Population Studies of the Heritable Influences on the Mind and Brain

Philip R. Jansen

Population Studies of the Heritable Influences on the Mind and Brain

Philip R. Jansen

ISBN: 978-94-028-1554-2

The work presented in this thesis was made possible by financial support of the Sophia Foundation for Scientific Research (SSWO: Sophia Stiching voor Wetenschappelijk onderzoek), grant number S14-27.

Cover design: *'The colorful mind and the calculating brain'* by Ilona A. Dekkers
Layout: Philip R. Jansen
Printing: Ipskamp Printing, Enschede, the Netherlands

Population Studies of the Heritable Influences on the Mind and Brain

Populatie studies naar de erfelijke invloeden op de geest en het brein

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

vrijdag 5 juli 2019 om 15.30 uur

Philip Rombout Jansen
geboren te Utrecht

Erasmus
University
Rotterdam

Promotiecommissie

Promotoren:	Prof. dr. F.C. Verhulst Prof. dr. D. Posthuma
Overige leden:	Prof. dr. A.G. Uitterlinden Prof. dr. M.C. O'Donovan Prof. dr. H.E. Hulshoff Pol
Copromotoren	Dr. T.J.H. White Dr. T.J.C. Polderman

Paranimfen:
Victor A. Jansen
Jorim J. Tielbeek
Miguel C. Jansen (reserve)

Table of Contents

Chapter 1:	General introduction	7
Part I:	**Population-based Genetic Studies of Complex Traits**	
Chapter 2:	Polygenic Scores for Schizophrenia and Educational Attainment are Associated with Behavioural Problems in Early Childhood in the General Population *(published in Journal of Child Psychology and Psychiatry)*	23
Chapter 3:	Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways *(published in Nature Genetics)*	33
Chapter 4:	Meta-analysis of Genome-wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways *(published in Nature Genetics)*	59
Chapter 5:	Genome-wide Association Meta-analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence *(published in Nature Genetics)*	91
Part II:	**Brain Imaging and Genetics**	
Chapter 6:	Incidental Findings on Brain Imaging in the General Pediatric Population *(published in the New England Journal of Medicine)*	123
Chapter 7:	Cross-trait Analysis of Brain Volume and Intelligence Identifies Shared Genomic Loci and Genes *(Manuscript in preparation)*	133
Chapter 8:	Common Polygenic Variations for Psychiatric Disorders and Cognition in Relation to Brain Morphology in the General Pediatric Population *(published in Journal of the American Academy of Child and Adolescent Psychiatry)*	165
Chapter 9:	Polygenic Scores for Neuropsychiatric Traits and White Matter Microstructure in the Pediatric Population *(published in Biological Psychiatry: Cognitive Neuroscience and Neuroimaging)*	175
Part III		
Chapter 10:	General discussion and summary	189
Summary	Dutch summary	203
Addendum	Publication list	207
	PhD portfolio	211

Addendum Authors and affiliations 215

Dankwoord (word of gratitude) 217

Acknowledgements 221

About the author 223

Chapter 1: General Introduction

Introduction

Genetic studies have led to many novel insights into genetic determinants of a wide variety of human behavior. Recent genome-wide association studies (GWAS) have tremendously increased the rapid pace of scientific discovery, leading to thousands of novel loci being discovered that explain both normal variation and disease risk[1]. Although we have learned much about which genes and trait mechanisms are important in explaining individual variation, much of the heritability of these traits is yet unknown. This thesis aims to extend these existing novel discoveries by demonstrating novel loci, genes, pathways and mechanisms that underlie a range of human complex behavioral phenotypes.

1. Population genetic studies of complex traits

1.1 From genes to behavior

Human behavior has long been known to be strongly influenced by genetic factors[2]. After groundbreaking work by Gregor Mendel describing fundamental laws of inheritance in his 'Experiments on Plant Hybridization' in 1866 **(Fig. 1a)**, it was his half-cousin Sir Francis Galton who was among the first to study the heredity of human behavior[3]. Early work by Galton in the second half of the 19th century mainly focused on addressing the question whether behavior in offspring resembles that of their parents[4]. As such, he was among the first to consider the role of relatedness in the presence of certain behavioral characteristics[5]. Galton proposed that twins may be used to study the role of 'nature and nurture' (a term introduced by Galton himself) in the origin of behavioral differences[6]. However, he did not consider the comparison between monozygotic (i.e. identical) and dizygotic (i.e., fraternal) twins[7] as he was unaware of the genetic differences between these two types of twins. The modern twin study design emerged several decades later when, in 1924, (**Fig. 1a**) two researchers, Curtis Merriman[8] and Hermann Siemens[9] independently published descriptions of estimates of heritability by comparing twins[7]. Using multiple sophisticated twin modelling designs that were developed in the following decades, it was concluded that all behavioral traits are more or less heritable, often referred to as the 'first law of behavioral genetics'[10].

