beta=0.041$; GHQ $z\beta=0.043$). Results from all P_T are presented in Table 5.3.2.2.1., with variance explained by PGRS visualized in Figure 5.3.2.2.1.

Trait	PGRS P_T	$z\beta(SE)$	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
SoS	1	0.086 (0.013)	6.69	44.8 (1, 15741)	2.28E-11	0.25
	0.5	0.084 (0.013)	6.49	42.1 (1, 15710)	9.00E-11	0.24
	0.1	0.069 (0.013)	5.33	28.4 (1, 15900)	1.01E-07	0.16
	0.05	0.069 (0.013)	5.34	28.5 (1, 16073)	9.58E-08	0.16
	0.01	0.053 (0.013)	4.09	16.7 (1, 15924)	4.32E-05	0.09
Gen3	1	0.055 (0.014)	4.01	16.0 (1, 8733)	6.25E-05	0.17
	0.5	0.053 (0.014)	3.90	15.2 (1, 8700)	9.79E-05	0.16
	0.1	0.045 (0.014)	3.25	10.5 (1, 8740)	0.001	0.11
	0.05	0.049 (0.014)	3.55	12.6 (1, 8783)	3.91E-04	0.14
	0.01	0.040 (0.014)	2.90	8.4 (1, 8784)	0.004	0.09
Gen4	1	-0.055 (0.015)	-3.56	12.7 (1, 8886)	3.67E-04	0.14
	0.5	-0.053 (0.015)	-3.45	11.9 (1, 8853)	5.66E-04	0.13
	0.1	-0.044 (0.016)	-2.85	8.1 (1, 8892)	0.004	0.09
	0.05	-0.051 (0.015)	-3.30	10.9 (1, 8941)	9.59E-04	0.12
	0.01	-0.042 (0.016)	-2.73	7.5 (1, 8932)	0.006	0.08

Table 5.3.2.2.1. Association with and variance in trait score explained by polygenic risk score for major depressive disorder. P-values which survived multiple testing correction are highlighted in bold.

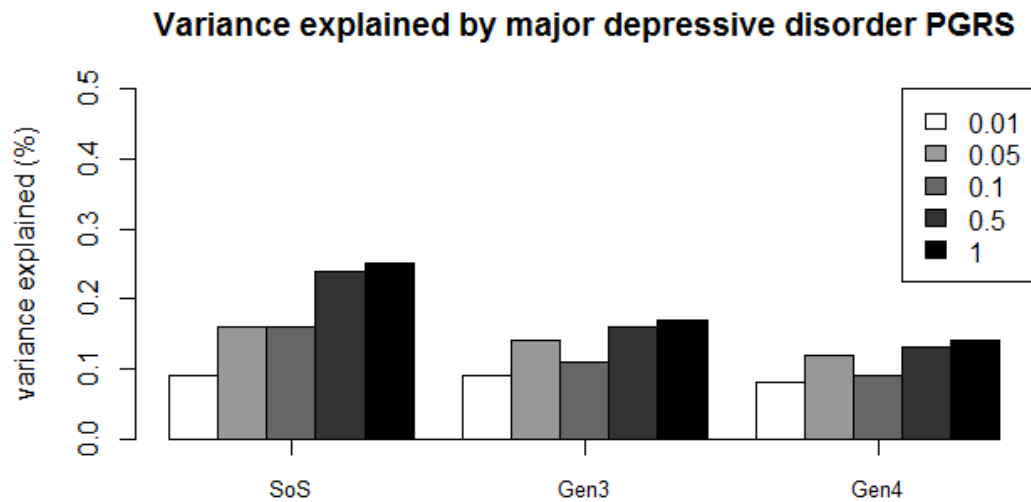


Figure 5.3.2.2.1. Percentage variance explained in candidate endophenotypes in GS:SFHS by MDD PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

5.3.2.2.2. Association with autism PGRS

For all composite traits, both positive and negative associations with autism PGRS were observed at different P_T . Autism PGRS were not significantly associated with any composite trait at any P_T . The marker weights for the autism PGRS were derived from the PGR Cross-Disorder Group analysis, which do not report which P_T explained the most variance in the discovery dataset. Therefore $P_T \leq 1$, which includes all SNPs, will be summarized here. At this P_T , the variance explained by autism PGRS was 0.004% for Gen3, 0.008% for Gen4 and 0.000005% for SoS. These findings are unsurprising, given the lack of association between constituent traits or MDD and autism PGRS. Results from all P_T are presented in Table 5.3.2.2.2., with variance explained by PGRS visualized in Figure 5.3.2.2.2.

Trait	PGRS P _T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
SoS	1	-3.68E-04 (0.013)	-0.03	8.18E-04 (1, 15976)	0.98	4.55E-06
	0.5	0.002 (0.013)	0.18	0.03 (1, 15977)	0.86	1.76E-04
	0.1	0.013 (0.013)	1	1.0 (1, 16033)	0.32	0.006
	0.05	0.014 (0.013)	1.11	1.2 (1, 15973)	0.27	0.007
	0.01	0.004 (0.013)	0.27	0.07 (1, 16094)	0.79	4.15E-04
Gen3	1	-0.008 (0.14)	-0.61	0.4 (1, 8816)	0.54	0.004
	0.5	-0.005 (0.014)	-0.4	0.2 (1, 8813)	0.69	0.002
	0.1	0.008 (0.014)	0.59	0.3 (1, 8768)	0.56	0.004
	0.05	0.007 (0.014)	0.51	0.3 (1, 8751)	0.61	0.003
	0.01	-2.15E-04 (0.014)	-0.02	2.49E-04 (1, 8712)	0.99	2.65E-06
Gen4	1	0.013 (0.015)	0.84	0.7 (1, 8943)	0.40	0.008
	0.5	0.011 (0.015)	0.74	0.5 (1, 8943)	0.46	0.006
	0.1	-0.004 (0.015)	-0.24	0.06 (1, 8912)	0.81	6.07E-04
	0.05	-0.002 (0.015)	-0.12	0.01 (1, 8875)	0.91	1.43E-04
	0.01	2.21E-04 (0.015)	0.01	2.06E-04 (1, 8855)	0.99	2.21E-06

Table 5.3.2.2.2. Association with and variance in trait score explained by polygenic risk score for autism. P-values which survived multiple testing correction are highlighted in bold.

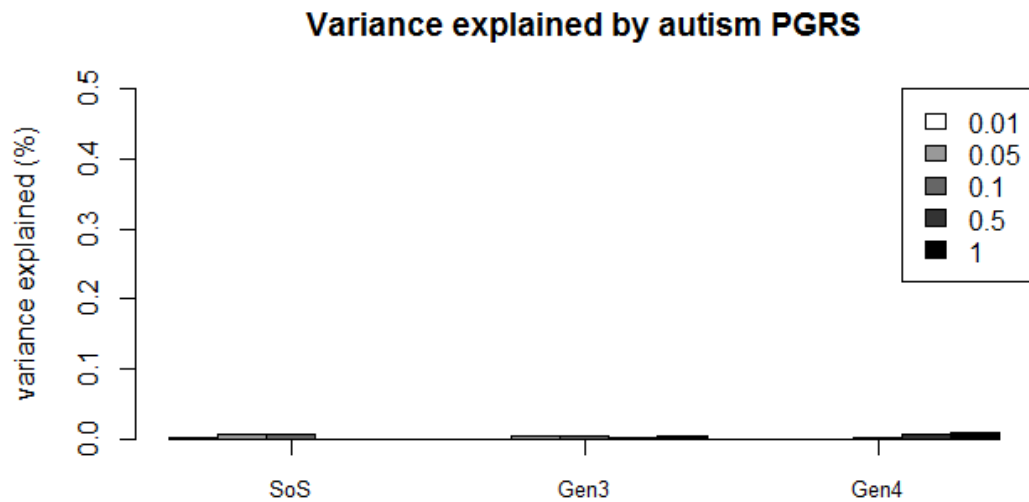


Figure 5.3.2.2.2. Percentage variance explained in candidate endophenotypes in GS:SFHS by autism PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

5.3.2.2.3. Association with bipolar PGRS

PGRS for bipolar disorder were positively associated with SoS and Gen3, and negatively associated with Gen4 in GS:SFHS. PGRS for bipolar disorder were significantly associated with SoS at all P_T thresholds, no P_T for Gen3 or Gen4 survived multiple testing correction. As $P_T \leq 0.5$ explained the most variance in the discovery dataset⁵¹, this is the P_T which will be summarized here. At this P_T , the variance explained by PGRS for bipolar disorder was 0.09% for SoS, 0.06% for Gen3 and 0.04% for Gen4. In previous Chapters, no trait has significantly associated with bipolar disorder, therefore SoS is the only trait presented in this thesis to do so. This is interesting, given the phenotypic and genetic overlap between bipolar disorder and MDD, as discussed previously in Section 3.3.4.3. Results from all P_T are presented in Table 5.3.2.2.3. with variance explained by PGRS visualized in Figure 5.3.2.2.3.

Trait	PGRS P _T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
SoS	1	0.050 (0.013)	3.87	15.0 (1, 16055)	1.07E-04	0.08
	0.5	0.051 (0.013)	3.99	15.9 (1, 16032)	6.65E-05	0.09
	0.1	0.058 (0.013)	4.52	20.5 (1, 16129)	6.13E-06	0.11
	0.05	0.053 (0.013)	4.10	16.8 (1, 16056)	4.09E-05	0.009
	0.01	0.057 (0.013)	4.39	19.3 (1, 15897)	1.12E-05	0.11
Gen3	1	0.033 (0.014)	2.38	5.6 (1, 8920)	0.02	0.06
	0.5	0.033 (0.014)	2.41	5.8 (1, 8919)	0.02	0.06
	0.1	0.040 (0.014)	2.95	8.7 (1, 8918)	0.003	0.09
	0.05	0.037 (0.014)	2.72	7.4 (1, 8881)	0.006	0.08
	0.01	0.034 (0.014)	2.49	6.2 (1, 8830)	0.01	0.07
Gen4	1	-0.030 (0.015)	-1.96	3.8 (1, 9049)	0.05	0.04
	0.5	-0.031 (0.015)	-1.99	4.0 (1, 9051)	0.05	0.04
	0.1	-0.033 (0.015)	-2.15	4.6 (1, 9062)	0.03	0.05
	0.05	-0.032 (0.015)	-2.04	4.2 (1, 9030)	0.04	0.05
	0.01	-0.033 (0.015)	-2.15	4.6 (1, 8978)	0.03	0.05

Table 5.3.2.2.3. Association with and variance in trait score explained by polygenic risk score for bipolar disorder. P-values which survived multiple testing correction are highlighted in bold.

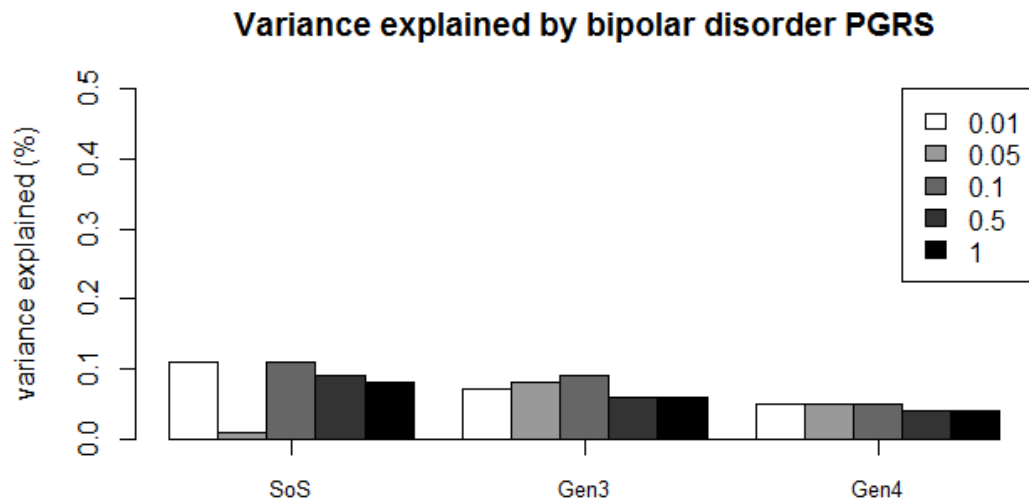


Figure 5.3.2.2.3. Percentage variance explained in candidate endophenotypes in GS:SFHS by bipolar disorder PGRS at five p -value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

5.3.2.2.4. Association with schizophrenia PGRS

PGRS for schizophrenia were positively associated with SoS and Gen3. For Gen4, both positive and negative associations were observed at different P_T . Associations at all P_T thresholds survived multiple testing correction SoS and Gen3. Associations with Gen4 survived multiple testing correction at $P_T P_T \leq 1$. As $P_T \leq 0.1$ explained the most variance in the discovery dataset²³, this is the P_T which will be summarized here. At this P_T , the variance explained by schizophrenia PGRS was 0.22% for Gen3, 0.21% for SoS and 0.17% for Gen4. SoS and Gen3 had a higher percentage of variance explained by schizophrenia PGRS than any of their constituent traits. SoS was also more strongly associated with schizophrenia PGRS than its constituent traits or MDD as a binary trait. Results from all P_T are presented in Table 5.3.2.2.4. with variance explained by PGRS visualized in Figure 5.3.2.2.4.

Trait	PGRS P_T	$z\beta$ (SE)	Z-ratio	F statistic (df, dendif)	P-value	% variance explained
SoS	1	0.077 (0.013)	5.73	32.8 (1, 15692)	1.04E-08	0.20
	0.5	0.078 (0.013)	5.76	33.2 (1, 15683)	8.54E-09	0.20
	0.1	0.079 (0.014)	5.79	33.6 (1, 15723)	7.06E-09	0.21
	0.05	0.076 (0.014)	5.58	31.1 (1, 15719)	2.46E-08	0.20
	0.01	0.082 (0.014)	5.79	33.6 (1, 15811)	6.99E-09	0.23
Gen3	1	0.065 (0.014)	4.58	21.0 (1, 8545)	4.69E-06	0.24
	0.5	0.064 (0.014)	4.50	20.2 (1, 8547)	6.94E-06	0.23
	0.1	0.063 (0.014)	4.37	19.1 (1, 8563)	1.26E-05	0.22
	0.05	0.064 (0.014)	4.44	19.8 (1, 8560)	8.93E-06	0.24
	0.01	0.057 (0.015)	3.83	14.7 (1, 8621)	1.28E-04	0.19
Gen4	1	-0.061 (0.016)	-3.82	14.6 (1, 8703)	1.33E-04	0.17
	0.5	-0.060 (0.016)	-3.74	14.0 (1, 8707)	1.82E-04	0.16
	0.1	0.061 (0.016)	-3.79	14.4 (1, 8723)	1.49E-04	0.17
	0.05	-0.061 (0.016)	-3.75	14.1 (1, 8717)	1.77E-04	0.17
	0.01	-0.054 (0.017)	-3.23	10.4 (1, 8782)	0.001	0.13

Table 5.3.2.2.4. Association with and variance in trait score explained by polygenic risk score for schizophrenia. P-values which survived multiple testing correction are highlighted in bold.

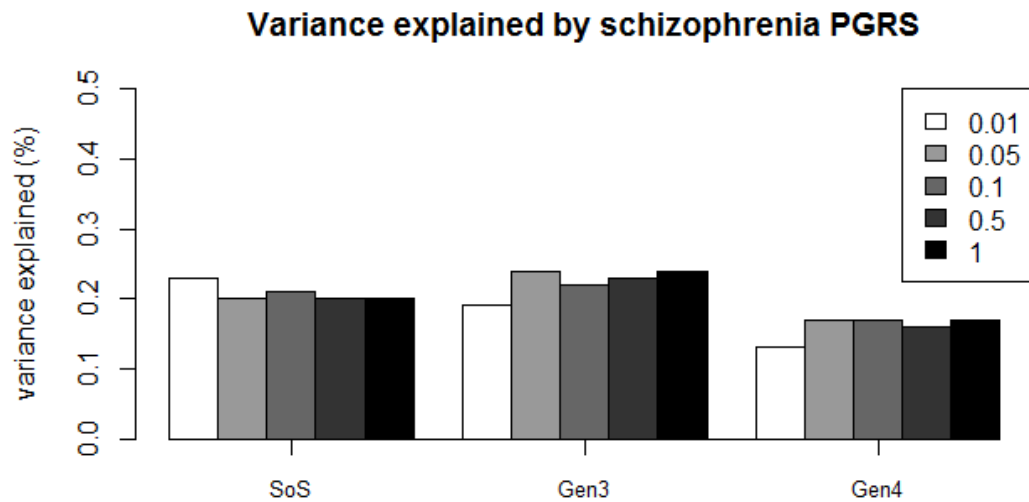


Figure 5.3.2.2.4. Percentage variance explained in candidate endophenotypes in GS:SFHS by schizophrenia PGRS at five p-value thresholds ($P_T \leq 1$, $P_T \leq 0.5$, $P_T \leq 0.1$, $P_T \leq 0.05$, $P_T \leq 0.01$).

5.3.2.3. Genome-wide association analysis

Genome-wide association analysis of SoS, Gen3 and Gen4 yielded no genome-wide significant hits. The ratio of the observed median χ^2 to that expected by chance, denoted λ , indicated that the observed $-\log(P)$ for these traits were slightly inflated to those expected under the null, with λ (SE) values for SoS = 1.033 (9.80e-06), Gen3 = 1.016 (5.60e-06) and Gen4 = 1.014 (4.81e-06). For SoS, as stated in Section 5.3.1.2.2., the ratio from LD Score regression indicated that this inflation could be due to confounding by population and family structure, rather than polygenicity¹⁶⁵, in GS:SFHS. Quantile-quantile plots of observed $-\log(P)$ against expected $-\log(P)$ and Manhattan plots of genomic coordinates against $-\log(P)$ are shown for SoS, Gen3, and Gen4 in Figures 5.3.2.3.1. - 5.2.3.2.4. The 20 most significantly associated SNPs for each composite trait are shown in Tables 5.3.2.3.1. - 5.2.3.2.4. Biological annotation of these SNPs can be found in Tables 5.3.2.3.4. - 5.3.2.3.6.

5.3.2.3.1. Summary statistics from GWAS of SoS

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
1	rs11205387	150482738	A	G	0.27	0.099	0.020	8.01e-07
10	rs10762465	73414206	G	A	0.32	0.087	0.019	6.20e-06
9	rs10976822	8180646	A	G	0.31	0.086	0.020	1.08e-05
2	rs16846962	142475665	A	G	0.04	0.199	0.045	1.12e-05
5	rs4869652	36346515	A	G	0.46	0.080	0.018	1.26e-05
6	rs9918490	56566149	A	G	0.10	0.132	0.030	1.37e-05
10	rs1665688	73417236	A	G	0.31	0.084	0.019	1.54e-05
1	rs2820871	163508344	G	A	0.21	-0.095	0.022	1.84e-05
1	rs2274127	150252984	A	C	0.48	0.078	0.018	1.84e-05
1	rs2147324	150255587	A	G	0.48	0.077	0.018	1.92e-05
1	rs12029359	150432802	A	G	0.21	0.093	0.022	2.22e-05
17	rs7220348	78885858	A	C	0.36	-0.080	0.019	2.22e-05
10	rs10823805	73386821	A	G	0.28	0.085	0.020	2.26e-05
12	rs10770914	22831185	G	A	0.46	-0.077	0.018	2.33e-05
1	rs12046899	150336468	A	C	0.21	0.093	0.022	2.39e-05
9	rs13287379	8229283	A	G	0.19	0.098	0.023	2.67e-05
18	rs1788571	71915587	G	A	0.29	0.083	0.020	2.70e-05
1	rs13294	150484987	A	G	0.41	0.077	0.018	2.79e-05
16	rs8053462	77340081	G	A	0.06	0.155	0.037	2.86e-05
9	rs1781263	8230268	G	A	0.19	0.097	0.023	2.90e-05

Table 5.3.2.3.1. Top 20 SNPs from GWAS of Sum of Scores, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

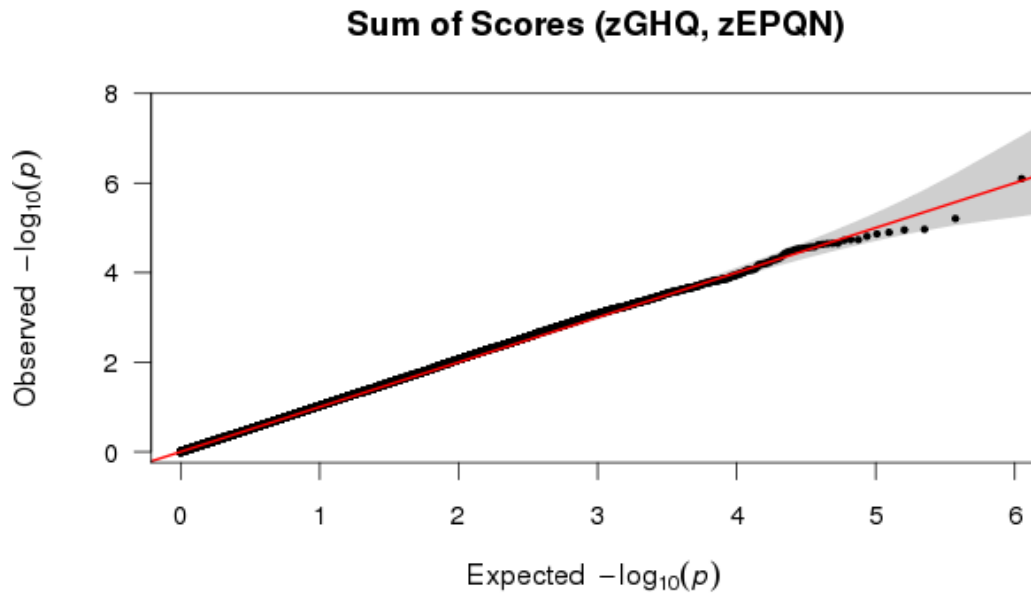


Figure 5.3.2.3.1.A. Quantile-quantile plot for Sum of Scores. $\lambda(\text{SE}) = 1.033$ ($9.80\text{e-}06$).

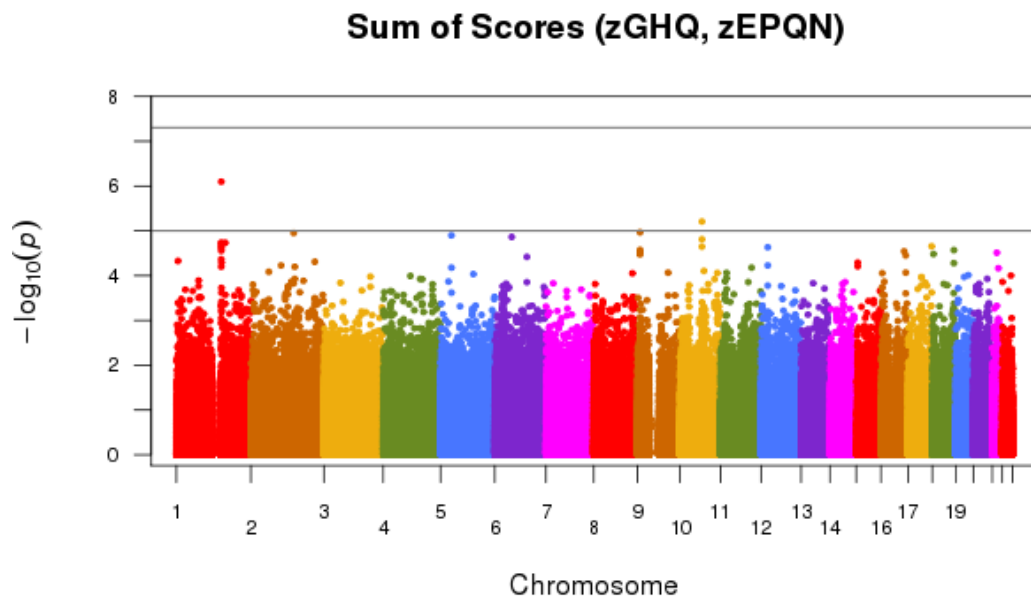


Figure 5.3.2.3.1.B. Manhattan plot for Sum of Scores, showing no genome-wide significant hits.

5.3.2.3.2. Summary statistics from GWAS of Gen3

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
11	rs7941928	2315404	A	G	0.37	0.093	0.020	2.61e-06
16	rs4887826	73535254	G	A	0.07	0.176	0.038	4.47e-06
4	rs6815215	36004792	G	A	0.25	0.102	0.022	4.72e-06
3	rs9875041	7125230	C	A	0.40	-0.089	0.020	6.03e-06
5	rs2161375	172176904	A	G	0.49	0.086	0.019	7.11e-06
1	rs12132546	73841900	A	G	0.12	-0.129	0.029	9.18e-06
2	rs16983681	17768968	A	G	0.05	0.195	0.045	1.14e-05
11	rs1870761	122851563	A	C	0.41	-0.085	0.019	1.36e-05
1	rs4661250	223399328	G	A	0.13	0.124	0.028	1.37e-05
11	rs11218912	122851155	A	G	0.39	-0.085	0.020	1.44e-05
1	rs12139950	73959879	A	G	0.12	-0.125	0.029	1.50e-05
16	rs6564386	77134281	G	A	0.46	-0.082	0.019	1.74e-05
3	rs9311975	7105375	A	G	0.39	-0.084	0.020	1.98e-05
18	rs17084114	68952897	A	G	0.02	0.310	0.073	2.03e-05
14	rs12883888	24161094	A	G	0.47	-0.081	0.019	2.31e-05
15	rs4965905	102147111	A	G	0.10	0.134	0.032	2.33e-05
9	rs10818964	127190340	A	G	0.36	0.084	0.020	2.33e-05
22	rs5756492	37424991	A	G	0.30	0.088	0.021	2.43e-05
9	rs16927294	126985334	C	A	0.28	-0.090	0.021	2.44e-05
1	rs9425020	74167281	G	A	0.11	-0.126	0.030	3.11e-05

Table 5.3.2.3.2. Top 20 SNPs from GWAS of Gen3, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

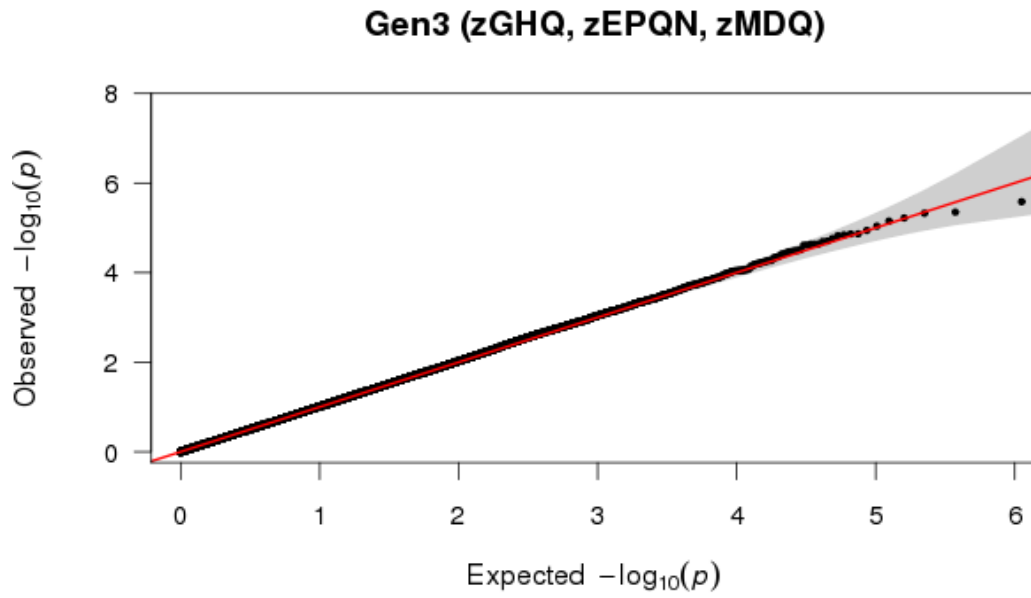


Figure 5.3.2.3.2.A. Quantile-quantile plot for Gen3. $\lambda(\text{SE}) = 1.016$ ($5.60\text{e-}06$).

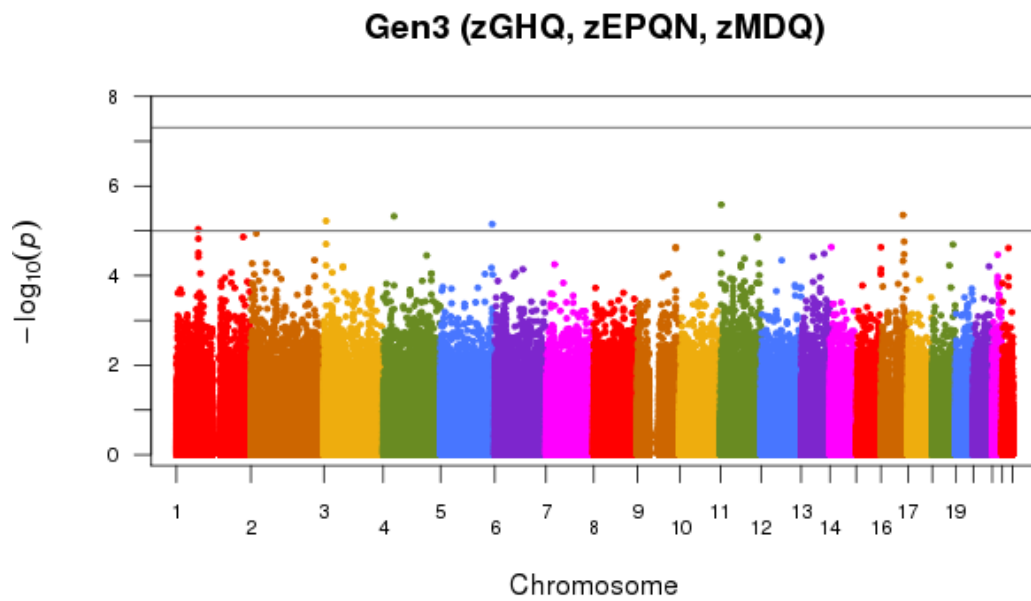


Figure 5.3.2.3.2.B. Manhattan plot for Gen3, showing no genome-wide significant hits.

5.3.2.3.3. Summary statistics from GWAS of Gen4

CHR	SNP	BP	A1	A2	FREQ	B	SE	P
11	rs7941928	2315404	A	G	0.37	-0.110	0.022	9.09e-07
16	rs6564387	77138585	A	G	0.11	0.163	0.034	1.54e-06
5	rs2161375	172176904	A	G	0.49	-0.100	0.022	3.36e-06
16	rs4887826	73535254	G	A	0.07	-0.196	0.043	5.93e-06
15	rs11630119	40665627	A	G	0.21	0.119	0.027	9.27e-06
14	rs12883888	24161094	A	G	0.47	0.094	0.021	1.13e-05
11	rs7944004	2311152	A	C	0.49	0.095	0.022	1.15e-05
11	rs11022157	2322829	A	C	0.35	0.099	0.023	1.33e-05
16	rs6564386	77134281	G	A	0.46	0.092	0.021	1.90e-05
5	rs6555905	170024636	A	G	0.16	-0.126	0.030	2.25e-05
12	rs11177355	69148522	A	G	0.07	-0.179	0.042	2.39e-05
4	rs6815215	36004792	G	A	0.25	-0.106	0.025	2.51e-05
11	rs17139714	80005512	A	G	0.06	-0.187	0.044	2.51e-05
3	rs9875041	7125230	C	A	0.40	0.093	0.022	2.60e-05
2	rs7586869	26190923	G	A	0.42	-0.092	0.022	2.62e-05
9	rs1928481	79646546	C	A	0.22	-0.110	0.026	2.66e-05
2	rs6433223	171461637	G	A	0.16	0.122	0.029	2.66e-05
1	rs11205353	150224505	G	A	0.29	0.100	0.024	3.02e-05
3	rs1456682	147845656	G	A	0.21	-0.111	0.027	3.07e-05
4	rs10029731	140816848	A	G	0.15	0.128	0.031	3.10e-05

Table 5.3.2.3.3. Top 20 SNPs from GWAS of Gen4, showing chromosome (CHR), SNP ID (SNP), base-pair position in kilobases (BP), major allele (A1), minor allele (A2), minor allele frequency (MAF), beta regression coefficient (B) and its corresponding standard error (SE), and association p-value (P).

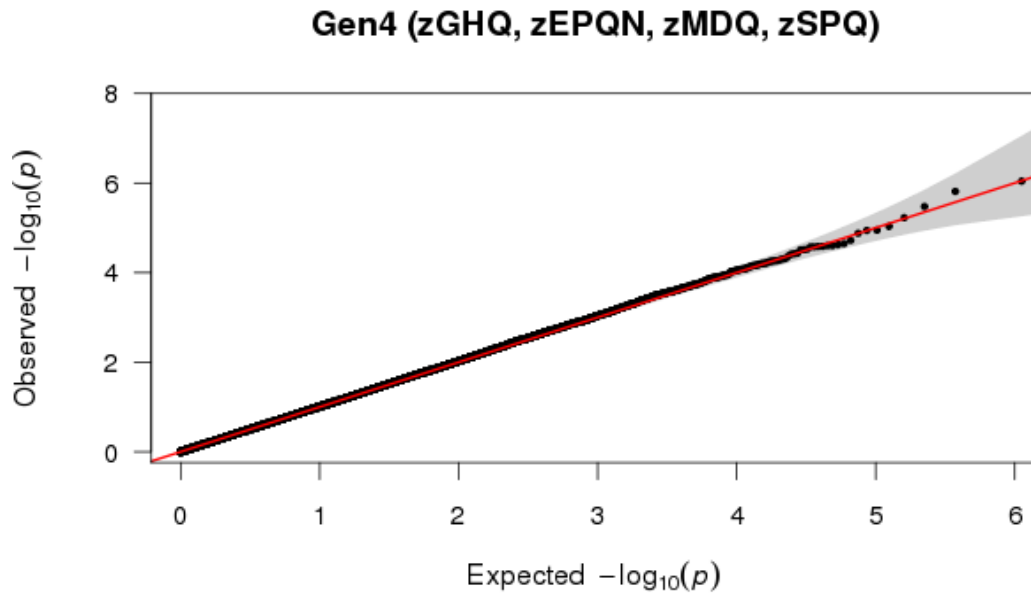


Figure 5.3.2.3.3.A. Quantile-quantile plot for Gen4, $\lambda(\text{SE}) = 1.014$ ($4.81\text{e-}06$).

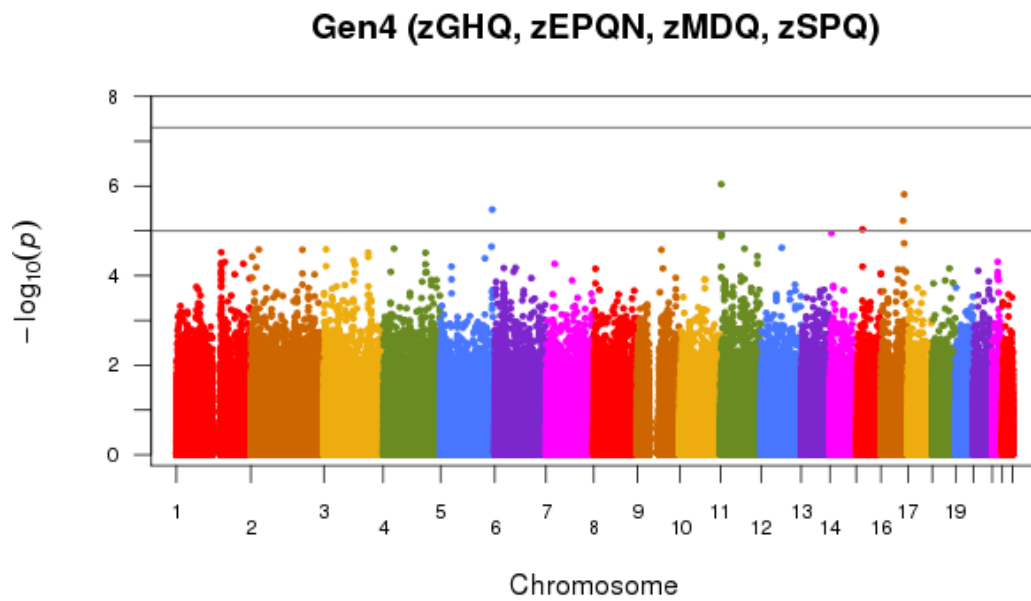


Figure 5.3.2.3.3.B. Manhattan plot for Gen4, showing no genome-wide significant hits.

5.3.2.3.4. Biological annotation of top hits from SoS GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs11205387	-	1	1:148749362	ECM1 (intronic)	enzyme binding; protein C-terminus binding	breast, thyroid, and other cancers; ulcerative colitis
rs11205387	rs12029359	0.65	1:148699426	KIAA0460 (intronic)	-	-
rs11205387	rs1625468	0.54	1:148588110	PRPF3 (intronic)	poly(A) RNA binding; identical protein binding	retinitis pigmentosa
rs11205387	rs3738487	0.52	1:148743685	TARS2 (intronic)	nucleotide binding; ligase activity; aminoacyl-tRNA	Combined oxidative phosphorylation deficiency
rs10762465	-	1	10:73084212	CDH23 (intronic)	calcium ion binding	hearing loss
rs10762465	rs10823805	0.8	10:73056827			
rs10762465	rs1665688	0.79	10:73087242			
rs16846962	-	1	2:142192135	LRP1B (intronic)	calcium ion binding; low-density lipoprotein receptor activity	endocervical carcinoma
rs9918490	-	1	6:56674108	DST (intronic)	calcium ion binding; actin binding	neuropathies; epidermolysis bullosa simplex

Table 5.3.2.3.4. Biological annotation of the top 20 SNPs from GWAS of SoS showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r2	CHR:POS	Gene	Gene function	Disease association
rs9918490	rs9396201	0.84	6:56285532	COL21A1 (intronic)	degradation of ECM; ERK signaling	some cancers
rs2274127	-		1:148519608	C1orf54 (intronic)	-	-
rs2274127	rs2147324	1	1:148522211	C1orf51,MRPS21 (5'UTR)	represses circadian associated transcription; poly(A) RNA binding; structural constituent of ribosome	
rs2274127	rs2298161	0.55	1:148504248	APH1A (intronic)	endopeptidase activity	Alzheimer's Disease
rs7220348	-	1	7:76500453	KIAA1303 (intronic)	binding; protein complex binding	some cancers (via mTOR pathway)
rs10770914	-	1	12:22722452	ETNK1 (intronic) (3'UTR)	transferase activity; ethanolamine kinase activity	leukemia
rs10770914	rs2129613	0.87	12:22731003			
rs1788571	rs1790815	0.92	18:70077756	CYB5A (intronic)	enzyme binding; cytochrome-c oxidase activity	methemoglobinemia
rs8053462	-	1	16:75897582	ADAMTS (intronic)	peptidase activity; metallopeptidase activity	Quebec platelet disorder

Table 5.3.2.3.4. Continued. Biological annotation of the top 20 SNPs from GWAS of SoS.

Four of the 20 highest ranking SNPs from GWAS of SoS were in intergenic regions. Of the 16 remaining SNPs, no query SNP or proxy SNP tagged genes directly linked to psychiatric phenotypes, and therefore will not be explored in greater depth here.

5.3.2.3.5. Biological annotation of top hits from Gen3 GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs7941928	rs2074019	0.68	11:2275443	C11orf21 (3'UTR)	-	beckwith-wiedemann syndrome (bws)
rs7941928	rs10831733	0.64	11:2281016	TSPAN32 (intronic)	immune system; NAD metabolism	bws; some cancers
rs9875041	-	1	3:7125230	GRM7 (intronic)	G-protein coupled receptor activity; PDZ domain binding	depression
rs9875041	rs9311975	0.87	3:7080375			
rs16983681	-	1	2:17768968	VSNL1 (intronic)	calcium ion binding	Alzheimer's disease
rs1870761	-	1	11:122356773	BSX (intronic)	transcription factor activity, sequence-specific DNA binding	Jacobsen syndrome
rs1870761	rs11218912	0.83	11:122356365			
rs1870761	rs2061603	0.53	11:122326244	C11orf63 (intronic)	-	-

Table 5.3.2.3.5. Biological annotation of the top 20 SNPs from GWAS of Gen3 showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r²	CHR:POS	Gene	Gene function	Disease association
rs4661250	-	1	1:221465951	SUSD4 (intronic)	inhibits classical and alternative complementary immune system	-
rs5756492	-	1	22:35754937	MPST (intronic)	thiosulfate and 3-mercaptopyruvate sulfurtransferase activity	mercaptolactate- cysteine disulfiduria
rs5756492	rs5750370	0.53	22:35742913	TST (intronic)	thiosulfate sulfurtransferase activity; 5S rRNA binding	staphyloenterotoxemia
rs16927294	rs2282086	0.69	9:126064894	NEK6 (intronic)	transferase activity; protein tyrosine kinase activity	possibly tumorigenesis

Table 5.3.2.3.5. Continued. Biological annotation of the top 20 SNPs from GWAS of Gen3.

Eleven of the 20 highest ranking SNPs from GWAS of GHQ were in intergenic regions. Of the 9 remaining SNPs, 3 tagged biologically interesting genes and will be discussed here. The 4th highest ranking SNP rs9875041 (also 14th highest in GWAS of Gen4) and 13th highest ranking SNP rs9311975, are intronic variants in the *GRM7* (Metabotropic glutamate receptor 7 precursor) gene. *GRM7* has previously been associated with MDD via linkage analysis in a large study of families with severe, recurrent MDD³⁸ and a smaller study of families of heavy smokers. This gene has also been implicated as a potential candidate gene for MDD via association analysis, yielding a meta-analysis association p-value of $p=1.11E-06$ in a study combining data from three MDD case-control cohorts of European ancestry: STAR*D, GenRED, and GAIN-MDD³¹⁰. Interestingly, this region of the genome was also identified for MDD in GS:SFHS in a GWAS of a continuous measure of MDD created by performing Factor Analysis on SCID questions. However, the highest ranking SNP in this region identified by that analysis, rs162212 ($p=3.10E-06$), was not in LD or within a 500kb window of the SNPs described here. The regional LD plot for rs9875041 and rs9311975 is shown in Figure 5.3.2.3.5.A.