In 2015 (**Fig. 1a**), a large meta-analysis summarized the results of all twin studies until that time of 17,804 traits[11]: the influence of genetic factors estimated over 2,748 twin study publications was 49%, an equal contribution of 'nature' and 'nurture' to the variation in human characteristics. After the importance of heredity in behavioral differences was ascertained in the twin study era, the first attempts of identifying specific genes involved in complex traits were carried out in candidate gene studies that aimed to link single genes to complex behavioral traits based on prior knowledge or expert opinion[12]. Although the concept of finding genes for disease by comparing genetic variants between patients and controls was promising, efforts from this type of hypothesis-driven gene finding studies in behavioral genetics would often not replicate in subsequent analyses[13], likely due to low statistical power and insufficient correction for the confounding effects of population stratification[14] (i.e. false-positive associations due to allele frequency and phenotypic differences between ethnicities).

More recently, a paradigm shift took place when around the year 2007 (**Fig. 1a**) a hypothesis-free design that simultaneously probes variants across the entire width of the genome in a genome-wide association study (GWAS)[15] design became more common, that quickly replaced prior hypothesis-driven approaches.

1.2 Genome-wide association studies

In the last decade, GWAS has proven to be the most important driver of scientific discovery in the field of complex trait genetics[1,16], and to date thousands of significant associations between single nucleotide polymorphisms (SNPs) and behavior have been found. Although these studies provide major steps forward in deciphering the genetic architecture of complex traits, for the majority of these traits most genomic loci are still undiscovered[17], leading to a gap between the estimated and explained heritability ('missing heritability'[18]). Given the small effect sizes of variants that were observed, it has been suggested that most of the variation that can theoretically be explained by SNPs is in fact hidden and may only be discovered when sample sizes increase[19,20].

In a period of just 10 years, the sample sizes of GWAS in humans has increased by approximately 200-fold, from a couple of thousand individuals in the first well-performed GWAS in 2007[21], to over a million in recent studies of educational attainment[22], blood pressure[23] and alcohol use[24]

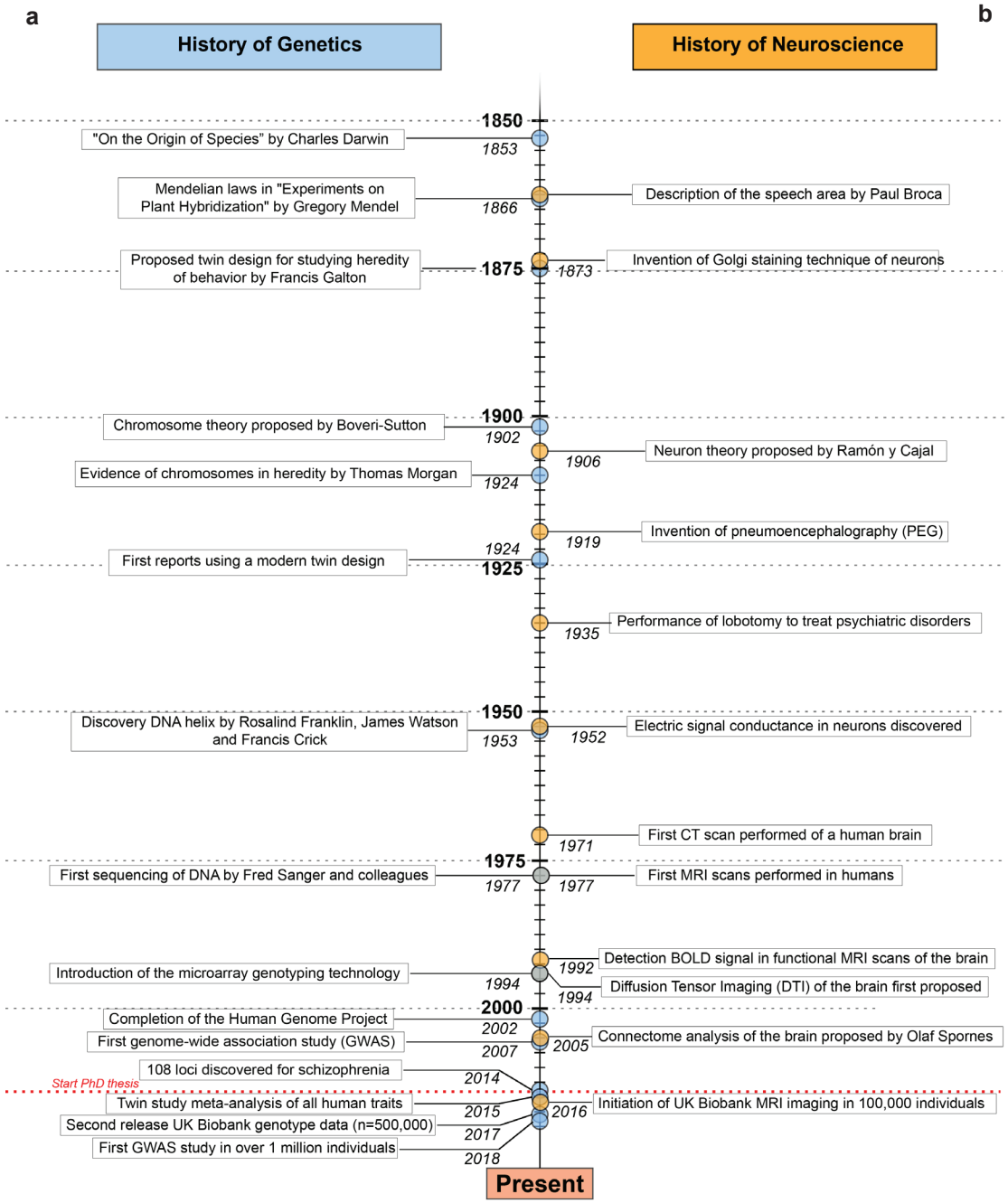


Fig. 1 | 'Standing on the shoulders of giants', a timeline of events in scientific discovery. A chronological overview of scientific milestones in the field of (**a**) genetics, and (**b**) neuroscientific research.

(**Fig. 1a**), equivalent to a population doubling time every 1.3 years (a trend more rapid than Moore's law[25], the predicted doubling time of the number of transistors on a chip, a measure of processor speed).

Recently, two important big data initiatives have strongly accelerated this trend, and have led to a recent 'Cambrian explosion' in genetic discoveries in the last two to three years:

The first is The UK Biobank (UKB)[26], a large population-based study that was initiated in 2005 collects health-related data in over 500,000 individuals. UKB released genotype data of the full cohort to researchers with

an approved application in 2017[27] (**Fig 1a**). This unique dataset of 500,000 genotypes in combination with the high dimensionality of phenotypic data collected in UKB participants allows well-powered GWAS on a massive collection of behavioral outcomes. The data release was a highly anticipated event by many research groups around the world that had approved access to the data, which subsequently continued the 'gold rush' in gene finding that was predicted ten years earlier[28,29]. UKB has proven to be a key data source not only in finding novel genes for certain behavior, but also for novel findings in the fields of neuroscience, neuroimaging and epidemiology.

The second is 23andMe, Inc., a direct-to-consumer genotyping company[30], that offers health information based on genome-wide genotyping. Customers of 23andMe are offered the possibility to opt-in for scientific research by completing online questionnaires related to their health. The customer database currently contains over 3 million individuals, many of which opted-in for participating in scientific research. Genome-wide analyses based on this data are performed on many traits by the 23andMe research team, and the full GWAS results are made available to researchers with an approved research project. Given the enormous and still increasing scale of the customer population, with potential sample sizes of millions of individuals worldwide, direct-to-consumer genotyping companies have been an important source of GWAS results for many behavioral traits and will continue to be in the years to come.