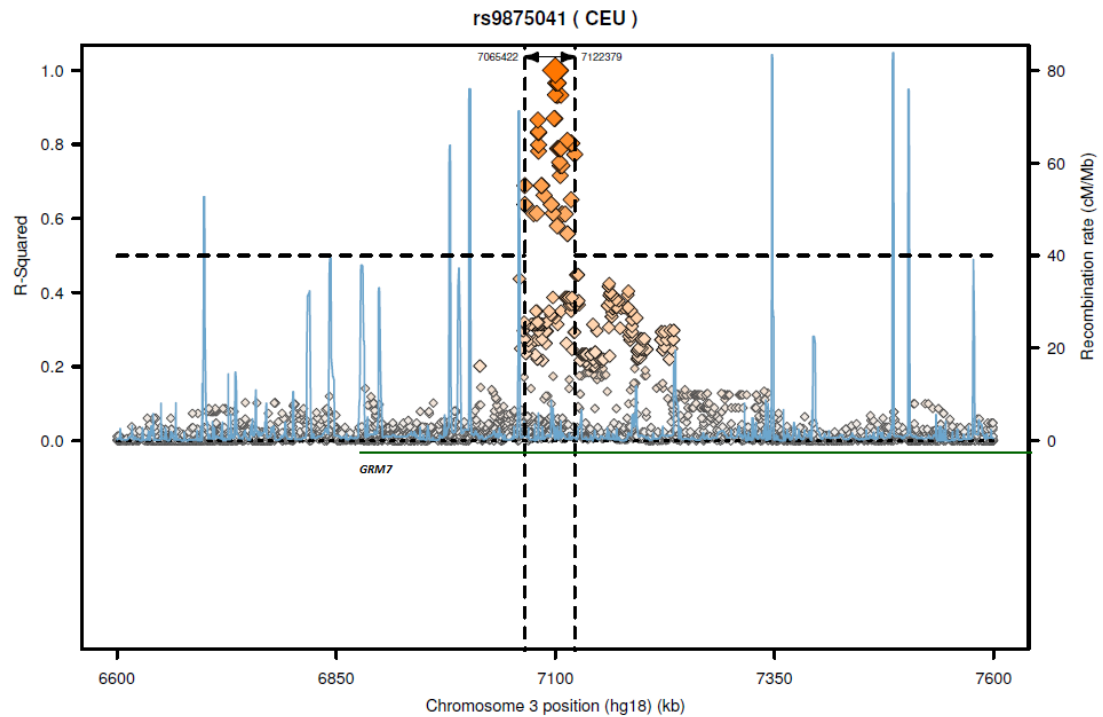


Figure 5.3.2.3.5. Regional LD plot for rs9875041 - ranked 4/20 in GWAS of Gen3, and 14/20 in GWAS of Gen4. This SNP is in high LD ($r^2=0.87$) with rs9311975 - ranked 13/20 in GWAS of Gen3.

The 7th highest ranking SNP, rs16983681, is an intronic variant in the *VSNL1* (Visinin-like protein 1) gene. This gene has shown association with schizophrenia case-control status and improved performance in the Wisconsin Card Sorting Task (WCST) in a sample of 627 cases and 541 controls. The WCST is used to assess frontal cortical function via testing executive functions, such as strategic planning, organized searching, utilizing environmental feedback to shift cognitive sets, directing behavior towards achieving a goal and modulating impulsive responding³¹¹. *VSNL1* was identified via genome-wide linkage as being associated with schizophrenia in the Korean population³¹². Differential gene expression of *VSNL1* and cytoarchitectural abnormalities have been observed in the frontal cortex of schizophrenia and bipolar disorder patients. The gene product of *VSNL1*, VILIP1, has also been identified as a peripheral marker for Alzheimer's disease³¹³. The regional LD plot for rs16983681 is shown in Figure 5.3.2.3.5.B.

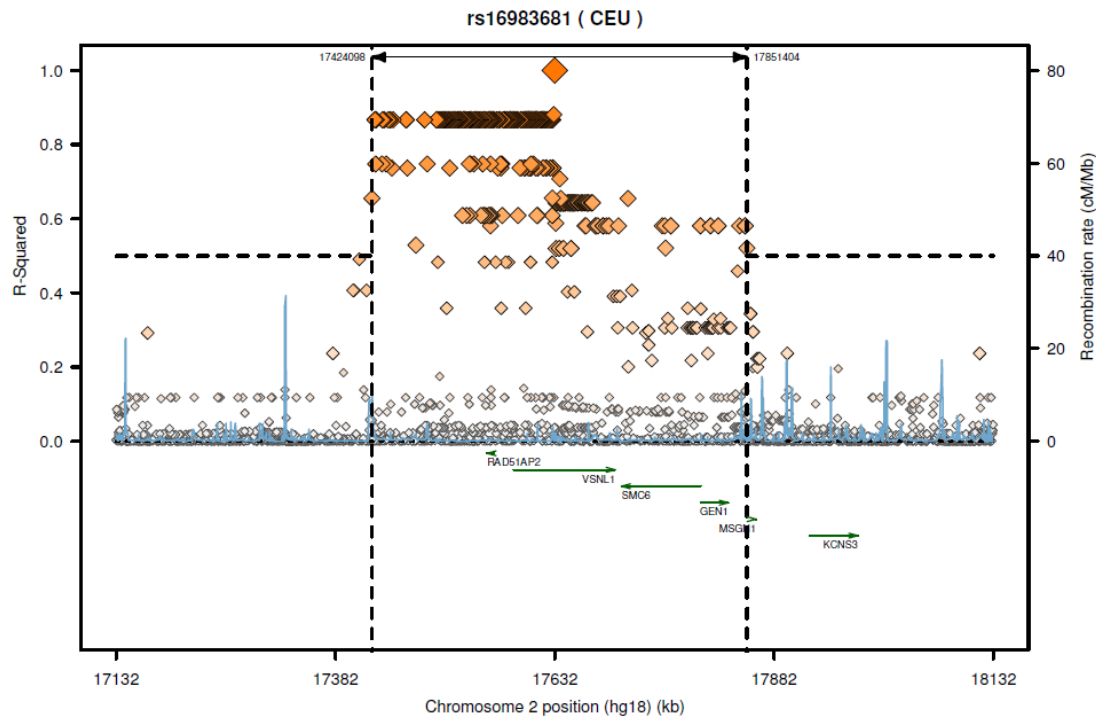


Figure 5.3.2.3.5.B. Regional LD plot for rs16983681 - ranked 7/20 in GWAS of Gen3.

The last SNP that will be discussed in this Section is rs4661250, ranked 9th and an intronic SNP in *SUSD4* (Sushi domain containing protein 4). *SUSD4* is highly expressed in the white matter on oligodendrocytes/axons³¹⁴. White matter integrity has been shown to be reduced in depression patients, with high neuroticism and low extraversion significantly mediated the relationship between fractional anisotropy and depressive symptoms³¹⁵. This is interesting given that neuroticism is one of the traits used to create Gen3. *SUSD4* has also been shown to play a role in inhibiting complementary pathways in the innate immune system³¹⁶. This is interesting, given the vast body of literature seeking to characterize the immune theory of depression and other psychiatric disorders³¹⁷. The regional LD plot for rs4661250 is shown in Figure 5.3.2.3.5.C.

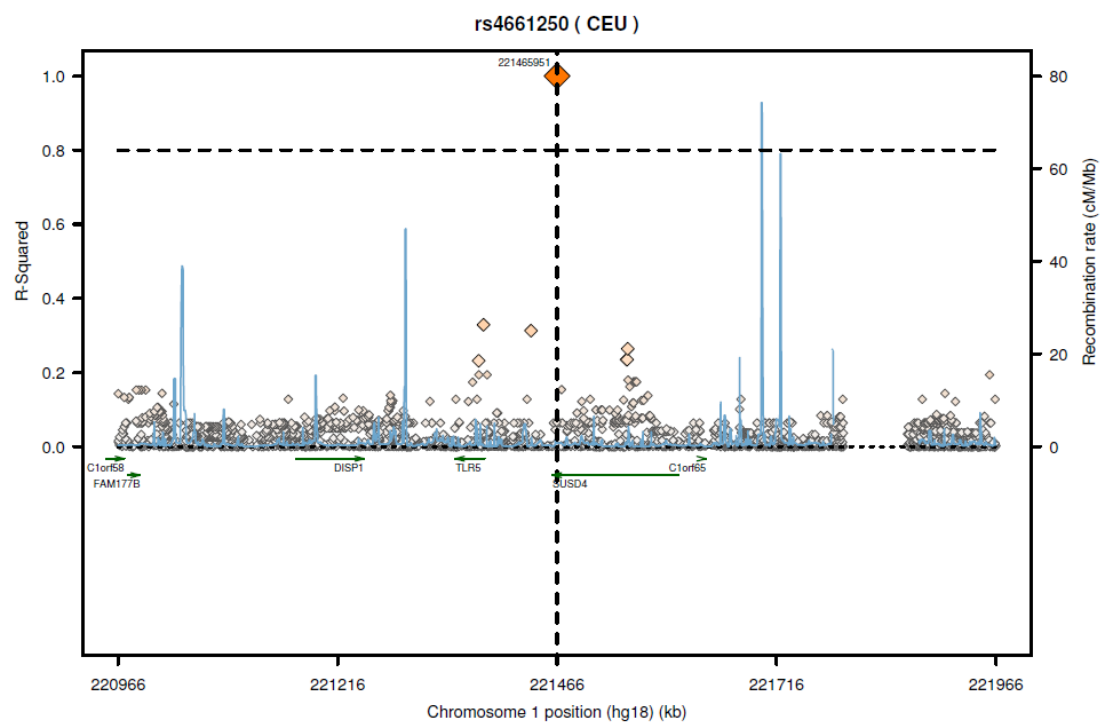


Figure 5.3.2.3.5.C. Regional LD plot for rs4661250 - ranked 9/20 in GWAS of Gen3.

5.3.2.3.6. Biological annotation of top hits from Gen4 GWAS

SNP	Proxy	r ²	CHR:POS	Gene	Gene function	Disease association
rs7941928	rs2074019	0.68	11:2275443	C11orf21 (3'UTR)	-	beckwith-wiedemann syndrome (bws)
rs7941928	rs10831733	0.64	11:2281016	TSPAN32 (intronic)	immune system; NAD metabolism	bws; some cancers
rs11630119	rs7176592	0.74	15:38464289	C15orf23 (intronic) (3'UTR)	metaphase-to-anaphase transition; required for chromosome alignment	-
rs11630119	rs11541641	0.74	15:38473379			
rs11630119	rs875767	0.59	15:38494748	IVD (intronic)	flavin adenine dinucleotide binding; acyl-CoA dehydrogenase activity	isovaleric acidemia; bone degeneration disease
rs6555905	-	1	5:169957214	KCNIP (intronic)	calcium ion binding; potassium channel regulator activity	epithelial myoepithelial carcinoma
rs11177355	-	1	12:67434789	SLC35E3 (intronic)	Putative transporter	-

Table 5.3.2.3.6. Biological annotation of the top 20 SNPs from GWAS of Gen4 showing query SNP ID (SNP), proxy SNP ID (Proxy), linkage disequilibrium squared correlation coefficient between pairs of loci (r²), chromosome and base-pair position (CHR:POS), name of gene and functional annotation in parentheses (Gene), function of gene if known, putative disease association of gene.

SNP	Proxy	r²	CHR:POS	Gene	Gene function	Disease association
rs9875041	-	1	3:7125230	GRM7 (intronic)	G-protein coupled receptor activity; PDZ domain binding	depression
rs7586869	-	1	2:26044427	KIF3C (intronic)	ATPase activity; microtubule motor activity	-
rs6433223	-	1	2:171169883	MYO3B (intronic)	transferase activity; protein tyrosine kinase activity	Bardet-biedl syndrome
rs11205353	rs3850839	0.96	1:148500906	CA14 (intronic)	carbonate dehydratase activity	proximal renal tubular acidosis
rs11205353	rs9126	0.5	1:148526024	MRPS21,C1orf51 (3'UTR)	represses circadian associated transcription; poly(A) RNA binding; structural constituent of ribosome	
rs10029731	-	1	4:141036298	MAML3 (intronic)	transcription coactivator activity	malignant triton tumor

Table 5.3.2.3.6. Continued. Biological annotation of the top 20 SNPs from GWAS of Gen4.

Eleven of the 20 highest ranking SNPs from GWAS of Gen4 were in intergenic regions. Of the 9 remaining SNPs, only one query SNP or proxy SNP tagged genes directly linked to psychiatric phenotypes, rs9875041, and this SNP has already been discussed in the context of high ranking SNPs in the GWAS of Gen3.

5.3.3. Replication in UKBiobank

5.3.3.1. Paired samples sign test

The ORs of SNPs in GS:SFHS and UKB were tested for consistent directionality using SNPs with a p-value of $<1E-04$ from GWAS of the three composite traits (SoS, Gen3, and Gen4) in GS:SFHS to perform a paired samples sign test. This resulted in testing 51 SNPs for SoS, 63 SNPs for Gen3, and 63 SNPs for Gen4. For these SNPs, the replication samples did not produce ORs in the same directions as the discovery analysis more frequently than expected by chance (unadjusted p-value = 0.05). Full results from the paired samples sign test are shown in Table 5.3.3.1.

Trait	UKB definition	nTrials	nSuccesses	Probability of success (CI)	P-value
SoS	broad	51	24	0.47 (0.33, 0.62)	0.78
	narrow		26	0.51 (0.37, 0.65)	1
	strict		26	0.51 (0.37, 0.65)	1
Gen3	broad	63	30	0.48 (0.35, 0.61)	0.80
	narrow		34	0.54 (0.41, 0.67)	0.61
	strict		33	0.52 (0.39, 0.65)	0.80
Gen4	broad	63	36	0.57 (0.44, 0.70)	0.31
	narrow		36	0.57 (0.44, 0.70)	0.31
	strict		35	0.56 (0.42, 0.68)	0.45

Table 5.3.3.1. Paired samples sign test between composite traits in GS:SFHS and UKB MDD definitions, with nTrials indicating the number of SNPs $p < 1E-04$, and nSuccesses indicating how many times ORs were in the same direction in both samples.

5.3.3.2. Bivariate LD Score regression

Results from bivariate LD Score regression of summary statistics from GWAS of composite traits in GS:SFHS and MDD in UKB are shown in Table 5.3.3.2. Broad, narrow and strict definitions of MDD in UKB were positively correlated with SoS and Gen3, and negatively correlated with Gen4 in GS:SFHS. Unlike the binary definition of MDD, the genetic correlation between these traits and MDD in UKB were constrained between -1 and 1. However, no correlation survived multiple testing correction ($p \leq 0.002$). Standard errors are large for all correlations, this could be reduced by using imputed data to increase the number of SNPs used in the analysis.

Trait	UKB definition	rg (SE)	P-value	h2_liab (SE)	h2_int (SE)	gcov_int (SE)
SoS	Broad	0.74 (0.26)	0.004	0.07 (0.03)	1.008 (0.008)	-0.002 (0.005)
	Narrow	0.74 (0.36)	0.04	0.07 (0.03)	1.008 (0.008)	-0.005 (0.006)
	Strict	0.83 (0.29)	0.005	0.07 (0.03)	1.008 (0.008)	-0.005 (0.005)
Gen3	Broad	0.96 (0.66)	0.15	0.05 (0.06)	1.004 (0.008)	0.001 (0.005)
	Narrow	0.88 (0.67)	0.19	0.05 (0.06)	1.004 (0.008)	-0.004 (0.005)
	Strict	0.88 (0.60)	0.14	0.05 (0.06)	1.004 (0.008)	-0.002 (0.005)
Gen4	Broad	-0.85 (0.58)	0.14	0.05 (0.06)	1.002 (0.008)	-0.002 (0.005)
	Narrow	-0.72 (0.60)	0.23	0.05 (0.06)	1.002 (0.008)	0.001 (0.005)
	Strict	-0.69 (0.51)	0.18	0.05 (0.06)	1.002 (0.008)	-1.99e-04 (0.005)

Table 5.3.3.2. Bivariate LD Score regression between the composite traits in GS:SFHS and the broad, narrow and strict MDD phenotype definition in UK Biobank.

5.4. Discussion

5.4.1. Summary of findings

5.4.1.1. Were composite traits better candidate endophenotypes than single traits?

Composite traits had larger effect sizes ($z\beta$) in unaffected relatives and cases compared to single traits. SoS also had a stronger association ($p < 3.85E-285$) with affection status than its constituent traits. Gen3 and Gen4 had higher heritability estimates than their constituent traits, SoS had a heritability similar to EPQN. The genetic correlation between SoS and MDD was between values observed for EPQN and GHQ. Gen3 and Gen4 both stronger genetic correlations with MDD than all their constituent traits, except MDQ. All composite traits have higher and more statistically significant phenotypic associations with MDD than single quantitative traits. Therefore, composite traits generally fulfilled endophenotype criteria in an even more satisfactory way than the single traits which they are derived from.

5.4.1.2. Did composite traits rank more highly than single traits using measures of genetic covariation?

All composite traits had higher coheritability estimates of the liability and observed scale relative to their constituent traits. Both Gen3 and Gen4 have a coheritability on the observed scale which is higher than the observed heritability estimate of MDD, suggesting that these traits are better predictors of disease liability than the binary definition of MDD. No single trait achieved this.

Using variance explained by MDD PGRS as the ranking method, variance explained in SoS and Gen3 were 2-3 times greater than the variance explained in GS:SFHS binary MDD definitions. Compared to the analysis of single traits, there is a modest improvement in variance explained for SoS relative to EPQN (41.2% increase) and GHQ (26.3% increase), and effect sizes for SoS ($z\beta = 0.084$) are double that of either constituent trait (EPQN $z\beta = 0.041$; GHQ $z\beta = 0.043$).

Gen4 was not significantly associated with MDD PGRS at this threshold. Therefore, SoS and Gen3 performed better than or equal to their constituent traits, but Gen4 did not.

5.4.1.3. Did composite traits have increased power to detect genetic effects compared to single traits or binary MDD?

Surprisingly, despite outperforming single traits in endophenotype criteria and ranking methods, composite traits actually performed more poorly in genome-wide association analysis in terms of the number of more extreme test statistics yielded for each trait. Some biologically interesting regions, including the previously implicated *GRM7*, were amongst the highest ranked SNPs, however. Therefore composite traits were more strongly associated with the polygenic component of MDD as evidenced by their superior performance in the MDD PGRS analysis, but this did not correspond to increased success in GWAA.

5.4.2. Limitations

5.4.2.1. Trait selection for composite traits

EPQN, GHQ, MDQ and SPQ were taken forward as the highest ranking endophenotypes to be used in the creation of composite traits. Although MDQ and SPQ had a markedly reduced sample size, about half that of other available measures, their high genetic correlation and ranking in coheritability methods made them an appealing choice. However, in hindsight, it would perhaps have been fruitful to also try a composite trait which used information from EPQN, GHQ and DSC, as DSC generally performed well in analyses in Chapter 4, and has the advantage of a sample size more equivalent to EPQN and GHQ. Another option would be to generate a composite trait using information from EPQN, GHQ and LM2D, as these traits all performed well when ranked by variance explained by MDD PGRS. Although given the lack of success in GWAA of SoS, perhaps this just indicates that these traits are just capturing heterogeneity rather than disease specific polygenicity!

5.4.3. Conclusions

Multivariate methods did not yield increased statistical power relative to univariate methods in this instance. This could be in part due to traits brought forward to generate the composite traits. It would be good to explore different trait selection, especially using imputed data, as the composite trait with the highest number of observations suffered from the same possible confounding as single traits.

Chapter 6: Discussion

6.1. Summary

The aim of this thesis was to assess whether quantitative traits could be used to aid genetic discovery for major depressive disorder (MDD) using data from Generation Scotland: The Scottish Family Health Study (GS:SFHS). To answer this question in full, depression as a binary trait; measures of mood, personality and cognitive ability; and a quantitative composite trait derived from traits genetically correlated with depression were analyzed and compared in terms of statistical power and biological insight.

6.2. Overview of findings

6.2.1. Analysis of MDD as a binary trait

Analysis of MDD as a binary trait generally established that MDD within GS:SFHS is consistent with the MDD literature to date in terms of heritability, prevalence, association with polygenic risk scores and notable lack of genome-wide association hits. However, it was surprising how many biologically plausible regions came out of the top ranking GWAS, and it begs the question of whether GS:SFHS could yield biological insight if better powered, or whether comprehensive literature searching generally gives results. Either way, similar regions of the genome were highly ranked using regional heritability analysis, which has improved statistical power relative to single SNP methods. However, regional heritability analysis has only been performed on undifferentiated MDD, so perhaps performing this technique on more familial subset, or on a larger sample, collaboratively, could yield interesting results. Similarly, genes involved with calcium signalling appeared prevalent throughout biological annotation (across all Chapters) therefore pathway analysis, and utilizing gene ontology databases, could be an interesting route forward in light of these results.

6.2.2. Analysis of quantitative traits

Analyzing the quantitative traits in the context of the endophenotype approach seemed to generally work well. As GS:SFHS has such a wealth of phenotypes, perhaps it would be more conducive to enter a wider range of candidate phenotypes.

6.2.3. Analysis of composite traits

The results from the composite trait analysis were quite surprising, in that pedigree based analyses suggested that these traits were an improvement yet association methods suggested that they were in fact less powerful than single trait methods. As stated previously, it could perhaps prove fruitful to try different composite phenotypes. Also, as extensively polygenic traits do not lend themselves so well to single SNP association analysis, perhaps also trying more powerful gene detection strategies, such as regional heritability, would prove fruitful.

6.3. Methodological remarks and limitations

6.3.1. Model selection

6.3.1.1. Mixed linear model

One of the main limitations of this project was trying to compare quantitative traits to binary traits, running all analyses on the linear scale. Whilst programs which run analyses, such as association, on the liability scale are very time consuming, if this was decided upon at the project outset, running all traits on liability scale could be achievable and perhaps yield more valid comparisons.

6.3.3.1. Use of genotyped data

Although imputed data has its pitfalls, as PGRS and replication attempts in this thesis used external data, imputed data could be very useful in providing a better sample overlap for these analyses.

6.4. Implications for the future

6.4.1. The endophenotype concept in psychiatric genetics

Candidate endophenotypes for psychiatric traits are often not any simpler in terms of genetic architecture. Consequently the same challenges that make psychiatric genetics difficult, namely those pertaining to statistical power and heterogeneity, are likely to make endophenotype genetics difficult. A lot of explanatory power is lost by binary classifications (case/control) of quantitative phenotypes (liability to disease). Therefore, although statistical power is still problematic for quantitative complex traits, such traits do have increased discriminatory power to detect genetic effects relative to binary classifications of disease when the prevalence equals the proportion of cases⁹⁶, such as MDD in GS:SFHS^{140,142}. However, endophenotypes do not always get us closer to genuine understanding of the illness/disorder, especially in the absence of cellular traits, longitudinal data, tests of confounders. Identifying quantitative traits which could act as proximal measures of binary outcome could still prove useful in the future, in particular in combination with other methods of improving statistical power.