Although GWAS has been crucial in the search for genes related to behavior, points of critique and shortcomings to this approach exist that need to be overcome to release its full potential. Much of the genetic signal observed in GWAS has been found to be located within or near genes without an obvious link to functional mechanisms that may explain differences in phenotypes, which lead to difficulties interpreting these genes. Also, the overwhelming polygenic findings often do not lead to a clear coherence in gene functions between the many discovered genes identified by GWAS[31]. A possible explanation has been proposed by a putative 'omnigenic model'[32] that describes core-genes in a gene regulatory network of disease biology and many more peripheral genes in the network. The much larger number of peripheral genes that correlate with these core genes in the network may theoretically explain a significant part of the GWAS, which may explain the observation that many traits are highly polygenic. This proposed model has been topic of debate in the genetics community, as it is argued to be an oversimplified representation of the complex biology of behavior[33] and provides only a vague distinction between core and peripheral genes. Although the validity of the omnigenic has been disputed, it emphasizes that a more sophisticated network analyses of the interplay between genes may aid in the interpretation of GWAS findings[34]. This requires more advanced methods and integration of high dimensional omics data to study the effects of genetic variants and genes on various levels of biological complexity[35].

1.3 Genetics of psychiatric disorders

For psychiatric disorders in particular, the last years have seen a tremendous increase in the understanding of the genetic architecture of psychiatric disorders, in which GWAS has served a major role. These debilitating disorders with often devastating consequences due to insufficient treatment strategies are long known to be among the most heritable traits[36]. Early applications of the GWAS methods did not prove to be successful, and clearly showed that collaborative efforts were necessary, imposed by nature due to the small effect sizes of individual genetic variants, to reach sufficient power for discovering these variants.

A major breakthrough in psychiatric genetics was achieved by a milestone paper in 2014 carried out by the Psychiatric Genomics Consortium (PGC[37]) that reported 108 locations in the DNA that influence the risk of schizophrenia[38.] This report, that has been cited almost 3,000 times in just a couple of years, laid the foundation for a new understanding of the genetics of schizophrenia. Importantly, the study paved the way for numerous follow-up studies that could build forth on these insights by studying various aspects of the disease, ranging from the functional impact of these loci[39] to epidemiological distributions of genetic risk[40], based on the publicly available GWAS results.

Similar successes followed for other common psychiatric disorders, including 44 risk loci for major depression[41], 12 risk loci for attention-deficit hyperactivity disorder[42] and the first risk locus for autism spectrum disorder[43]. These GWAS studies confirmed the highly complex genetic architecture of psychiatric disorders, influenced predominantly by many common variants of small effect and difficult to detect rare variants[44]. Continuation in this rapid pace of discovery of novel variants, genes and the pathways that modify disease risk of psychiatric disorders may hopefully lead to necessary improvements treatment options for patients that suffer from the incapacitating consequences from these disorders.

1.4 Post-GWAS annotation

The pace of discovery in recent GWAS studies has been difficult to keep up by (functional) follow-up studies to

better understand the characteristics of identified genetic variants[45–47]. An exciting area of genetic research emerged over the previous years that strives to extract more information about the genetics of complex traits from GWAS summary statistics alone[48]. This has the clear advantage of making individual level genotypes unnecessary by reducing the data to SNP associations alone (SNP effect sizes and/or *P*-values). A major development in this area is the LD Score regression software, which is able to estimate genetic overlap between traits based on GWAS results[49] and to partition SNP heritability over functional annotations of SNPs[50] and continuous SNP measures[51].

To further facilitate the process of performing follow-up analyses based on GWAS as the starting point, the online GWAS annotation tool FUMA[52] was developed in 2017 that automates the process of GWAS follow-up analysis, including functional annotation, gene-mapping and gene-expression analysis. This analysis pipeline has the clear advantage of automating each follow-up step, a reduction of the required time to run these analyses and having all individual datasets integrated into one easy to use tool.

The development of exciting novel follow-up analyses to fully understand the detected genetic signal, using GWAS as the starting point, is an important area of current-day genetic research and it is expected that more and more information can be extracted from GWAS results to better understand the biology of a particular trait or disease[53]. Among these follow-up analyses, the analysis of gene-expression from sequencing of messenger RNA (mRNA) in relevant postmortem tissue types has been central to understanding the role of genes implicated by GWAS. More recently, it has become technologically feasible to probe mRNA levels within single neuronal cell-types by single-cell RNA sequencing[54] (scRNA-seq). By integrating gene findings from GWAS with expression data in classes of neurons, neuronal cell-types can be identified that are likely to be involved in trait mechanisms on the microscopic scale. Availability of public RNA sequencing data has facilitated the analysis of neuronal enrichment of genetic signal[55]. Analysis of gene-expression data in psychiatric disorders has identified specific cell-types related to disease, such as cortical interneurons and pyramidal neurons in schizophrenia[56], and high expression of genes related to Parkinson's disease[57] in substantia nigra neurons.