6.4.2. Other methods of improving statistical power

6.4.2.1. Redefining the depressed phenotype

Redefining the depressed phenotype is where a lot of effort is going just now, with efforts to reduce heterogeneity through stratification of the disorders into subtypes. This strategy has yielded results, in the previously mentioned CONVERGE study, and also by refining the phenotype using age at disease onset as the stratifier³¹⁸.

6.4.2.2. Increasing the sample size

Since performing the analysis for this thesis, 15 genome-wide significant loci have been identified for depression in a study utilizing self-report data from

75,607 individuals reporting clinical diagnosis of depression and 231,747 individuals reporting no history of depression through 23andMe¹⁹⁸. The Manhattan plot from this study is shown in Figure 6.4.2.2.

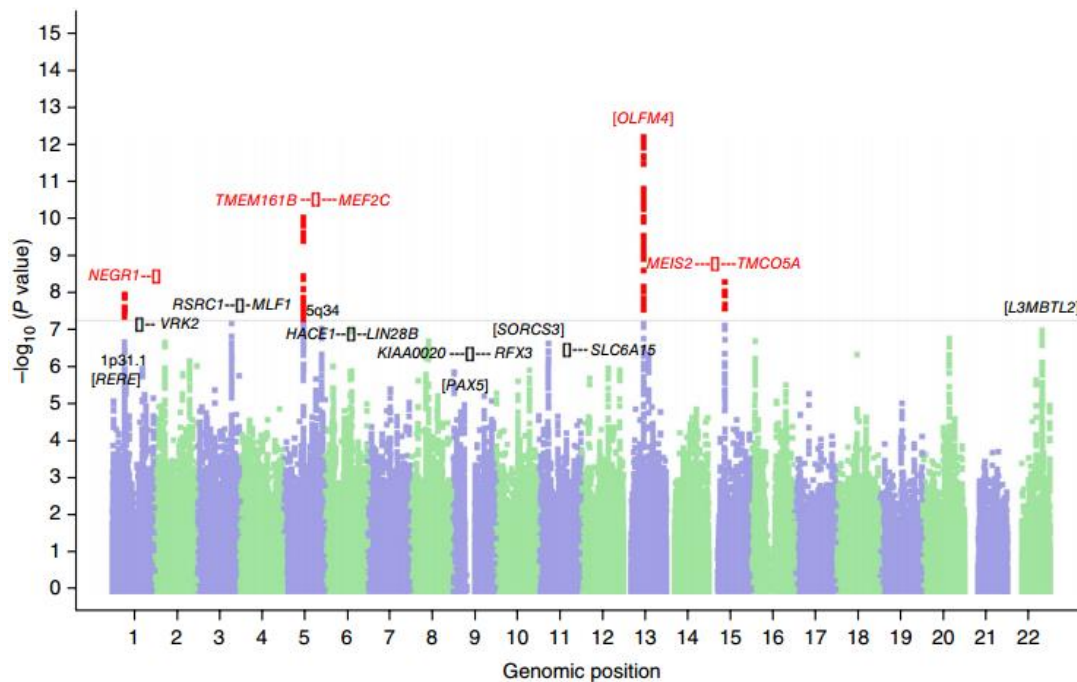


Figure 6.4.2.2. Manhattan plot from the largest available GWAS of MDD showing genetic variants grouped by chromosome and within-chromosome position along the X axis, and the association each variant with case/control status, given as $-\log_{10}$ of the p-value¹⁹⁸.

This study is an impressive feat in terms of sample size. However, it does not do anything to address disease heterogeneity, which in my mind is the biological problem that needs urgently tackled to get us closer to clinical utility.

6.5. Concluding remarks

Since I started this project in 2012, psychiatric genetics has become one of the most exciting and rapidly expanding areas in medical genetics. I have often referred to the race to "solve depression" as being akin to the moon landing. There have been many years of studies telling us what depression is not, I am hopeful that the next phase of collaborative research can tell us more about what depression is.

References

1. Kessler, R. C. *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**, 3095–3105 (2003).
2. WHO. GLOBAL HEALTH ESTIMATES 2014 SUMMARY TABLES: YLD BY CAUSE, AGE AND SEX, 2000-2012. *Glob. Summ. Estim. Years Lost due to Disabil.* 2000-2012. (2014). at <http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html>
3. Bloom, D. E. *et al.* The Global Economic Burden of Noncommunicable Diseases. *PGDA Work. Pap.* (2012). at <<http://ideas.repec.org/p/gdm/wpaper/8712.html>>
4. World Health Organization. Mental Health Action Plan 2013-2020. *WHO Libr. Cat. DataLibrary Cat. Data* 1–44 (2013). at <http://apps.who.int/iris/bitstream/10665/89966/1/9789241506021_eng.pdf>
5. Corey-Lisle, P. K., Nash, R., Stang, P. & Swindle, R. Response, partial response, and nonresponse in primary care treatment of depression. *Arch. Intern. Med.* **164**, 1197–204 (2004).
6. Saito, M. *et al.* Evaluation of the DSM-IV and ICD-10 criteria for depressive disorders in a community population in Japan using item response theory. *Int. J. Methods Psychiatr. Res.* **19**, 211–22 (2010).
7. APA. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Pub, 2013). at <<https://books.google.com/books?hl=en&lr=&id=-JivBAAAQBAJ&pgis=1>>
8. Bromet, E. *et al.* Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med.* **9**, 90 (2011).
9. Kessler, R. C. & Bromet, E. J. The epidemiology of depression across cultures. *Annu. Rev. Public Health* **34**, 119–38 (2013).
10. Weissman, M. M. *et al.* Cross-National Epidemiology of Major Depression and Bipolar Disorder. *JAMA J. Am. Med. Assoc.* **276**, 293–299 (1996).
11. Kendler, K. S., Kuhn, J. & Prescott, C. A. The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *Am. J. Psychiatry* **161**, 631–6 (2004).
12. Brown, G. & Harris, T. *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. (Tavistock: London., 1978).
13. Kessler, R. C. The effects of stressful life events on depression. *Annu. Rev. Psychol.* **48**, 191–214 (1997).
14. Kendler, K. S., Karkowski, L. M. & Prescott, C. A. Causal relationship between stressful life events and the onset of major depression. *Am. J. Psychiatry* **156**, 837–41 (1999).
15. Nanni, V., Uher, R. & Danese, A. Childhood Maltreatment Predicts Unfavorable Course of Illness and Treatment Outcome in Depression: A Meta-Analysis. *Am. J. Psychiatry* **169**, 141–151 (2012).
16. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic epidemiology of major

- depression: review and meta-analysis. *Am. J. Psychiatry* **157**, 1552–62 (2000).
17. Falconer, D. & Mackay, T. F. C. *Introduction to quantitative genetics*. (Harlow, England; New York: Prentice Hall, 1996).
 18. Flint, J. & Kendler, K. S. The Genetics of Major Depression. *Neuron* **81**, 484–503 (2014).
 19. Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–94 (2013).
 20. Kendler, K. S., Gardner, C. O., Neale, M. C. & Prescott, C. a. Genetic risk factors for major depression in men and women: similar or different heritabilities and same or partly distinct genes? *Psychol. Med.* **31**, 605–616 (2001).
 21. Kendler, K. S., Gatz, M., Gardner, C. O. & Pedersen, N. L. A Swedish national twin study of lifetime major depression. *Am. J. Psychiatry* **163**, 109–114 (2006).
 22. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
 23. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
 24. Bassett, A. S. *et al.* The schizophrenia phenotype in 22q11 deletion syndrome. *Am. J. Psychiatry* **160**, 1580–6 (2003).
 25. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–9 (2009).
 26. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
 27. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–69 (2008).
 28. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–3 (1993).
 29. Lifton, R. P. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lect.* **100**, 71–101
 30. Porteous, D. J. *et al.* DISC1 as a genetic risk factor for schizophrenia and related major mental illness: response to Sullivan. *Mol. Psychiatry* **19**, 141–3 (2014).
 31. Bertram, L., Lill, C. M. & Tanzi, R. E. The genetics of Alzheimer disease: back to the future. *Neuron* **68**, 270–81 (2010).
 32. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–56 (2000).
 33. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–7 (1996).
 34. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**, 1590–605 (2008).
 35. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).
 36. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.* **11**, 2417–23

- (2002).
37. Lesch, K. P. Gene-environment interaction and the genetics of depression. in *J. Psychiatry Neurosci.* **29**, 174–184 (2004).
 38. Breen, G. *et al.* A genome-wide significant linkage for severe depression on chromosome 3: The Depression Network study. *Am. J. Psychiatry* **168**, 840–847 (2011).
 39. Zubenko, G. S. *et al.* Genome-wide linkage survey for genetic loci that influence the development of depressive disorders in families with recurrent, early-onset, major depression. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **123B**, 1–18 (2003).
 40. Camp, N. J. *et al.* Genome-wide linkage analyses of extended Utah pedigrees identifies loci that influence recurrent, early-onset major depression and anxiety disorders. *Am. J. Med. Genet. - Neuropsychiatr. Genet.* **135 B**, 85–93 (2005).
 41. Holmans, P. *et al.* Genomewide significant linkage to recurrent, early-onset major depressive disorder on chromosome 15q. *Am. J. Hum. Genet.* **74**, 1154–67 (2004).
 42. Holmans, P. *et al.* Genetics of recurrent early-onset major depression (GenRED): Final genome scan report. *Am. J. Psychiatry* **164**, 248–258 (2007).
 43. Levinson, D. F. *et al.* Genetics of recurrent early-onset major depression (GenRED): significant linkage on chromosome 15q25-q26 after fine mapping with single nucleotide polymorphism markers. *Am. J. Psychiatry* **164**, 259–64 (2007).
 44. Middeldorp, C. M. *et al.* Suggestive linkage on chromosome 2, 8, and 17 for lifetime major depression. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **150B**, 352–8 (2009).
 45. Schol-Gelok, S. *et al.* A genome-wide screen for depression in two independent Dutch populations. *Biol. Psychiatry* **68**, 187–96 (2010).
 46. Wray, N. R. *et al.* Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol. Psychiatry* **17**, 36–48 (2012).
 47. Ripke, S. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
 48. Hemani, G. *et al.* Inference of the genetic architecture underlying bmi and height with the use of 20,240 sibling pairs. *Am. J. Hum. Genet.* **93**, 865–875 (2013).
 49. consortium, C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588 (2015).
 50. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
 51. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–983 (2011).
 52. McGuffin, P., Katz, R., Watkins, S. & Rutherford, J. A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Arch. Gen. Psychiatry* **53**, 129–36 (1996).

53. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–89 (2010).
54. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
55. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–7 (2010).
56. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42**, 973–7 (2010).
57. Major Depressive Disorder Working Group of the Psychiatric, G. C. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* **18**, 497–511 (2013).
58. Lubke, G. H. *et al.* Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol. Psychiatry* **72**, 707–9 (2012).
59. Brown, G. W., Bifulco, A. & Harris, T. O. Life events, vulnerability and onset of depression: some refinements. *Br. J. Psychiatry* **150**, 30–42 (1987).
60. Kendler, K. S. *et al.* Stressful life events, genetic liability, and onset of an episode of major depression in women. *Am. J. Psychiatry* **152**, 833–842 (1995).
61. SILBERG, J. Genetic moderation of environmental risk for depression and anxiety in adolescent girls. *Br. J. Psychiatry* **179**, 116–121 (2001).
62. Angold, A., Prendergast, M., Cox, A. & Al, E. The Child and Adolescent Psychiatric Assessment (CAPA). *Psychol. Med.* **25**, 739–753 (1995).
63. McClelland, G. H. & Judd, C. M. Statistical difficulties of detecting interactions and moderator effects. *Psychol. Bull.* **114**, 376–90 (1993).
64. Uher, R. Gene-environment interactions in severe mental illness. *Front. psychiatry* **5**, 48 (2014).
65. Kim, J. M. *et al.* Interactions Between Life Stressors and Susceptibility Genes (5-HTTLPR and BDNF) on Depression in Korean Elders. *Biol. Psychiatry* **62**, 423–428 (2007).
66. Brown, G. W. *et al.* Functional polymorphism in the brain-derived neurotrophic factor gene interacts with stressful life events but not childhood maltreatment in the etiology of depression. *Depress. Anxiety* **31**, 326–34 (2014).
67. Bradley, R. G. *et al.* Influence of child abuse on adult depression: moderation by the corticotropin-releasing hormone receptor gene. *Arch. Gen. Psychiatry* **65**, 190–200 (2008).
68. Polanczyk, G. *et al.* Protective effect of CRHR1 gene variants on the development of adult depression following childhood maltreatment: replication and extension. *Arch. Gen. Psychiatry* **66**, 978–85 (2009).
69. Jokela, M. *et al.* Serotonin receptor 2A gene and the influence of childhood maternal nurturance on adulthood depressive symptoms. *Arch Gen Psychiatry* **64**, 356–360 (2007).
70. Cicchetti, D., Rogosch, F. A. & Sturge-Apple, M. L. Interactions of child maltreatment and serotonin transporter and monoamine oxidase A

- polymorphisms: depressive symptomatology among adolescents from low socioeconomic status backgrounds. *Dev. Psychopathol.* **19**, 1161–80 (2007).
71. Melas, P. A. *et al.* Genetic and epigenetic associations of MAOA and NR3C1 with depression and childhood adversities. *Int. J. Neuropsychopharmacol.* **16**, 1513–28 (2013).
 72. Bet, P. M. *et al.* Glucocorticoid receptor gene polymorphisms and childhood adversity are associated with depression: New evidence for a gene-environment interaction. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **150B**, 660–9 (2009).
 73. Caspi, A. *et al.* Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301**, 386–389 (2003).
 74. Karg, K., Burmeister, M., Shedden, K. & Sen, S. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. *Arch. Gen. Psychiatry* **68**, 444–54 (2011).
 75. Brown, G. W. *et al.* Serotonin transporter length polymorphism, childhood maltreatment, and chronic depression: a specific gene-environment interaction. *Depress. Anxiety* **30**, 5–13 (2013).
 76. Fergusson, D. M., Horwood, L. J., Miller, A. L. & Kennedy, M. A. Life stress, 5-HTTLPR and mental disorder: findings from a 30-year longitudinal study. *Br. J. Psychiatry* **198**, 129–35 (2011).
 77. Duncan, L. E. & Keller, M. C. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am. J. Psychiatry* **168**, 1041–9 (2011).
 78. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).
 79. Maier, R. *et al.* Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **96**, 283–94 (2015).
 80. Cross-Disorder Group of the Psychiatric Genomics, C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* (2013). doi:10.1016/S0140-6736(12)62129-1
 81. Hamshere, M. L. *et al.* Polygenic dissection of the bipolar phenotype. *Br. J. Psychiatry* **198**, 284–8 (2011).
 82. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19**, 1017–24 (2014).
 83. Whalley, H. C. *et al.* Dissection of Major Depressive Disorder using polygenic risk scores for Schizophrenia in two independent cohorts. *unpublished*
 84. Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955 (2015).
 85. Hamshere, M. & Langley, K. High loading of polygenic risk for ADHD in children with comorbid aggression. *Am. J. ...* 909–916 (2013). doi:10.1176/appi.ajp.2013.12081129
 86. McIntosh, A. M. *et al.* Polygenic risk for schizophrenia is associated with

- cognitive change between childhood and old age. *Biol. Psychiatry* **73**, 938–943 (2013).
87. Middeldorp, C. M. *et al.* The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data. *Transl. Psychiatry* **1**, e50 (2011).
 88. Bois, C. *et al.* Polygenic risk for schizophrenia is associated with increased gyrification in a familial high risk cohort. *unpublished*
 89. Whalley, H. C. *et al.* Polygenic risk and white matter integrity in individuals at high risk of mood disorder. *Biol. Psychiatry* **74**, 280–286 (2013).
 90. Whalley, H. C. *et al.* The influence of polygenic risk for bipolar disorder on neural activation assessed using fMRI. *Transl. Psychiatry* **2**, e130 (2012).
 91. Whalley, H. C. *et al.* Impact of cross-disorder polygenic risk on frontal brain activation with specific effect of schizophrenia risk. *Schizophr. Res.* **161**, 484–489 (2015).
 92. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–15 (2013).
 93. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–8 (2007).
 94. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2014).
 95. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–46 (2014).
 96. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34**, 254–257 (2010).
 97. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
 98. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–9 (2010).
 99. Wray, N. R. & Maier, R. Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability. *Curr. Epidemiol. Reports* **1**, 220–227 (2014).
 100. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
 101. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–74 (2009).
 102. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109**, 1193–1198 (2012).
 103. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
 104. Chanock, S. J. *et al.* Replicating genotype-phenotype associations. *Nature*

- 447**, 655–60 (2007).
105. Fried, E. I. & Nesse, R. M. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J. Affect. Disord.* **172**, 96–102 (2015).
 106. Lyons, M. J. *et al.* A registry-based twin study of depression in men. *Arch. Gen. Psychiatry* **55**, 468–72 (1998).
 107. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13**, 72 (2015).
 108. BECK, A. T., WARD, C. H., MENDELSON, M., MOCK, J. & ERBAUGH, J. An inventory for measuring depression. *Arch. Gen. Psychiatry* **4**, 561–71 (1961).
 109. Devins, G. M. *et al.* Measuring depressive symptoms in illness populations: Psychometric properties of the Center for Epidemiologic Studies Depression (CES-D) Scale. *Psychol. Health* **2**, 139–156 (1988).
 110. Dorogatis, L. R. Symptom Checklist-90-R Administration, Scoring, and Procedures Manual (3rd Edition). *Natl. Comput. Syst. Inc, Minneapolis, MN* (1994).
 111. Jang, K. L., Livesley, W. J., Taylor, S., Stein, M. B. & Moon, E. C. Heritability of individual depressive symptoms. *J. Affect. Disord.* **80**, 125–33 (2004).
 112. Lux, V. & Kendler, K. S. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol. Med.* **40**, 1679–90 (2010).
 113. Fried, E. I., Nesse, R. M., Zivin, K., Guille, C. & Sen, S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychol. Med.* **44**, 2067–76 (2014).
 114. Regier, D. A. *et al.* DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).
 115. Kendler, K. S., Aggen, S. H. & Neale, M. C. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. *JAMA psychiatry* **70**, 599–607 (2013).
 116. Kendler, K. S. *et al.* The structure of the genetic and environmental risk factors for six major psychiatric disorders in women. Phobia, generalized anxiety disorder, panic disorder, bulimia, major depression, and alcoholism. *Arch. Gen. Psychiatry* **52**, 374–83 (1995).
 117. Hammen, C. Stress and depression. *Annu. Rev. Clin. Psychol.* **1**, 293–319 (2005).
 118. Barth, J., Schumacher, M. & Herrmann-Lingen, C. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. *Psychosom. Med.* **66**, 802–13 (2004).
 119. Ayerbe, L., Ayis, S., Crichton, S. L., Rudd, A. G. & Wolfe, C. D. A. Explanatory factors for the increased mortality of stroke patients with depression. *Neurology* **83**, 2007–12 (2014).
 120. Satin, J. R., Linden, W. & Phillips, M. J. Depression as a predictor of disease progression and mortality in cancer patients: a meta-analysis. *Cancer* **115**, 5349–61 (2009).

121. Anderson, R. J., Freedland, K. E., Clouse, R. E. & Lustman, P. J. The prevalence of comorbid depression in adults with diabetes: a meta-analysis. *Diabetes Care* **24**, 1069–1078 (2001).
122. Moretti, R., Torre, P., Antonello, R. M., Cazzato, G. & Bava, A. Depression and Alzheimer's disease: Symptom or comorbidity? *Am. J. Alzheimers. Dis. Other Dement.* **17**, 338–344 (2002).
123. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–41 (2015).
124. Glahn, D. C. *et al.* High dimensional endophenotype ranking in the search for major depression risk genes. *Biol. Psychiatry* **71**, 6–14 (2012).
125. Liu, J., Pei, Y., Papasian, C. J. & Deng, H.-W. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet. Epidemiol.* **33**, 217–27 (2009).
126. Gottesman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* **160**, 636–645 (2003).
127. Glahn, D. C. *et al.* Arguments for the sake of endophenotypes: Examining common misconceptions about the use of endophenotypes in psychiatric genetics. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **165**, 122–130 (2014).
128. Dick, D. M. *et al.* Endophenotypes successfully lead to gene identification: results from the collaborative study on the genetics of alcoholism. *Behav. Genet.* **36**, 112–26 (2006).
129. Hinrichs, A. L. *et al.* Functional variant in a bitter-taste receptor (hTAS2R16) influences risk of alcohol dependence. *Am. J. Hum. Genet.* **78**, 103–11 (2006).
130. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
131. Sing, C. F., Stengård, J. H. & Kardia, S. L. R. Genes, environment, and cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* **23**, 1190–6 (2003).
132. Comuzzie, A. G. *et al.* A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. *Nat. Genet.* **15**, 273–6 (1997).
133. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
134. Mitchell, B. D. *et al.* Genes influencing variation in serum osteocalcin concentrations are linked to markers on chromosomes 16q and 20q. *J. Clin. Endocrinol. Metab.* **85**, 1362–6 (2000).
135. Kammerer, C. M. *et al.* Quantitative trait loci on chromosomes 2p, 4p, and 13q influence bone mineral density of the forearm and hip in Mexican Americans. *J. Bone Miner. Res.* **18**, 2245–52 (2003).
136. Kiel, D. P. *et al.* Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med. Genet.* **8 Suppl 1**, S14 (2007).
137. Hasler, G., Drevets, W. C., Manji, H. K. & Charney, D. S. Discovering endophenotypes for major depression. *Neuropsychopharmacology* **29**,

- 1765–81 (2004).
138. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med. Genet.* **7**, 74 (2006).
 139. www.generationscotland.org/. www.generationscotland.org/. at <www.generationscotland.org/>
 140. 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).
 141. First, M. B. *et*, Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV)*. for *DSMIV* (1997). at <http://scholar.google.es/scholar?q=related:H5J6D58b8soJ:scholar.google.com/&hl=ca&as_sdt=0,5#0>
 142. 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).
 143. Wechsler, D. Wechsler Memory Scale- (Third Ed.). *San Antonio, TX Psychol. Corp.* (1997).
 144. Lezak, M. D. *Neuropsychological assessment (3rd ed.)*. *Neuropsychol. Assess.* (3rd *ed.*) (1995). at <<http://search.ebscohost.com/login.aspx?direct=true&db=psych&AN=1995-97708-000&site=ehost-live>>
 145. Wechsler, D. Wechsler Adult Intelligence Scale III UK. *London Psychol. Corp.* (1998).
 146. Raven, J., Raven, J. C. & Court, J. *Manual for Raven's progressive matrices and vocabulary scales. Raven Man.* (1998).
 147. Eysenck, H. J. & Eysenck, S. B. G. Manual for the Eysenck Personality Questionnaire. (*EPQ-R Adult*) 21 (1994). doi:10.1177/014662168000400106
 148. Eysenck, B. G., Eysenck, H. J. & Barrett, P. VERSION OF THE PSYCHOTICISM SCALE. *Pers. individ. Diff.* **6**, 21–29 (1985).
 149. Hirschfeld, R. M. *et al.* Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *Am J Psychiatry* **157**, 1873–1875 (2000).
 150. Hirschfeld, R. M. A. The Mood Disorder Questionnaire: A Simple, Patient-Rated Screening Instrument for Bipolar Disorder. *Prim. Care Companion J. Clin. Psychiatry* **4**, 9–11 (2002).
 151. Raine, A. & Benishay, D. The SPQ-B: A Brief Screening Instrument for Schizotypal Personality Disorder. *J. Pers. Disord.* **9**, 346–355 (1995).
 152. Goldberg, D. P. & Hillier, V. F. A scaled version of the General Health Questionnaire. *Psychol. Med.* **9**, 139–145 (1979).
 153. Box, G. & Cox, D. An analysis of transformations. *J. R. Stat. Soc. Ser. B* **26**, 211–252 (1964).
 154. Venables, W. & Ripley, B. *Modern applied statistics with S.* (Springer,

- 2002).
155. R Development Core Team. R: A language and environment for statistical computing. *R Found. Stat. Comput.* (2016). at <<https://www.r-project.org/>>
 156. Kerr, S. M. *et al.* Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Med. Genet.* **14**, 38 (2013).
 157. Gunderson, K. L. Whole-genome genotyping on bead arrays. *Methods Mol. Biol.* **529**, 197–213 (2009).
 158. Li, M., Reilly, M. P., Rader, D. J. & Wang, L.-S. Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics* **26**, 798–806 (2010).
 159. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* **7**, 385–94 (2006).
 160. ENIGMA2 Genetics Support Team. ENIGMA2 1KGP Cookbook (v3) for MDS and Imputation. *Enigm. 1KGP Cookb. [Online]. Enhancing Neuroimaging Genet. through Meta-Analysis Consort.* 15 (2013). at <http://enigma.ini.usc.edu/wp-content/uploads/2012/07/old/ENIGMA2_1KGP_cookbook_v3.doc>
 161. 1000 Genomes Project Consortium, T. 1000 G. P. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
 162. 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).
 163. Everitt, B. & Hothorn, T. *An Introduction to Applied Multivariate Analysis with R.* (Springer, 2011).
 164. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 437 (2015).
 165. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
 166. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
 167. 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**, 100–6 (2014).
 168. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits.* (Sinauer, 1998).
 169. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
 170. Visscher, P. M. *et al.* Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genet.* **10**, e1004269 (2014).
 171. Cortes, A. *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat.*

- Genet.* **45**, 730–8 (2013).
172. Kessler, R. C., Davis, C. G. & Kendler, K. S. Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. *Psychol. Med.* **27**, 1101–19 (1997).
 173. Butler, D., Cullis, B., Gilmour, A. & Gogel, B. ASReml-R reference manual, release 3. *Tech. report, Dep. Prim. Ind. Queensl.* (2007).
 174. Visscher, P. M. A Note on the Asymptotic Distribution of Likelihood Ratio Tests to Test Variance Components. *Twin Res. Hum. Genet.* **9**, 490–495 (2006).
 175. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–50 (2003).
 176. Newmann, J. P. Aging and depression. *Psychol. Aging* **4**, 150–65 (1989).
 177. Kessler, R., Foster, C., Webster, P. & House, J. The relationship between age and depressive symptoms in two national surveys. *Psychol Aging* **7**, 119–126 (1992).
 178. Rodriguez, G. & Goldman, N. An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *J. R. Stat. Soc. Ser. A (Statistics Soc.)* **158**, 73 (1995).
 179. Hadfield, J. D. *Journal of Statistical Software.* **33**, (2010).
 180. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. (2006).
 181. 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**, 100–6 (2014).
 182. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
 183. Kenward, M. G. & Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–97 (1997).
 184. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
 185. Turner, S. D. *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.* *bioRxiv* (Cold Spring Harbor Labs Journals, 2014). doi:10.1101/005165
 186. Nakagawa, S. & Schielzeth, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**, 133–142 (2013).
 187. www.ukbiobank.ac.uk/. www.ukbiobank.ac.uk/. at [<http://www.ukbiobank.ac.uk/>](http://www.ukbiobank.ac.uk/)
 188. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, e1001779 (2015).
 189. 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).
190. UK Biobank. Resource 155580; Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource. (2015). at <<http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580>>
 191. Fernandez-Pujals, A.-M. *et al.* Epidemiology and Heritability of depression, stratified by age of onset in Generation Scotland; a large family and population-based cohort. *unpublished*
 192. Revelle, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <http://CRAN.R-project.org/package=psych> Version = 1.6.6. (2016).
 193. Altamura, A. . *et al.* Lifetime prevalence of brief recurrent depression (results from a community survey). *Eur. Neuropsychopharmacol.* **5**, 99–102 (1995).
 194. Andrade, L. *et al.* The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *Int. J. Methods Psychiatr. Res.* **12**, 3–21 (2003).
 195. Bijl, R. V, Ravelli, A. & van Zessen, G. Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc. Psychiatry Psychiatr. Epidemiol.* **33**, 587–95 (1998).
 196. Scottish Government. Mid-Year Population Estimates Scotland, Mid-2015 and Corrected Population Estimates for Mid-2012, Mid-2013 and Mid-2014 Population estimates by sex, age and administrative area. (2016).
 197. Lyness, J. M. *et al.* Older Age and the Underreporting of Depressive Symptoms. *J. Am. Geriatr. Soc.* **43**, 216–221 (1995).
 198. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* (2016). doi:10.1038/ng.3623
 199. Bulik-Sullivan, B. K. LD Score documentation. at <<https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation>>
 200. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–9 (2008).
 201. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
 202. Ryu, J.-H. & Chun, J.-S. Opposing Roles of WNT-5A and WNT-11 in Interleukin-1beta Regulation of Type II Collagen Expression in Articular Chondrocytes. *J. Biol. Chem.* **281**, 22039–22047 (2006).
 203. Ridout, K. K. *et al.* Interleukin 1B gene (IL1B) variation and internalizing symptoms in maltreated preschoolers. *Dev. Psychopathol.* **26**, 1277–87 (2014).
 204. Kapelski, P. *et al.* Association study of functional polymorphisms in interleukins and interleukin receptors genes: IL1A, IL1B, IL1RN, IL6, IL6R, IL10, IL10RA and TGFB1 in schizophrenia in Polish population. *Schizophr. Res.* **169**, 1–9 (2015).
 205. Fatjó-Vilas, M. *et al.* Effect of the Interleukin-1 β Gene on Dorsolateral

- Prefrontal Cortex Function in Schizophrenia: A Genetic Neuroimaging Study. *Biol. Psychiatry* **72**, 758–765 (2012).
206. Smoller, J. W. Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
 207. Lee, M. T. M. *et al.* Genome-wide association study of bipolar I disorder in the Han Chinese population. *Mol. Psychiatry* **16**, 548–56 (2011).
 208. Jan, W.-C. *et al.* Exploring the associations between genetic variants in genes encoding for subunits of calcium channel and subtypes of bipolar disorder. *J. Affect. Disord.* **157**, 80–6 (2014).
 209. Walker, R. M. *et al.* DNA methylation in a Scottish family multiply affected by bipolar disorder and major depressive disorder. *Clin. Epigenetics* **8**, 5 (2016).
 210. Forstner, A. J. *et al.* Genome-wide analysis implicates microRNAs and their target genes in the development of bipolar disorder. *Transl. Psychiatry* **5**, e678 (2015).
 211. Forstner, A. J. *et al.* Genome-wide analysis implicates microRNAs and their target genes in the development of bipolar disorder. *Transl. Psychiatry* **5**, e678 (2015).
 212. Soldatov, N. M. CACNB2: An Emerging Pharmacological Target for Hypertension, Heart Failure, Arrhythmia and Mental Disorders. *Curr. Mol. Pharmacol.* **8**, 32–42 (2015).
 213. Dong, C. *et al.* Role of peroxisome proliferator-activated receptors gene polymorphisms in type 2 diabetes and metabolic syndrome. *World J. Diabetes* **6**, 654–61 (2015).
 214. Tang, L., Lü, Q., Cao, H., Yang, Q. & Tong, N. PPARD rs2016520 polymorphism is associated with metabolic traits in a large population of Chinese adults. *Gene* **585**, 191–195 (2016).
 215. Watkins, C. C., Sawa, A. & Pomper, M. G. Glia and immune cell signaling in bipolar disorder: insights from neuropharmacology and molecular imaging to clinical application. *Transl. Psychiatry* **4**, e350 (2014).
 216. Scholtysek, C. *et al.* PPAR β/δ governs Wnt signaling and bone turnover. *Nat. Med.* **19**, 608–613 (2013).
 217. Zandi, P. P. *et al.* Association study of Wnt signaling pathway genes in bipolar disorder. *Arch. Gen. Psychiatry* **65**, 785–93 (2008).
 218. Endicott, J. *et al.* The Global Assessment Scale. *Arch. Gen. Psychiatry* **33**, 766 (1976).
 219. Heath, A. C. *et al.* A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. *Biol. Psychiatry* **70**, 513–8 (2011).
 220. Mühleisen, T. W. *et al.* Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* **5**, (2014).
 221. Kajitani, T. *et al.* Cloning and characterization of granulosa cell high-mobility group (HMG)-box protein-1, a novel HMG-box transcriptional regulator strongly expressed in rat ovarian granulosa cells. *Endocrinology* **145**, 2307–18 (2004).
 222. Shirali, M. Using Haplotype Mapping to Uncover the Missing Heritability:

- A Simulation Study. (2014).
223. Zeng, Y. *et al.* Genome-wide regional heritability mapping identifies a locus within the TOX2 gene associated with Major Depressive Disorder. *unpublished* (2016).
 224. Schlaepfer, T. E. *et al.* Deep brain stimulation to reward circuitry alleviates anhedonia in refractory major depression. *Neuropsychopharmacology* **33**, 368–77 (2008).
 225. Eicher, J. D. *et al.* Characterization of the DYX2 locus on chromosome 6p22 with reading disability, language impairment, and IQ. *Hum. Genet.* **133**, 869–81 (2014).
 226. Eicher, J. D. *et al.* Dyslexia and language impairment associated genetic markers influence cortical thickness and white matter in typically developing children. *Brain Imaging Behav.* **10**, 272–82 (2016).
 227. Li, P.-Y. *et al.* Tspan9 inhibits the proliferation, migration and invasion of human gastric cancer SGC7901 cells via the ERK1/2 pathway. *Oncol. Rep.* **36**, 448–54 (2016).
 228. Massie, M. J. Prevalence of depression in patients with cancer. *J. Natl. Cancer Inst. Monogr.* **2004**, 57–71 (2004).
 229. Luo, H., Cowen, L., Yu, G., Jiang, W. & Tang, Y. SMG7 is a critical regulator of p53 stability and function in DNA damage stress response. *Cell Discov.* **2**, 15042 (2016).
 230. Martin, L. *et al.* Identification and characterization of small molecules that inhibit nonsense-mediated RNA decay and suppress nonsense p53 mutations. *Cancer Res.* **74**, 3104–13 (2014).
 231. Italiano, D., Lena, A. M., Melino, G. & Candi, E. Identification of NCF2/p67phox as a novel p53 target gene. *Cell Cycle* **11**, 4589–4596 (2012).
 232. Shimohama, S. *et al.* Activation of NADPH Oxidase in Alzheimer's Disease Brains. *Biochem. Biophys. Res. Commun.* **273**, 5–9 (2000).
 233. Starkstein, S. E., Mizrahi, R. & Power, B. D. Depression in Alzheimer's disease: phenomenology, clinical correlates and treatment. *Int. Rev. Psychiatry* **20**, 382–8 (2008).
 234. Filatova, E. V *et al.* Expression analysis of genes of ubiquitin-proteasome protein degradation system in MPTP-induced mice models of early stages of Parkinson's disease. *Dokl. Biochem. Biophys.* **456**, 116–8 (2014).
 235. McNaught, K. S. P., Olanow, C. W., Halliwell, B., Isacson, O. & Jenner, P. Failure of the ubiquitin-proteasome system in Parkinson's disease. *Nat. Rev. Neurosci.* **2**, 589–594 (2001).
 236. Nilsson, F. M., Kessing, L. V & Bolwig, T. G. Increased risk of developing Parkinson's disease for patients with major affective disorder: a register study. *Acta Psychiatr. Scand.* **104**, 380–6 (2001).
 237. Ishihara, L. & Brayne, C. A systematic review of depression and mental illness preceding Parkinson's disease. *Acta Neurol. Scand.* **113**, 211–20 (2006).
 238. Nuti, A. *et al.* Psychiatric comorbidity in a population of Parkinson's disease patients. *Eur. J. Neurol.* **11**, 315–20 (2004).
 239. Evangelou, M. *et al.* A method for gene-based pathway analysis using

- genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet. Epidemiol.* **38**, 661–70 (2014).
240. Barnard, K. D., Skinner, T. C. & Peveler, R. The prevalence of co-morbid depression in adults with Type 1 diabetes: systematic literature review. *Diabet. Med.* **23**, 445–448 (2006).
 241. Lloyd, S. E., Maytham, E. G., Grizenkova, J., Hummerich, H. & Collinge, J. A Copine family member, Cpne8, is a candidate quantitative trait gene for prion disease incubation time in mouse. *Neurogenetics* **11**, 185–91 (2010).
 242. Zeidler, M. *et al.* New variant Creutzfeldt-Jakob disease: psychiatric features. *Lancet (London, England)* **350**, 908–10 (1997).
 243. Consultation, W. The Revision of the Surveillance Case Definition for Variant Creutzfeldt-Jakob Disease (vCJD) World Health Organization. (2001).
 244. Reinhardt, P. *et al.* Genetic correction of a LRRK2 mutation in human iPSCs links parkinsonian neurodegeneration to ERK-dependent changes in gene expression. *Cell Stem Cell* **12**, 354–67 (2013).
 245. Wang, K.-S. *et al.* A meta-analysis of two genome-wide association studies identifies 3 new loci for alcohol dependence. *J. Psychiatr. Res.* **45**, 1419–25 (2011).
 246. Zuo, L. *et al.* Genome-wide association discoveries of alcohol dependence. *Am. J. Addict.* **23**, 526–39 (2014).
 247. Chen, X., Cho, K., Singer, B. H. & Zhang, H. The nuclear transcription factor PKNX2 is a candidate gene for substance dependence in European-origin women. *PLoS One* **6**, e16002 (2011).
 248. Wang, K.-S., Zhang, Q., Liu, X., Wu, L. & Zeng, M. PKNX2 is associated with formal thought disorder in schizophrenia: a meta-analysis of two genome-wide association studies. *J. Mol. Neurosci.* **48**, 265–72 (2012).
 249. Kang, E. *et al.* Interaction between FEZ1 and DISC1 in regulation of neuronal development and risk for schizophrenia. *Neuron* **72**, 559–71 (2011).
 250. Lipska, B. K. *et al.* Expression of DISC1 binding partners is reduced in schizophrenia and associated with DISC1 SNPs. *Hum. Mol. Genet.* **15**, 1245–58 (2006).
 251. Gu, Z., Flemington, C., Chittenden, T. & Zambetti, G. P. ei24, a p53 response gene involved in growth suppression and apoptosis. *Mol. Cell. Biol.* **20**, 233–41 (2000).
 252. Zhao, X. *et al.* Apoptosis factor EI24/PIG8 is a novel endoplasmic reticulum-localized Bcl-2-binding protein which is associated with suppression of breast cancer invasiveness. *Cancer Res.* **65**, 2125–9 (2005).
 253. Gentile, M., Ahnström, M., Schön, F. & Wingren, S. Candidate tumour suppressor genes at 11q23-q24 in breast cancer: evidence of alterations in PIG8, a gene involved in p53-induced apoptosis. *Oncogene* **20**, 7753–60 (2001).
 254. Mazumder Indra, D. *et al.* Inactivation of CHEK1 and EI24 is associated with the development of invasive cervical carcinoma: clinical and

- prognostic implications. *Int. J. cancer* **129**, 1859–71 (2011).
255. Cancer Research UK. Breast cancer statistics. at
<<http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero>>
 256. Devineni, A. V, Eddison, M. & Heberlein, U. The novel gene tank, a tumor suppressor homolog, regulates ethanol sensitivity in *Drosophila*. *J. Neurosci.* **33**, 8134–43 (2013).
 257. Schuckit, M. A. Low level of response to alcohol as a predictor of future alcoholism. *Am. J. Psychiatry* **151**, 184–9 (1994).
 258. Morean, M. E. & Corbin, W. R. Subjective response to alcohol: a critical review of the literature. *Alcohol. Clin. Exp. Res.* **34**, 385–95 (2010).
 259. Thompson, L. H. & West, M. G. XRCC1 keeps DNA from getting stranded. *Mutat. Res.* **459**, 1–18 (2000).
 260. Kang, C. H. *et al.* The prognostic significance of ERCC1, BRCA1, XRCC1, and betaIII-tubulin expression in patients with non-small cell lung cancer treated by platinum- and taxane-based neoadjuvant chemotherapy and surgical resection. *Lung Cancer* **68**, 478–83 (2010).
 261. MacAuley, A. & Ladiges, W. C. Approaches to determine clinical significance of genetic variants. *Mutat. Res.* **573**, 205–20 (2005).
 262. Pettan-Brewer, C. *et al.* Tumor growth is suppressed in mice expressing a truncated XRCC1 protein. *Am. J. Cancer Res.* **2**, 168–77 (2012).
 263. Sujitha, S. P. *et al.* DNA Repair Gene (XRCC1) Polymorphism (Arg399Gln) Associated with Schizophrenia in South Indian Population: A Genotypic and Molecular Dynamics Study. *PLoS One* **11**, e0147348 (2016).
 264. Czarny, P. *et al.* Impact of Single Nucleotide Polymorphisms of Base Excision Repair Genes on DNA Damage and Efficiency of DNA Repair in Recurrent Depression Disorder. *Mol. Neurobiol.* 1–10 (2016). doi:10.1007/s12035-016-9971-6
 265. Vidal, A. E., Boiteux, S., Hickson, I. D. & Radicella, J. P. XRCC1 coordinates the initial and late stages of DNA abasic site repair through protein-protein interactions. *EMBO J.* **20**, 6530–9 (2001).
 266. Yamamori, T. *et al.* SIRT1 deacetylates APE1 and regulates cellular base excision repair. *Nucleic Acids Res.* **38**, 832–845 (2010).
 267. Marchesi, C. *et al.* Inactivation of endothelial proprotein convertase 5/6 decreases collagen deposition in the cardiovascular system: role of fibroblast autophagy. *J. Mol. Med. (Berl)*. **89**, 1103–11 (2011).
 268. Jang, H. B. *et al.* Intake levels of dietary polyunsaturated fatty acids modify the association between the genetic variation in PCSK5 and HDL cholesterol. *J. Med. Genet.* **51**, 782–8 (2014).
 269. Nudel, R. *et al.* Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes. Brain. Behav.* **13**, 418–29 (2014).
 270. Leonard, L. B. *Children with Specific Language Impairment*. (The MIT Press, 2014). at
<https://books.google.co.uk/books?id=GLTPAwAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false>
 271. Pettigrew, K. A. *et al.* Further evidence for a parent-of-origin effect at the