1.5 Polygenic scoring

Among the primary aims of GWAS is to increase understanding of the role of the genome in illnesses, which could ultimately lead to new cures. In addition, knowing which variants influence disease risk may be used to estimate an individual's total genetic risk of disease, and identify groups that are most at risk and may benefit most from preventive measures[58.] An estimate of overall genetic risk, based on GWAS variants, can be derived by calculation of a polygenic risk score (PRS)[59], a sum of all risk alleles weighted by their effect size from GWAS. Whereas early PRS studies intuitively included only variants that passed a stringent genome-wide significance threshold[60] ($P < 5 \times 10^{-8}$), later studies showed that much more variation could be explained by a PRS that also included variants well below the significance line, suggesting that much of the genetic signal is concealed in variants that do not (yet) reach the Bonferroni threshold[38,61,62].

PRSs have shown to be promising for future clinical applications for common diseases, including breast cancer[63], diabetes and cardiovascular disease[64], showing comparable relative risk on the highest part of the continuum of genetic predisposition as several rare variants with a Mendelian (monogenic) pattern of disease risk[64]. The overall genetic load has several useful applications in genetic research, including new opportunities for studying gene-environment interactions[65] (PRS x E) and finding genetic overlap between seemingly unrelated traits[66]. Moreover, using a PRS for diseases in the general population can lead to novel hypotheses about associations between genetic predisposition for disease such as psychiatric disorders, and disease-related manifestations in the general population[59].

2. Genetics of complex traits (Part 1)

In the first part of the thesis, we study the genetics of three phenotypes related to behavior, which include: insomnia and sleep-related traits, neuroticism and depression, and intelligence.

2.1 Insomnia and sleep related traits

Sleep disorders including insomnia are known to be related to all sorts of health outcomes[67], ranging from disorders related to metabolic syndrome, including obesity[68], and type 2 diabetes[69], to overall happiness[70] and longevity[71].

In the past few years, large GWAS studies have led to important new insights into the genetics of sleep, including novel genes for insomnia[72–74], showing the strongest genetic signal located in the *MEIS1* gene. Also, many additional loci have been found for traits related to sleep, including sleep duration[75] and being a morning person[76]. The heritability of insomnia that can be explained by common variants (h^2_{SNP}) has been reported to be low, approximately 9%[73], as estimated by LD Score regression[77], which suggest that many more samples are needed to find more loci that influence the risk of insomnia.

2.2 Neuroticism and depression

Personality traits and mental health are strongly interrelated[78] which is reflected in strong overlap in genetic etiology[79]. Personality traits have historically been described according to a five-factor model along five axes of phenotypic variation ('The Big Five')[80]. Among these axes, neuroticism has been declared as the most prevalent trait as it plays a role in almost all personality types[81]. This personality feature is characterized by a tendency to have a negative perception and emotional response to negative stimuli[82] and predisposes to the development of depression[83]. Although this trait is considered universal[84], the term neuroticism may have origins in the Freudian concept of 'neurosis'[85–87], and leads back to the Greek word νεῦρον ('neuron', meaning 'tendon'), which shares this origin with the English word 'nervous' via the Latin word 'nervus'.

Neurotic personality traits are considered highly heritable[11], with heritability increasing with age,[88] and highly stable over a lifetime[88,89]. Subsequent gene finding studies of neuroticism test scores have shown that this personality trait is a highly polygenic trait[90] and that significant overlap exists in the polygenic background of the Big Five dimensions[91]. The increase in scale of GWAS in neuroticism research has led to a larger number of loci being discovered[79,92–94] and confirmed shared genetic risk with psychiatric disorders[79]. Thus far, these gene discovery efforts have not converged on a biologically interpretable model of neuroticism, which requires additional statistical power and detailed functional follow-up analysis.