- NOP9 locus on language-related phenotypes. *J. Neurodev. Disord.* **8**, 24 (2016).
272. Li, P. *et al.* LTB4 promotes insulin resistance in obese mice by acting on macrophages, hepatocytes and myocytes. *Nat. Med.* **21**, 239–47 (2015).
 273. Agarwal, A. *et al.* Dysregulated expression of neuregulin-1 by cortical pyramidal neurons disrupts synaptic plasticity. *Cell Rep.* **8**, 1130–45 (2014).
 274. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* **71**, 877–92 (2002).
 275. Georgieva, L. *et al.* Support for neuregulin 1 as a susceptibility gene for bipolar disorder and schizophrenia. *Biol. Psychiatry* **64**, 419–27 (2008).
 276. Munafò, M. R., Thiselton, D. L., Clark, T. G. & Flint, J. Association of the NRG1 gene and schizophrenia: a meta-analysis. *Mol. Psychiatry* **11**, 539–546 (2006).
 277. Litwack, E. D., Babey, R., Buser, R., Gesemann, M. & O’Leary, D. D. . Identification and characterization of two novel brain-derived immunoglobulin superfamily members with a unique structural organization. *Mol. Cell. Neurosci.* **25**, 263–274 (2004).
 278. van den Oord, E. J. C. G. *et al.* Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Arch. Gen. Psychiatry* **65**, 1062–71 (2008).
 279. Heck, A. *et al.* Evidence for associations between MDGA2 polymorphisms and harm avoidance: replication and extension of a genome-wide association finding. *Psychiatr. Genet.* **21**, 257–60 (2011).
 280. Hettema, J. M., van den Oord, E. J. C. G., An, S.-S., Kendler, K. S. & Chen, X. Follow-up association study of novel neuroticism gene MAMDC1. *Psychiatr. Genet.* **19**, 213–4 (2009).
 281. Ormel, J., Oldehinkel, A. J. & Vollebergh, W. Vulnerability before, during, and after a major depressive episode: a 3-wave population-based study. *Arch. Gen. Psychiatry* **61**, 990–6 (2004).
 282. Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. A longitudinal twin study of personality and major depression in women. *Arch. Gen. Psychiatry* **50**, 853–62 (1993).
 283. Koehler, K. *et al.* Mutations in GMPPA cause a glycosylation disorder characterized by intellectual disability and autonomic dysfunction. *Am. J. Hum. Genet.* **93**, 727–34 (2013).
 284. Liu, X. *et al.* Disruption of striated preferentially expressed gene locus leads to dilated cardiomyopathy in mice. *Circulation* **119**, 261–8 (2009).
 285. Agrawal, P. B. *et al.* SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *Am. J. Hum. Genet.* **95**, 218–26 (2014).
 286. Steptoe, A., Mohabir, A., Mahon, N. G. & McKenna, W. J. Health related quality of life and psychological wellbeing in patients with dilated cardiomyopathy. *Heart* **83**, 645–50 (2000).
 287. Hirschfeld, R. M. A., Lewis, L. & Vornik, L. A. Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals

- with bipolar disorder. *J. Clin. Psychiatry* **64**, 161–74 (2003).
288. Difflorio, A. & Jones, I. Is sex important? Gender differences in bipolar disorder. *Int. Rev. Psychiatry* **22**, 437–52 (2010).
 289. Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet (London, England)* **361**, 865–72 (2003).
 290. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
 291. Cox, D. R. (David R. & Snell, E. J. *Analysis of binary data*. (Chapman and Hall, 1989).
 292. NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
 293. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genet. Epidemiol.* **36**, 214–224 (2012).
 294. Flint, J. & Munafò, M. R. The endophenotype concept in psychiatric genetics. *Psychol. Med.* **37**, 163–80 (2007).
 295. Derks, E. M., Vorstman, J. A. S., Ripke, S., Kahn, R. S. & Ophoff, R. A. Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: A polygenic risk score analysis. *PLoS One* **7**, (2012).
 296. Glahn, D. C. & Blangero, J. Why endophenotype development requires families. *Chinese Sci. Bull.* **56**, 3382–3384 (2011).
 297. Clarke, T.-K. *et al.* Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Mol. Psychiatry* **21**, 419–425 (2016).
 298. Kirchheiner, J. *et al.* Bupropion and 4-OH-bupropion pharmacokinetics in relation to genetic polymorphisms in CYP2B6. *Pharmacogenetics* **13**, 619–26 (2003).
 299. Kiang, T. K. L. *et al.* Contribution of CYP2C9, CYP2A6, and CYP2B6 to Valproic Acid Metabolism in Hepatic Microsomes from Individuals with the CYP2C9*1/*1 Genotype. *Toxicol. Sci.* **94**, 261–271 (2006).
 300. Yuce-Artun, N. *et al.* Influence of CYP2B6 and CYP2C19 polymorphisms on sertraline metabolism in major depression patients. *Int. J. Clin. Pharm.* **38**, 388–94 (2016).
 301. Jaquenoud Sirot, E. *et al.* Multicenter study on the clinical effectiveness, pharmacokinetics, and pharmacogenetics of mirtazapine in depression. *J. Clin. Psychopharmacol.* **32**, 622–9 (2012).
 302. Uchihara, Y., Tanaka, K., Asano, T., Tamura, F. & Mizushima, T. Superoxide dismutase overexpression protects against glucocorticoid-induced depressive-like behavioral phenotypes in mice. *Biochem. Biophys. Res. Commun.* **469**, 873–7 (2016).
 303. Szebeni, A. *et al.* Shortened telomere length in white matter oligodendrocytes in major depression: potential role of oxidative stress. *Int. J. Neuropsychopharmacol.* **17**, 1579–89 (2014).
 304. Gałęcki, P., Szemraj, J., Bieńkiewicz, M., Florkowski, A. & Gałęcka, E. Lipid

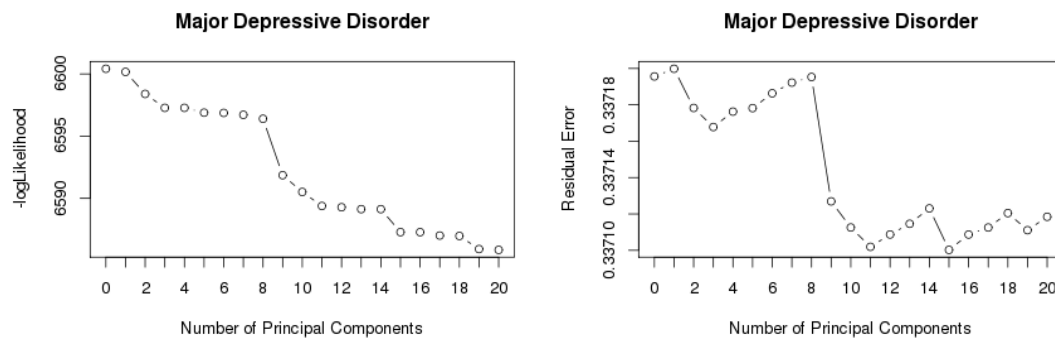
- peroxidation and antioxidant protection in patients during acute depressive episodes and in remission after fluoxetine treatment. *Pharmacol. Reports* **61**, 436–447 (2009).
305. Dalla, C., Antoniou, K., Papadopoulou-Daifoti, Z., Balthazart, J. & Bakker, J. Oestrogen-deficient female aromatase knockout (ArKO) mice exhibit depressive-like symptomatology. *Eur. J. Neurosci.* **20**, 217–28 (2004).
 306. Stimmel, G. L. & Gutierrez, M. A. Pharmacologic treatment strategies for sexual dysfunction in patients with epilepsy and depression. *CNS Spectr.* **11**, 31–7 (2006).
 307. Carrier, N. *et al.* The Anxiolytic and Antidepressant-like Effects of Testosterone and Estrogen in Gonadectomized Male Rats. *Biol. Psychiatry* **78**, 259–69 (2015).
 308. Seliktar, N., Polek, C., Brooks, A. & Hardie, T. Cognition in breast cancer survivors: hormones versus depression. *Psychooncology.* **24**, 402–7 (2015).
 309. Walters, J. T. R. & Owen, M. J. Endophenotypes in psychiatric genetics. *Mol. Psychiatry* **12**, 886–890 (2007).
 310. 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).
 311. Braunewell, K. H. *et al.* Association of VSNL1 with schizophrenia, frontal cortical function, and biological significance for its gene product as a modulator of cAMP levels and neuronal morphology. *Transl. Psychiatry* **1**, e22 (2011).
 312. Hong, K. S. *et al.* Genome-widely significant evidence of linkage of schizophrenia to chromosomes 2p24.3 and 6q27 in an SNP-Based analysis of Korean families. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **150B**, 647–52 (2009).
 313. Lin, C.-W. *et al.* VSNL1 Co-Expression Networks in Aging Include Calcium Signaling, Synaptic Plasticity, and Alzheimer's Disease Pathways. *Front. psychiatry* **6**, 30 (2015).
 314. Tu, Z., Cohen, M., Bu, H. & Lin, F. Tissue distribution and functional analysis of Sushi domain-containing protein 4. *Am. J. Pathol.* **176**, 2378–84 (2010).
 315. McIntosh, A. M. *et al.* Neuroticism, depressive symptoms and white-matter integrity in the Lothian Birth Cohort 1936. *Psychol. Med.* **43**, 1197–206 (2013).
 316. Holmquist, E., Okroj, M., Nodin, B., Jirström, K. & Blom, A. M. Sushi domain-containing protein 4 (SUSD4) inhibits complement by disrupting the formation of the classical C3 convertase. *FASEB J.* **27**, 2355–66 (2013).
 317. Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W. & Kelley, K. W. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nat. Rev. Neurosci.* **9**, 46–56 (2008).
 318. Power, R. A. *et al.* Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biol. Psych.* 3095–3105

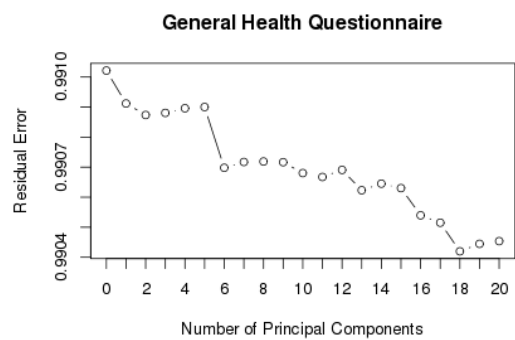
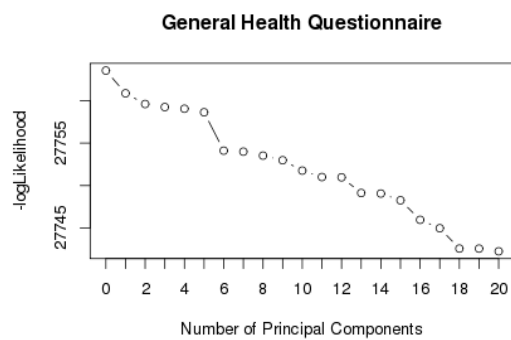
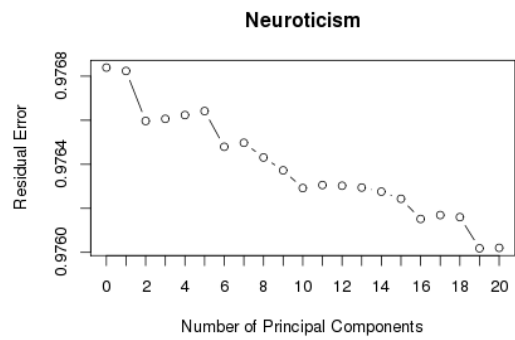
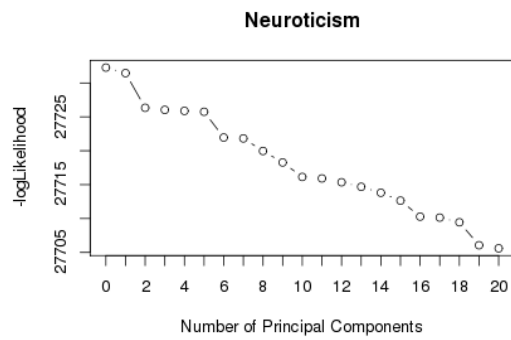
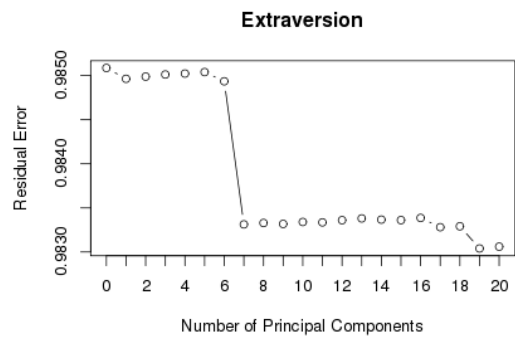
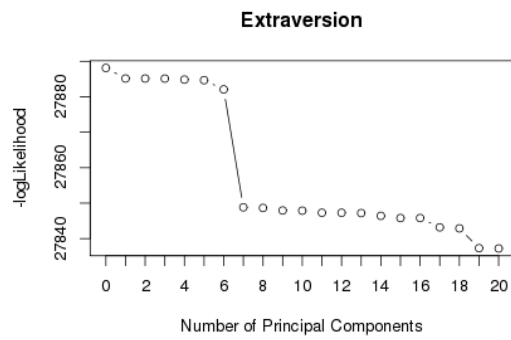
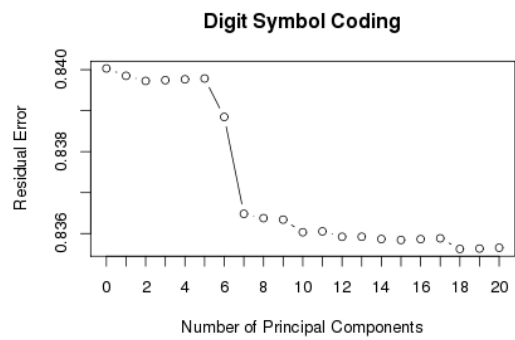
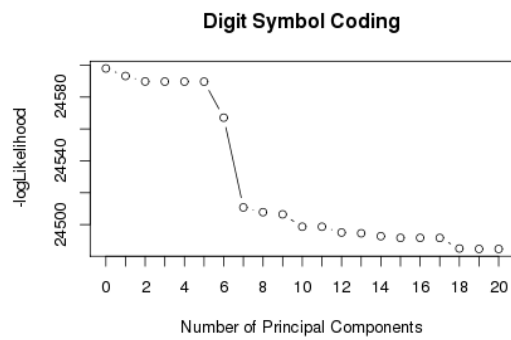
Appendices

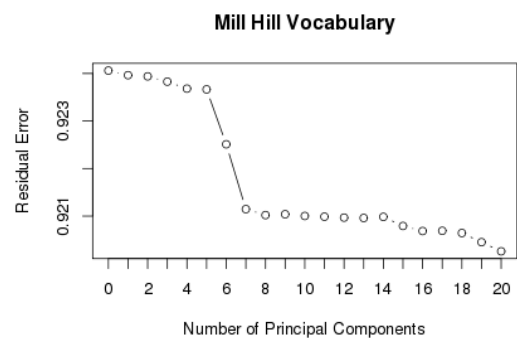
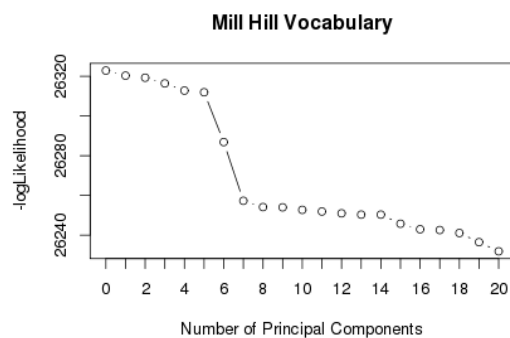
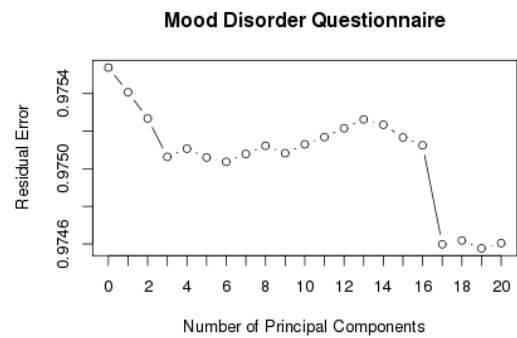
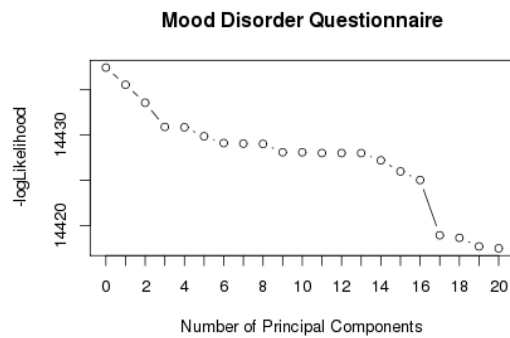
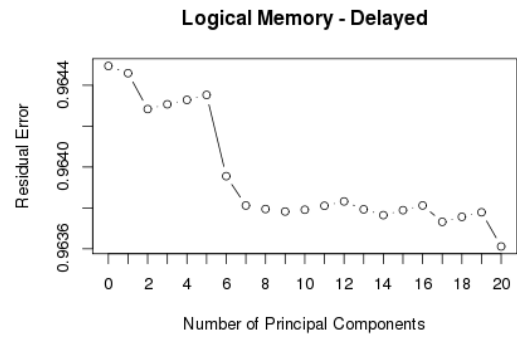
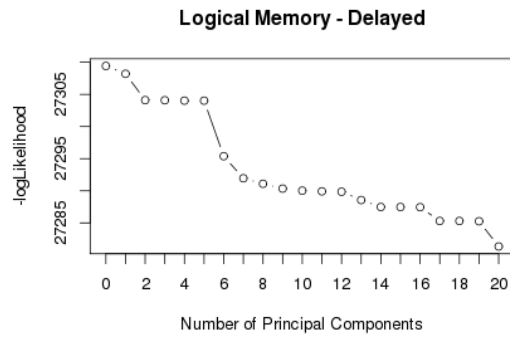
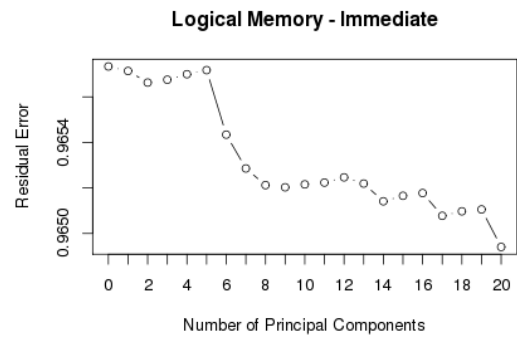
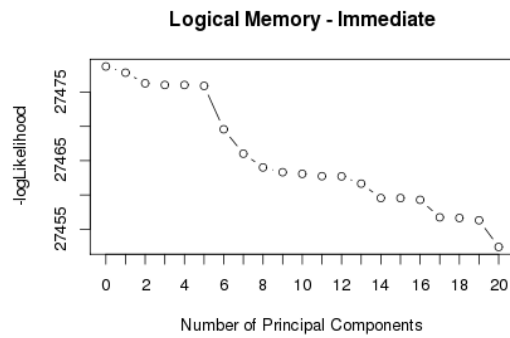
2.A. Principal components for population stratification

2.A.1. Scree plots

Shown in this Appendix are scree plots of the log-likelihood and residual error from linear regression models of the standardized traits on age, age², sex, and up to 20 principal components (PCs) for major depressive disorder, mood, personality and cognitive traits. PCs are included in the model in their order of importance based on the variance they explain in the kinship matrix.







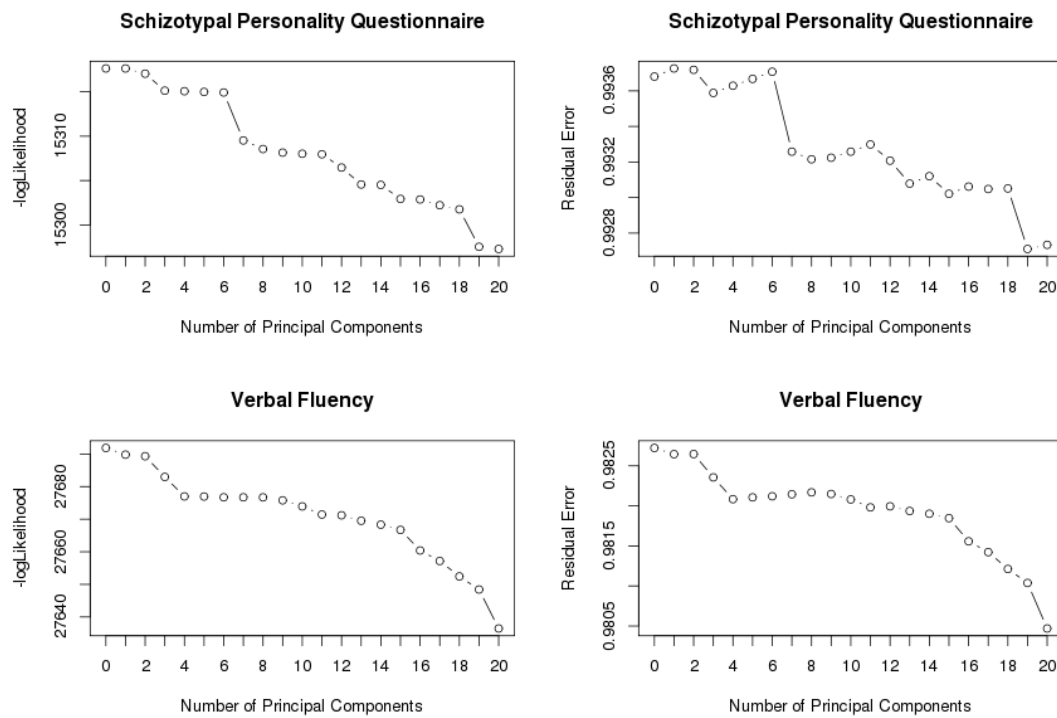


Figure 2.A.1. Scree plots of the log likelihood and residual error from regressing age, age2, sex and up to 20 PCs onto MDD, mood, personality and cognitive traits.

2.A.2. Significance of principal components

Significant PCs for each trait are shown in Table 2.A.2.

Trait	PCs which significantly improved model fit
MDD	9
DSC	1, 2, 6, 7, 8, 9, 10, 12, 18
EPQE	1, 6, 7, 17, 19
EPQN	2, 6, 8, 10, 16, 19
GHQ	1, 6, 13, 16, 18
LM1	6, 7, 8, 14, 17, 20
LM2D	2, 6, 7, 17, 20
MDQ	1, 2, 3, 17
MHV	1, 3, 4, 6, 7, 8, 15, 18, 19, 20

Table 2.B.2. Principal components which significantly improve model fit for MDD, mood, personality and cognitive traits.

2.B. Model selection fitting environmental components

2.B.1. Adding common household environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + household	positive	11727.59	11728.59	1	0.16
GHQ		positive	-9590.72	-9584.06	6.66	7.66E-04
EPQN		positive	-9309.97	-9308.59	1.38	0.12
MDQ		positive	-4860.15	-4857.92	2.23	0.05
SPQ		positive	-5217.25	-5197.69	19.56	1.90E-09
DSC		positive	-5957.29	-5956.92	0.37	0.33
EPQE		positive	-9450.97	-9449.62	1.35	0.12
LM1		boundary	-8972.54	-8972.54	0	0.50
LM2D		boundary	-8942.42	-8942.4	0.02	0.50
MHV		positive	-7325.22	-7292.15	33.07	5.62E-15
VF		positive	-8976.2	-8963.27	12.93	1.46E-06

Table 2.B.1. Likelihood ratio test results with the null model fitting an additive genetic component as a random effect (ped), and the alternative model fitting ped and common household environment (historic and recent fitted simultaneously) as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold.

2.B.2. Adding common sibling environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + sibling	boundary	11727.59	11727.59	0	0.50
GHQ		positive	-9590.72	-9590.56	0.16	0.42
EPQN		positive	-9309.97	-9309.69	0.28	0.36
MDQ		positive	-4860.15	-4860.1	0.05	0.50
SPQ		boundary	-5217.25	-5217.25	0	0.50
DSC		positive	-5957.29	-5955.05	2.24	0.05
EPQE		boundary	-9450.97	-9450.97	0	0.50
LM1		positive	-8972.54	-8967.1	5.44	2.16E-03
LM2D		positive	-8942.42	-8935.54	6.88	6.59E-04
MHV		positive	-7325.22	-7300.86	24.36	1.86E-11
VF		positive	-8976.2	-8974.99	1.21	0.13

Table 2.B.2. Likelihood ratio test results with the null model fitting ped as a random effect, and the alternative model fitting ped and shared sibling environment as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold.

2.B.3. Adding common spouse environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + spouse	positive	11727.59	11729.38	1.79	0.07
GHQ		positive	-9590.72	-9582.47	8.25	1.68E-04
EPQN		positive	-9309.97	-9309.54	0.43	0.32
MDQ		positive	-4860.15	-4847.26	12.89	1.46E-06
SPQ		positive	-5217.25	-5178.12	39.13	1.90E-15
DSC		positive	-5957.29	-5937.36	19.93	1.49E-09
EPQE		positive	-9450.97	-9447.5	3.47	2.09E-03
LM1		positive	-8972.54	-8970.21	2.33	0.05
LM2D		positive	-8942.42	-8939.36	3.06	0.02
<i>MHV</i>		?	-7325.22	-7173.59	151.63	<i>1.90E-15</i>
VF		positive	-8976.2	-8947.93	28.27	4.20E-13

Table 2.B.3. Likelihood ratio test results with the null model fitting ped as a random effect, and the alternative model fitting ped and shared spouse environment as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold. Traits for which the alternative model was erroneous are highlighted in italics.

2.B.4. Adding common spouse environment + common household environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + spouse + house	positive	11729.38	11729.71	0.33	0.34
GHQ		positive	-9582.47	-9579.75	2.72	0.03
EPQN		positive	-9309.54	-9308.55	0.99	0.16
MDQ		boundary	-4847.26	-4846.81	0.45	0.31
SPQ		positive	-5178.12	-5173.4	4.72	4.24E-03
DSC		boundary	-5937.36	-5937.36	0	0.50
EPQE		positive	-9447.5	-9447.28	0.22	0.39
LM1		boundary	-8970.21	-8970.21	0	0.50
LM2D		boundary	-8939.36	-8939.36	0	0.50
<i>MHV</i>		?/boundary	-7173.59	-7169.63	<i>3.96</i>	<i>8.10E-03</i>
VF		positive	-8947.93	-8943.91	4.02	7.95E-03

Table 2.B.4. Likelihood ratio test results with the null model fitting ped and spouse as random effects, and the alternative model fitting ped, spouse and household as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold. Traits for which the alternative model was erroneous are highlighted in italics.

2.B.5. Adding common spouse environment + common sibling environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + spouse + sibling	boundary	11729.38	11729.38	0	0.5
GHQ		positive	-9582.47	-9582.46	0.01	0.46
EPQN		positive	-9309.54	-9309.32	0.22	0.26
MDQ		boundary	-4847.26	-4847.26	0	0.5
SPQ		boundary	-5178.12	-5178.12	0	0.5
DSC		positive	-5937.36	-5936.97	0.39	0.19
EPQE		boundary	-9447.5	-9447.5	0	0.5
LM1		positive	-8970.21	-8965.7	4.51	1.34E-03
LM2D		positive	-8939.36	-8933.49	5.87	3.04E-04
<i>MHV</i>		singular	-7173.59	-7171.75	1.84	0.03
VF		positive	-8947.93	-8947.92	0.01	0.47

Table 2.B.5. Likelihood ratio test results with the null model fitting ped and spouse as random effects, and the alternative model fitting ped, spouse and sibling as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold. Traits for which the alternative model was erroneous are highlighted in italics.

2.B.6. Adding common household environment + common sibling environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + household + sibling	Boundary	11728.59	11728.59	0	0.50
GHQ		Positive	-9584.06	-9582.08	0.99	0.15
EPQN		Positive	-9308.59	-9307.52	0.54	0.28
MDQ		Positive	-4857.92	-4857.51	0.21	0.44
SPQ		Positive	-5197.69	-5197.08	0.31	0.39
DSC		Positive	-5956.92	-5953.86	1.54	0.05
EPQE		Positive	-9449.62	-9449.53	0.05	0.50
LM1		Positive	-8972.54	-8966.99	2.78	7.24E-03
LM2D		Boundary	-8942.40	-8935.13	3.64	2.09E-03
MHV		positive	-7292.15	-7238.38	26.72	2.45E-14
VF		Positive	-8963.27	-8956.81	3.24	3.48E-03

Table 2.B.6. Likelihood ratio test results with the null model fitting ped and household as random effects, and the alternative model fitting ped, household and sibling as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold.

2.A.7. Adding common spouse environment + common household environment + common sibling environment

Trait	Model	Constraint	Null	Full	χ^2	P-value
MDD	ped + spouse + household + sibling	boundary	11729.71	11729.71	0	0.50
GHQ		positive	-9579.75	-9579.03	0.72	0.23
EPQN		boundary	-9308.55	-9307.52	1.03	0.16
MDQ		boundary	-4846.81	-4846.81	0	0.50
SPQ		boundary	-5173.4	-5173.4	0	0.50
DSC		boundary	-5937.36	-5936.97	0.39	0.32
EPQE		boundary	-9447.28	-9447.28	0	0.50
LM1		boundary	-8965.7	-8965.7	0	0.50
LM2D		boundary	-8933.49	-8933.42	0.07	0.50
<i>MHV</i>		<i>singularity</i>	-	-	-	-
VF		positive	-8943.91	-8943.38	0.53	0.28

Table 2.A.7. Likelihood ratio test results with the null model fitting ped, spouse and household as random effects, and the alternative model fitting ped spouse, household and sibling as random effects. Traits for which the alternative model was a significantly improved fit are highlighted in bold. Traits for which the alternative model was erroneous are highlighted in italics.

3.A. Code for Taylor transformation to liability scale

```
# Read in linear regression betas and standard errors

Taylor<-read.table("MDD_GWAS_summary_stats.txt", header=TRUE,sep=",")

# Set prevalence of trait (e.g. MDD)

P=0.13

# Set b01 and se01 to be the Beta/SE output from linear scale

Taylor$b01<-Taylor$B

Taylor$se01<-Taylor$SE

# Estimate Odds Ratio from Beta (b01)

Taylor$ORn <- (P + Taylor$b01)/(1 - P - Taylor$b01)

Taylor$ORd <- P/(1-P)

Taylor$OR <- Taylor$ORn/Taylor$ORd

# Calculate logit

Taylor$blogit <- abs(log(Taylor$OR))*sign(Taylor$b01)

Taylor$vblogit= (1/(P-Taylor$b01) + 1/(1-P-Taylor$b01))^2 * Taylor$se01^2

Taylor$selogit = Taylor$blogit * Taylor$se01 / Taylor$b01

# Calculate confidence intervals

Taylor$CI_L=Taylor$OR-(1.96*Taylor$selogit)

Taylor$CI_U=Taylor$OR+(1.96*Taylor$selogit)
```


4. Variance component estimates for quantitative traits

4.A. Univariate analyses

4.A.1. Digit Symbol Coding

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.04	0.01	0.01	1.02	Positive
YHH.var	0.05	0.02	0.01	1.52	Positive
sib.var	0.07	0.02	0.009	2.45	Positive
ped.var	0.86	0.30	0.02	14.29	Positive
resid.var	1.00	0.35	0.01	25.13	Positive

Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	2.28E-07	5.73E-08	4.68E-09	12.25	Boundary
YHH.var	1.94E-07	4.88E-08	3.99E-09	12.25	Boundary
spouse.var	0.39	0.10	0.02	6.22	Positive
sib.var	0.03	0.008	0.009	0.89	Positive
ped.var	1.38	0.35	0.02	22.18	Positive
resid.var	1.00	0.25	0.02	12.25	Positive

Variance component estimates from univariate analysis of DSC, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.2. Extraversion

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.03	0.02	0.02	1.05	Positive
YHH.var	0.05	0.03	0.02	1.68	Positive
sib.var	0.009	0.006	0.01	0.42	Positive
ped.var	0.42	0.27	0.03	9.12	Positive
resid.var	1.00	0.65	0.02	32.33	Positive

Variance component estimates from univariate analysis of EPQE. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.004	0.002	0.02	0.15	Positive
YHH.var	0.02	0.01	0.02	0.66	Positive
spouse.var	0.09	0.05	0.02	2.14	Positive
sib.var	3.99e-07	2.41e-07	1.15e-08	20.96	Boundary
ped.var	0.50	0.30	0.03	11.53	Positive
resid.var	1.00	0.60	0.03	20.96	Positive

Variance component estimates from univariate analysis of EPQE, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.3. Neuroticism

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.05	0.03	0.02	1.95	Positive
YHH.var	0.04	0.03	0.02	1.49	Positive
sib.var	0.03	0.02	0.01	1.45	Positive
ped.var	0.35	0.23	0.03	7.72	Positive
resid.var	1.00	0.65	0.02	33.07	Positive

Variance component estimates from univariate analysis of EPQN. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.05	0.03	0.01	1.95	Positive
YHH.var	0.04	0.03	0.02	1.49	Positive
spouse.var	2.51e-07	1.63e-07	4.94e-09	33.07	Boundary
sib.var	0.03	0.02	0.01	1.45	Positive
ped.var	0.35	0.23	0.03	7.72	Positive
resid.var	1.00	0.65	0.02	33.07	Positive

Variance component estimates from univariate analysis of EPQN, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.4. General Health Questionnaire

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.09	0.06	0.02	3.56	Positive
YHH.var	0.09	0.07	0.02	3.62	Positive
sib.var	0.04	0.03	0.01	1.96	Positive
ped.var	0.16	0.11	0.03	3.67	Positive
resid.var	1.00	0.71	0.02	34.70	Positive

Variance component estimates from univariate analysis of GHQ. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.07	0.04	0.02	2.26	Positive
YHH.var	0.07	0.05	0.02	2.33	Positive
spouse.var	0.11	0.07	0.03	2.62	Positive
sib.var	0.03	0.02	0.02	1.19	Positive
ped.var	0.23	0.15	0.03	4.31	Positive
resid.var	1.00	0.65	0.03	21.39	Positive

Variance component estimates from univariate analysis of GHQ, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.5. Logical Memory Immediate

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.01	0.007	0.02	0.44	Positive
YHH.var	0.002	0.001	0.02	0.06	Positive
sib.var	0.07	0.04	0.01	3.16	Positive
ped.var	0.52	0.30	0.03	10.53	Positive
resid.var	1.00	0.58	0.02	31.20	Positive

Variance component estimates from univariate analysis of LM1. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	5.08e-07	2.76e-07	1.41e-08	19.61	Boundary
YHH.var	2.23e-07	1.21e-07	6.19e-09	19.61	Boundary
spouse.var	0.07	0.04	2.18e-02	1.69	Positive
sib.var	0.07	0.04	1.27e-02	2.98	Positive
ped.var	0.58	0.31	1.99e-02	15.84	Positive
resid.var	1.00	0.54	2.77e-02	19.61	Positive

Variance component estimates from univariate analysis of LM1, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.6. Logical Memory Delayed

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	5.75e-07	3.45e-07	9.68e-09	35.63	Boundary
YHH.var	0.02	0.01	0.02	0.90	Positive
sib.var	0.08	0.05	0.01	3.72	Positive
ped.var	0.45	0.27	0.02	12.08	Positive
resid.var	1.00	0.60	0.02	35.63	Positive

Variance component estimates from univariate analysis of LM2D. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	2.94e-07	1.64e-07	8.14e-09	20.16	Boundary
YHH.var	0.01	0.006	0.02	0.36	Positive
spouse.var	0.08	0.04	0.02	1.86	Positive
sib.var	0.08	0.04	0.01	3.35	Positive
ped.var	0.50	0.28	0.02	11.98	Positive
resid.var	1.00	0.56	0.03	20.16	Positive

Variance component estimates from univariate analysis of LM2D, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.7. Mood Disorder Questionnaire

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.10	0.06	0.03	2.37	Positive
YHH.var	0.06	0.04	0.03	1.36	Positive
sib.var	0.03	0.02	0.02	0.89	Positive
ped.var	0.32	0.20	0.04	4.55	Positive
resid.var	1.00	0.63	0.03	21.96	Positive

Variance component estimates from univariate analysis of MDQ. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.05	2.20e-02	0.02	0.97	Positive
YHH.var	3.52e-07	1.60e-07	1.48e-08	10.85	Boundary
spouse.var	4.28e-01	0.19	0.04	5.56	Positive
sib.var	2.74e-07	1.25e-07	1.15e-08	10.85	Boundary
ped.var	0.62	0.28	0.03	9.39	Positive
resid.var	1.00	0.46	0.04	10.85	Positive

Variance component estimates from univariate analysis of MDQ, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.8. Mill Hill Vocabulary

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.70	0.11	0.01	8.11	Positive
YHH.var	1.01	0.16	0.01	10.96	Positive
sib.var	0.66	0.10	0.01	9.92	Positive
ped.var	2.00	0.32	0.02	13.33	Positive
resid.var	1.00	0.16	0.01	10.67	Positive

Variance component estimates from univariate analysis of MHV. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	14.39	0.008	0.02	0.52	Positive
YHH.var	99.90	0.06	0.02	3.31	Positive
spouse.var	406.34	0.22	0.01	15.84	Positive
sib.var	58.81	0.03	0.01	2.72	Positive
ped.var	954.17	0.53	0.031	15.66	Positive
resid.var	1.00	0.0006	0.02	0.03	Positive

Variance component estimates from univariate analysis of MHV, including spouse (log likelihood did not converge), spouse.var: variance attributable to common spouse environment.

4.A.9. Schizotypal Personality Questionnaire

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.29	0.15	0.02	5.97	Positive
YHH.var	0.23	0.12	0.03	4.40	Positive
sib.var	0.04	0.02	0.02	1.08	Positive
ped.var	0.38	0.19	0.04	4.53	Positive
resid.var	1.00	0.51	0.03	18.13	Positive

Variance component estimates from univariate analysis of SPQ. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.26	0.08	0.02	3.05	Positive
YHH.var	0.14	0.04	0.03	1.54	Positive
spouse.var	0.83	0.25	0.03	7.44	Positive
sib.var	3.04e-07	9.08e-08	1.15e-08	7.90	Boundary
ped.var	1.08	0.32	0.04	8.45	Positive
resid.var	1.00	0.30	0.04	7.90	Positive

Variance component estimates from univariate analysis of SPQ, including spouse, spouse.var: variance attributable to common spouse environment.

4.A.10. Verbal Fluency

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.15	0.06	0.02	3.56	Positive
YHH.var	0.29	0.11	0.02	6.16	Positive
sib.var	0.12	0.05	0.01	3.56	Positive
ped.var	0.98	0.37	0.03	13.23	Positive
resid.var	1.00	0.38	0.02	20.33	Positive

Variance component estimates from univariate analysis of VF. Variance component estimates from univariate analysis of DSC. OHH: common old/historic household environment; YHH: common young/recent household environment; sib: common sibling environment; ped: additive genetic component; resid: residual; var: variance.

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.03	0.009	0.02	0.50	Positive
YHH.var	0.21	0.06	0.02	2.93	Positive
spouse.var	0.44	0.13	0.02	5.20	Positive
sib.var	0.05	0.01	0.01	1.05	Positive
ped.var	1.68	0.48	0.03	13.89	Positive
resid.var	1.00	0.28	0.03	10.48	Positive

Variance component estimates from univariate analysis of VF, including spouse, spouse.var: variance attributable to common spouse environment.

4.B. Bivariate analyses between quantitative traits and MDD

4.B.1. Digit Symbol Coding and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.22	0.22	0.006	36.18	?
ped.cov	0.005	0.005	0.008	0.61	?
ped.DSC.var	0.31	0.31	0.021	14.73	Positive
OHH.MDD.var	0.0002	0.0002	0.003	0.06	?
OHH.cov	-0.0006	-0.0006	0.004	-0.13	?
OHH.DSC.var	0.002	0.002	0.01	0.17	?
YHH.MDD.var	0.0005	0.0005	0.003	0.18	?
YHH.cov	-0.004	-0.004	0.004	-0.84	?
YHH.DSC.var	0.03	0.03	0.01	2.29	?
sib.MDD.var	0.001	0.001	0.002	0.89	?
sib.cov	-0.005	-0.005	0.003	-1.72	?
sib.DSC.var	0.02	0.02	0.009	1.59	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.0004	0.0004	0.004	0.09	?
resid.cov	0.00001	0.00001	0.005	0.003	?
resid.DSC.var	0.35	0.35	0.01	25.25	Positive

Variance component estimates from bivariate analysis of DSC and MDD (full model did not converge).

effect	gamma	component	std.error	z.ratio	constraint
sib.DSC	0.02	0.02	0.009	2.48	Positive
OHH.DSC	0.01	0.01	0.01	1.01	Positive
OHH.cov	0.002	0.002	0.003	0.73	Positive
OHH.MDD	0.002	0.002	0.002	1.18	Positive
YHH.DSC	0.02	0.02	0.01	1.57	Positive
YHH.cov	-0.004	-0.004	0.003	-1.05	Positive
YHH.MDD	0.003	0.003	0.002	1.39	Positive
ped.DSC	0.30	0.30	0.02	14.32	Positive
ped.cov	-0.02	-0.02	0.005	-3.61	Positive
ped.MDD	0.02	0.02	0.003	6.87	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.DSC	0.35	0.35	0.01	25.10	Positive
resid.cov	-0.0009	-0.0009	0.004	-0.24	Positive
resid.MDD	0.09	0.09	0.002	39.18	Positive

Variance component estimates from conditional bivariate analysis of DSC and MDD, where sib effects are only fitted for DSC.

4.B.2. Extraversion and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.19	Positive
ped.cov	-0.01	-0.01	0.007	-1.42	Positive
ped.EPQE.var	0.26	0.26	0.03	8.91	Positive
OHH.MDD.var	0.002	0.002	0.002	1.25	Positive
OHH.cov	-0.004	-0.004	0.004	-1.00	Positive
OHH.EPQE.var	0.02	0.02	0.02	1.21	Positive
YHH.MDD.var	0.003	0.003	0.002	1.27	Positive
YHH.cov	0.0001	0.0001	0.004	0.02	?
YHH.EPQE.var	0.03	0.03	0.02	1.82	Positive
sib.MDD.var	0.0002	0.0002	0.002	0.13	?
sib.cov	0.0006	0.0006	0.003	0.17	?
sib.EPQE.var	0.01	0.01	0.01	0.72	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.45	Positive
resid.cov	-0.01	-0.01	0.005	-2.56	Positive
resid.EPQE.var	0.64	0.64	0.02	32.20	Positive

Variance component estimates from bivariate analysis of EPQE and MDD (full model did not converge).

effect	gamma	component	std.error	z.ratio	constraint
sib.EPQE	0.007	0.007	0.01	0.52	Positive
OHH.EPQE	0.02	0.02	0.02	1.12	Positive
OHH.cov	-0.003	-0.003	0.004	-0.89	Positive
OHH.MDD	0.002	0.002	0.002	1.25	Positive
YHH.EPQE	0.03	0.03	0.02	1.71	Positive
YHH.cov	0.002	0.002	0.004	0.51	Positive
YHH.MDD	0.003	0.003	0.002	1.26	Positive
ped.EPQE	0.27	0.27	0.03	9.10	Positive
ped.cov	-0.01	-0.01	0.006	-1.97	Positive
ped.MDD	0.019	0.019	0.003	6.90	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.EPQE	0.65	0.65	0.02	32.27	Positive
resid.cov	-0.01	-0.01	0.005	-2.91	Positive
resid.MDD	0.09	0.09	0.002	39.14	Positive

Variance component estimates from bivariate analysis of EPQE and MDD.

4.B.3. Neuroticism and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.13	Positive
ped.cov	0.03	0.03	0.007	4.57	Positive
ped.EPQN.var	0.22	0.22	0.03	7.49	Positive
OHH.MDD.var	0.003	0.003	0.002	1.32	Positive
OHH.cov	0.002	0.002	0.004	0.43	Positive
OHH.EPQN.var	0.03	0.03	0.02	2.11	Positive
YHH.MDD.var	0.003	0.003	0.002	1.40	Positive
YHH.cov	0.008	0.008	0.004	1.71	Positive
YHH.EPQN.var	0.03	0.03	0.02	1.67	Positive
sib.MDD.var	0.0002	0.0002	0.002	0.11	?
sib.cov	0.002	0.002	0.004	0.59	?
sib.EPQN.var	0.02	0.02	0.01	1.73	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.54	Positive
resid.cov	0.04	0.04	0.005	8.28	Positive
resid.EPQN.var	0.65	0.65	0.02	33.02	Positive

Variance component estimates from bivariate analysis of EPQN and MDD

4.B.4. General Health Questionnaire and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.14	Positive
ped.cov	0.04	0.04	0.008	4.85	Positive
ped.GHQ.var	0.11	0.11	0.03	3.50	Positive
OHH.MDD.var	0.003	0.003	0.002	1.43	Positive
OHH.cov	0.009	0.009	0.004	2.06	Positive
OHH.GHQ.var	0.07	0.07	0.02	3.84	Positive
YHH.MDD.var	0.003	0.003	0.002	1.33	Positive
YHH.cov	0.008	0.008	0.005	1.77	Positive
YHH.GHQ.var	0.07	0.07	0.02	3.72	Positive
sib.MDD.var	0.0002	0.0002	0.002	0.12	?
sib.cov	-0.0006	-0.0006	0.004	-0.17	?
sib.GHQ.var	0.03	0.03	0.01	2.11	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.61	Positive
resid.cov	0.03	0.03	0.005	6.28	Positive
resid.GHQ.var	0.71	0.71	0.02	34.72	Positive

Variance component estimates from bivariate analysis of GHQ and MDD

4.B.5. Logical Memory Immediate and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.12	Positive
ped.cov	-0.004	-0.004	0.007	-0.54	Positive
ped.LM1.var	0.29	0.29	0.03	10.05	Positive
OHH.MDD.var	0.002	0.002	0.002	1.18	Positive
OHH.cov	0.002	0.002	0.004	0.39	Positive
OHH.LM1.var	0.01	0.01	0.02	0.66	Positive
YHH.MDD.var	0.003	0.003	0.002	1.30	Positive
YHH.cov	-0.0007	-0.0007	0.004	-0.18	Positive
YHH.LM1.var	0.004	0.004	0.02	0.26	Positive
sib.MDD.var	0.0004	0.0004	0.002	0.24	?
sib.cov	-0.002	-0.002	0.003	-0.50	?
sib.LM1.var	0.05	0.05	0.01	3.88	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.51	Positive
resid.cov	0.0007	0.0007	0.005	0.15	Positive
resid.LM1.var	0.58	0.58	0.02	31.07	Positive

Variance component estimates from bivariate analysis of LM1 and MDD

4.B.6. Logical Memory Delayed and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.003	6.90	Positive
ped.cov	-0.003	-0.003	0.007	-0.41	Positive
ped.LM2D.var	0.27	0.27	0.03	9.14	Positive
OHH.MDD.var	0.002	0.002	0.002	1.17	?
OHH.cov	0.002	0.002	0.004	0.54	?
OHH.LM2D.var	0.003	0.003	0.02	0.19	?
YHH.MDD.var	0.003	0.003	0.002	1.27	Positive
YHH.cov	0.001	0.001	0.004	0.35	Positive
YHH.LM2D.var	0.01	0.01	0.02	0.89	Positive
sib.MDD.var	5.55e-07	5.55e-07	NA	NA	Boundary
sib.cov	-3.24e-04	-3.24e-04	0.003	-0.10	Positive
sib.LM2D.var	0.05	0.05	0.01	3.61	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	39.20	Positive
resid.cov	-0.005	-0.005	0.005	-1.04	Positive
resid.LM2D.var	0.60	0.60	0.02	32.01	Positive

Variance component estimates from bivariate analysis of LM2D and MDD

4.B.7. Mood Disorder Questionnaire and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.26	Positive
ped.cov	0.06	0.06	0.01	5.31	Positive
ped.MDQ.var	0.20	0.20	0.04	4.43	Positive
OHH.MDD.var	0.002	0.002	0.002	1.04	Positive
OHH.cov	-0.004	-0.004	0.006	-0.75	Positive
OHH.MDQ.var	0.06	0.06	0.03	2.23	Positive
YHH.MDD.var	0.003	0.003	0.002	1.25	Positive
YHH.cov	-0.004	-0.004	0.006	-0.70	Positive
YHH.MDQ.var	0.04	0.04	0.03	1.48	Positive
sib.MDD.var	0.0003	0.0003	0.002	0.18	?
sib.cov	-0.002	-0.002	0.005	-0.31	?
sib.MDQ.var	0.03	0.03	0.02	1.40	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.63	Positive
resid.cov	0.02	0.02	0.007	2.76	Positive
resid.MDQ.var	0.63	0.63	0.03	22.04	Positive

Variance component estimates from bivariate analysis of MDQ and MDD

4.B.8. Mill Hill Vocabulary and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.003	6.93	Positive
ped.cov	-0.003	-0.003	0.007	-0.41	Positive
ped.MHV.var	0.32	0.32	0.02	13.36	Positive
OHH.MDD.var	0.002	0.002	0.002	1.14	Positive
OHH.cov	0.002	0.002	0.004	0.58	Positive
OHH.MHV.var	0.11	0.11	0.01	8.10	Positive
YHH.MDD.var	0.002	0.002	0.002	1.25	Positive
YHH.cov	0.001	0.001	0.004	0.25	Positive
YHH.MHV.var	0.16	0.16	0.01	10.94	Positive
sib.MDD.var	5.58e-07	5.58e-07	NA	NA	Boundary
sib.cov	-3.89e-04	-3.89e-04	0.003	-0.13	Positive
sib.MHV.var	0.11	0.11	0.01	9.93	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	39.21	Positive
resid.cov	0.005	0.005	0.004	1.08	Positive
resid.MHV.var	0.16	0.16	0.01	10.67	Positive

Variance component estimates from bivariate analysis of MHV and MDD

4.B.9. Schizotypal Personality Questionnaire and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.24	Positive
ped.cov	0.04	0.04	0.01	4.08	Positive
ped.SPQ.var	0.19	0.19	0.04	4.45	Positive
OHH.MDD.var	0.002	0.002	0.002	1.15	Positive
OHH.cov	0.006	0.006	0.006	1.10	Positive
OHH.SPQ.var	0.15	0.15	0.02	6.03	Positive
YHH.MDD.var	0.002	0.002	0.002	1.12	Positive
YHH.cov	0.005	0.005	0.006	0.85	Positive
YHH.SPQ.var	0.12	0.12	0.03	4.49	Positive
sib.MDD.var	0.0002	0.0002	0.002	0.13	?
sib.cov	-0.0002	-0.0002	0.005	-0.04	?
sib.SPQ.var	0.02	0.02	0.02	1.24	?
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.71	Positive
resid.cov	0.009	0.009	0.007	1.32	Positive
resid.SPQ.var	0.51	0.51	0.03	18.15	Positive

Variance component estimates from bivariate analysis of SPQ and MDD

4.B.10. Verbal Fluency and MDD

effect	gamma	component	std.error	z.ratio	constraint
ped.MDD.var	0.02	0.02	0.004	5.30	Positive
ped.cov	-0.006	-0.006	0.007	-0.89	Positive
ped.VF.var	0.37	0.37	0.03	13.24	Positive
OHH.MDD.var	0.002	0.002	0.002	1.12	Positive
OHH.cov	0.003	0.003	0.004	0.81	Positive
OHH.VF.var	0.06	0.06	0.02	3.55	Positive
YHH.MDD.var	0.003	0.003	0.002	1.19	Positive
YHH.cov	0.004	0.004	0.004	1.00	Positive
YHH.VF.var	0.11	0.11	0.02	6.16	Positive
sib.MDD.var	6.18e-05	6.18e-05	0.002	0.04	?
sib.cov	0.002	0.002	0.003	0.51	Positive
sib.VF.var	0.05	0.05	0.01	3.54	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDD.var	0.09	0.09	0.002	37.54	Positive
resid.cov	0.002	0.002	0.005	0.36	Positive
resid.VF.var	0.38	0.38	0.02	20.34	Positive

Variance component estimates from bivariate analysis of VF and MDD (full model did not converge)

effect	gamma	component	std.error	z.ratio	constraint
sib.VF	0.04	0.04	0.01	3.53	Positive
OHH.VF	0.06	0.06	0.02	3.55	Positive
OHH.cov	0.003	0.003	0.004	0.68	Positive
OHH.MDD	0.002	0.002	0.002	1.17	Positive
YHH.VF	0.11	0.11	0.02	6.16	Positive
YHH.cov	0.003	0.003	0.004	0.83	Positive
YHH.MDD	0.003	0.003	0.002	1.28	Positive
ped.VF	0.37	0.37	0.03	13.24	Positive
ped.cov	-0.004	-0.004	0.006	-0.67	Positive
ped.MDD	0.02	0.02	0.003	6.89	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.VF	0.38	0.38	0.019	20.34	Positive
resid.cov	0.003	0.003	0.005	0.56	Positive
resid.MDD	0.09	0.09	0.002	39.21	Positive

Variance component estimates from bivariate analysis of VF, MDD

5. Variance component estimates for composite traits

5.A. Bivariate analyses between quantitative traits

5.A.1. GHQ and EPQN

effect	gamma	component	std.error	z.ratio	constraint
OHH.GHQ.var	0.06	0.06	0.02	3.41	Positive
OHH.cov	0.03	0.03	0.01	2.22	Positive
OHH.EPQN.var	0.04	0.04	0.02	2.21	Positive
YHH.GHQ.var	0.07	0.07	0.02	3.61	Positive
YHH.cov	0.03	0.03	0.01	2.02	Positive
YHH.EPQN.var	0.03	0.03	0.02	1.52	Positive
sib.GHQ.var	0.03	0.03	0.01	1.85	Positive
sib.cov	0.01	0.01	0.01	1.14	Positive
sib.EPQN.var	0.02	0.02	0.01	1.59	Positive
ped.GHQ.var	0.12	0.12	0.03	3.92	Positive
ped.cov	0.15	0.15	0.01	6.55	Positive
ped.EPQN.var	0.22	0.22	0.03	7.66	Positive
resid.var	1.00	1.000	NA	NA	Fixed
resid.GHQ.var	0.71	0.71	0.02	34.78	Positive
resid.cov	0.23	0.23	0.02	14.91	Positive
resid.EPQN.var	0.65	0.65	0.02	33.01	Positive

Variance component estimates from bivariate analysis of GHQ and EPQN.

5.A.2. GHQ and MDQ

effect	gamma	component	std.error	z.ratio	constraint
OHH.GHQ.var	0.06	0.06	0.02	3.37	Positive
OHH.cov	0.05	0.05	0.02	2.61	Positive
OHH.MDQ.var	0.07	0.07	0.03	2.62	Positive
YHH.GHQ.var	0.07	0.07	0.02	3.57	Positive
YHH.cov	0.02	0.02	0.02	1.13	Positive
YHH.MDQ.var	0.04	0.04	0.03	1.64	Positive
sib.GHQ.var	0.03	0.03	0.01	1.89	Positive
sib.cov	0.01	0.01	0.01	0.97	Positive
sib.MDQ.var	0.02	0.02	0.02	0.79	Positive
ped.GHQ.var	0.12	0.12	0.03	3.89	Positive
ped.cov	0.14	0.14	0.03	4.64	Positive
ped.MDQ.var	0.20	0.20	0.04	4.64	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.GHQ.var	0.71	0.71	0.02	34.87	Positive
resid.cov	0.07	0.07	0.02	3.53	Positive
resid.MDQ.var	0.63	0.63	0.03	21.96	Positive

Variance component estimates from bivariate analysis of GHQ and MDQ.

5.A.3. GHQ and SPQ

effect	gamma	component	std.error	z.ratio	constraint
OHH.GHQ.var	0.06	0.06	0.02	3.38	Positive
OHH.cov	0.07	0.07	0.02	4.03	Positive
OHH.SPQ.var	0.15	0.17	0.02	6.08	Positive
YHH.GHQ.var	0.06	0.06	0.02	3.47	Positive
YHH.cov	0.06	0.06	0.02	3.15	Positive
YHH.SPQ.var	0.12	0.12	0.03	4.57	Positive
sib.GHQ.var	0.03	0.03	0.01	1.95	Positive
sib.cov	0.009	0.009	0.01	0.66	Positive
sib.SPQ.var	0.02	0.02	0.02	0.88	Positive
ped.GHQ.var	0.12	0.12	0.03	3.79	Positive
ped.cov	0.11	0.11	0.03	3.76	Positive
ped.SPQ.var	0.19	0.19	0.04	4.64	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.GHQ.var	0.72	0.72	0.02	35.064	Positive
resid.cov	0.09	0.09	0.02	4.60	Positive
resid.SPQ.var	0.51	0.51	0.03	18.23	Positive

Variance component estimates from bivariate analysis of SPQ, GHQ

5.A.4. EPQN and MDQ

effect	gamma	component	std.error	z.ratio	constraint
OHH.EPQN.var	0.03	0.03	0.02	2.01	Positive
OHH.cov	0.04	0.04	0.02	2.35	Positive
OHH.MDQ.var	0.07	0.07	0.02	2.65	Positive
YHH.EPQN.var	0.02	0.02	0.018	1.37	Positive
YHH.cov	0.02	0.02	0.02	1.08	Positive
YHH.MDQ.var	0.04	0.04	0.03	1.54	Positive
sib.EPQN.var	0.02	0.02	0.01	1.53	Positive
sib.cov	0.02	0.02	0.01	1.32	Positive
sib.MDQ.var	0.02	0.02	0.02	0.88	Positive
ped.EPQN.var	0.23	0.23	0.03	7.73	Positive
ped.cov	0.11	0.11	0.03	3.75	Positive
ped.MDQ.var	0.19	0.19	0.04	4.42	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.EPQN.var	0.65	0.65	0.02	33.14	Positive
resid.cov	0.14	0.14	0.02	7.37	Positive
resid.MDQ.var	0.63	0.63	0.03	22.31	Positive

Variance component estimates from bivariate analysis of EPQN and MDQ.

5.A.5. EPQN and SPQ

effect	gamma	component	std.error	z.ratio	constraint
OHH.EPQN.var	0.03	0.03	0.02	2.13	Positive
OHH.cov	0.05	0.05	0.02	2.99	Positive
OHH.SPQ.var	0.14	0.14	0.02	5.81	Positive
YHH.EPQN.var	0.02	0.02	0.02	1.38	Positive
YHH.cov	0.04	0.04	0.02	2.41	Positive
YHH.SPQ.var	0.12	0.12	0.03	4.53	Positive
sib.EPQN.var	0.02	0.02	0.01	1.55	Positive
sib.cov	-0.002	-0.002	0.01	-0.18	Positive
sib.SPQ.var	0.02	0.02	0.02	1.03	Positive
ped.EPQN.var	0.22	0.22	0.03	7.60	Positive
ped.cov	0.14	0.14	0.03	4.81	Positive
ped.SPQ.var	0.20	0.20	0.04	4.74	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.EPQN.var	0.65	0.65	0.02	33.25	Positive
resid.cov	0.18	0.18	0.02	9.17	Positive
resid.SPQ.var	0.51	0.51	0.03	18.52	Positive

Variance component estimates from bivariate analysis of EPQN and SPQ.

5.A.6. MDQ and SPQ

effect	gamma	component	std.error	z.ratio	constraint
OHH.MDQ.var	0.07	0.07	0.03	2.89	Positive
OHH.cov	0.08	0.08	0.02	3.97	Positive
OHH.SPQ.var	0.15	0.15	0.02	5.99	Positive
YHH.MDQ.var	0.03	0.03	0.03	1.34	Positive
YHH.cov	0.05	0.05	0.02	2.34	Positive
YHH.SPQ.var	0.11	0.11	0.03	4.27	Positive
sib.MDQ.var	0.02	0.02	0.02	0.97	Positive
sib.cov	0.02	0.02	0.02	1.19	Positive
sib.SPQ.var	0.02	0.02	0.02	1.13	Positive
ped.MDQ.var	0.19	0.19	0.04	4.40	Positive
ped.cov	0.15	0.15	0.03	4.25	Positive
ped.SPQ.var	0.20	0.20	0.04	4.64	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.MDQ.var	0.63	0.63	0.03	22.20	Positive
resid.cov	0.20	0.20	0.02	9.07	Positive
resid.SPQ.var	0.519	0.51	0.03	18.10	Positive

Variance component estimates from bivariate analysis of MDQ and SPQ.

5.B. Univariate analyses

5.B.1. Sum of Z-Scores (GHQ + EPQN)

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.08	0.15	0.05	3.09	Positive
YHH.var	0.07	0.13	0.05	2.50	Positive
sib.var	0.04	0.08	0.04	1.90	Positive
ped.var	0.36	0.66	0.09	7.38	Positive
resid.var	1.00	1.83	0.06	31.00	Positive

Variance component estimates from univariate analysis of SoS.

5.B.2. Genetic composite trait 3 (GHQ, EPQN, MDQ)

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.13	0.13	0.05	2.84	Positive
YHH.var	0.04	0.03	0.05	0.72	Positive
sib.var	0.05	0.05	0.04	1.37	Positive
ped.var	0.52	0.50	0.08	6.15	Positive
resid.var	1.00	0.97	0.05	18.72	Positive

Variance component estimates from univariate analysis of Gen3.

5.B.3. Genetic composite trait 4 (GHQ, EPQN, MDQ, SPQ)

effect	gamma	component	std.error	z.ratio	constraint
OHH.var	0.21	0.23	0.06	4.04	Positive
YHH.var	0.11	0.12	0.06	1.94	Positive
sib.var	0.05	0.05	0.04	1.21	Positive
ped.var	0.59	0.64	0.10	6.39	Positive
resid.var	1.00	1.08	0.06	16.93	Positive

Variance component estimates from univariate analysis of Gen4.

5.C. Bivariate analyses between composite traits and MDD

5.C.1. Sum of Z-Scores and MDD

effect	gamma	component	std.error	z.ratio	constraint
OHH.SoS	0.18	0.18	0.05	3.53	Positive
OHH.cov	0.01	0.01	0.007	1.40	Positive
OHH.MDD	0.003	0.003	0.002	1.46	Positive
YHH.SoS	0.16	0.16	0.05	2.92	Positive
YHH.cov	0.01	0.01	0.0083	1.88	Positive
YHH.MDD	0.003	0.003	0.002	1.46	Positive
sib.SoS	0.12	0.12	0.04	2.90	?
sib.cov	-0.002	-0.002	0.006	-0.25	?
sib.MDD	0.0004	0.0004	0.002	0.23	Boundary
ped.SoS	0.59	0.59	0.09	6.61	Positive
ped.cov	0.07	0.07	0.01	5.70	Positive
ped.MDD	0.02	0.02	0.004	5.01	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.SoS	1.81	1.81	0.06	30.88	Positive
resid.cov	0.07	0.07	0.009	8.45	Positive
resid.MDD	0.09	0.09	0.002	37.59	Positive

Variance component estimates from bivariate analysis of SoS, MDD

5.C.2. Genetic composite trait 3 and MDD

effect	gamma	component	std.error	z.ratio	constraint
sib.Gen3	0.06	0.06	0.03	1.75	Positive
OHH.Gen3	0.12	0.12	0.04	2.71	Positive
OHH.cov	0.002	0.002	0.008	0.20	Positive
OHH.MDD	0.002	0.002	0.002	1.25	Positive
YHH.Gen3	0.04	0.04	0.05	0.88	Positive
YHH.cov	0.001	0.001	0.008	0.16	Positive
YHH.MDD	0.003	0.003	0.002	1.33	Positive
ped.Gen3	0.50	0.50	0.08	6.34	Positive
ped.cov	0.09	0.09	0.01	7.83	Positive
ped.MDD	0.02	0.02	0.003	6.87	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.Gen3	0.97	0.97	0.05	19.12	Positive
resid.cov	0.05	0.05	0.009	5.76	Positive
resid.MDD	0.097	0.09	0.002	39.36	Positive

Variance component estimates from bivariate analysis of Gen3, MDD (conditional)

5.C.3. Genetic composite trait 4 and MDD

effect	gamma	component	std.error	z.ratio	constraint
sib.Gen4	0.07	0.07	0.04	1.63	Positive
OHH.Gen4	0.23	0.23	0.06	4.01	Positive
OHH.cov	-0.005	-0.005	0.009	-0.54	Positive
OHH.MDD	0.002	0.002	0.002	1.23	Positive
YHH.Gen4	0.13	0.13	0.06	2.16	Positive
YHH.cov	-0.003	-0.003	0.009	-0.29	Positive
YHH.MDD	0.003	0.003	0.002	1.27	Positive
ped.Gen4	0.63	0.63	0.10	6.45	Positive
ped.cov	-0.10	-0.10	0.01	-7.77	Positive
ped.MDD	0.02	0.02	0.003	6.88	Positive
resid.var	1.00	1.00	NA	NA	Fixed
resid.Gen4	1.08	1.08	0.06	17.23	Positive
resid.cov	-0.05	-0.05	0.01	-4.96	Positive
resid.MDD	0.09	0.09	0.002	39.42	Positive

Variance component estimates from bivariate analysis of Gen4, MDD