2.3 Intelligence

The first descriptions of heritability in an individual's mental capacities was heavily studied by the aforementioned geneticist Francis Galton[5]. The first twin studies on the topic of intelligence confirmed a major role of genetic influence. Comparing twin study heritability estimates in children, adolescents and adults showed that the variance explained by genetic factors is not stable over the course of life, but increases as we age[95].

After several unsuccessful genome-wide attempts were done to capture genetic variants associated with intelligence[96,97], the first study that published an extensive list of loci and genes involved in intelligence was published in 2017[98]. These results implicated 18 genomic loci, and 52 genes associated by positional mapping and gene-based association testing. Interestingly, gene-set analyses[99] showed that processes related to the regulation of cell development were significantly enriched for genetic signal, which may indicate the importance of these genes early in development, suggesting an early-life window during which genes linked to intelligence operate. Although heritability had been estimated to be higher in adults compared to children[95], the genetic effects between these age groups were highly correlated (r_g = 0.89) as estimated by LD Score regression[100], which shows that largely the same genetic factors influence intelligence over the life course.

The increase in the number of detected loci with increasing sample size, suggests that more genomic loci are likely to be found currently residing under the genome-wide significance threshold[101]. This motivates further expansion of the sample size and statistical power. Investigating normal cognitive functioning, one of the most important functions of the human brain, has significant implications for our understanding of the normal neurobiology of the brain, and of diseases that are characterized by impairment in normal cognitive functioning, including schizophrenia[102], autism spectrum disorder[103], and Alzheimer disease[104].

3. Brain imaging studies (Part 2)

Brain imaging techniques such as magnetic resonance imaging (MRI) have proven to be an important modality in both clinical setting as in understanding the inner workings of the brain. In the second part of the thesis, we used multimodal brain imaging techniques to study associations between genetic variation and brain morphology.

3.1 Magnetic resonance imaging of the brain

In parallel to discoveries in behavioral genetics, the field of neuroscience has gone through equally rapid scientific advances in the last centuries (**Fig. 1b**). A major limitation during most of this time was that the secrets of the brain's inner workings were hidden in the skull, a 'locked safe' that was preventing scientists from making inferences about the brain's neurobiology in living persons. Until the mid 70's, the only way to receive a glimpse of the brain was through the use of pneumoencephalography (PEG), a procedure invented in 1919 (**Fig. 1b**) that included draining the cerebrospinal fluid and inserting air into the ventricles to increase the visibility of the brain on an x-ray image. The procedure was very painful, had a significant risk of morbidity[105], and resulted in death in 0.2% of the patients[106].

This method was rendered obsolete by the invention of computed tomography (CT) by Sir Godfrey Hounsfield, and magnetic resonance imaging (MRI) by Sir Peter Mansfield and Paul Lauterbur in the 70s that could non-invasively visualize the brain while in the skull. Whereas the first brain imaging studies using MRI were concerned with imaging the structure of the brain, more applications for using the versatile MR signal for studying complex

features of the brain followed over the years, including functional MRI[107] (fMRI) for studying brain activity in 1992, diffusion tensor imaging[108] (DTI) for tract analysis in 1994, and graph theoretical measures to analyze the brain as a connected network in 2005 (**Fig. 1b**)[109].

Next to the tremendous impact of brain imaging on diagnosis and decision making in neurological patients, the wide availability of brain imaging techniques has become the cornerstone in neuroscience for scientific investigations in the living brain. In epidemiology, population-based brain imaging is an indispensable tool for studying traditional risk factors related to brain development at early age[110], and brain disorders that occur later in life[111]. Examples of such large-scale MRI data collection efforts in different age categories are the Generation R cohort in The Netherlands[110], consisting of 4,000 scanned children, up to 10,000 adolescents in the ABCD study in the USA[112], and up to a planned 100,000 middle-aged adults from the UK Biobank, initiated in 2016 (**Fig. 1b**)[113]. Insights into the brains of these large numbers of individuals provide unique clues of normal brain function and associations with insulting risk factors. In addition, the enormous scale of these imaging efforts provides sufficient sample size and statistical power to study the modest individual associations between genetic variants and the brain.

Although the yield of knowledge from these large datasets is invaluable, high resolution imaging in large numbers of brain scans also has the downside of more unexpected findings that may either be unharmful or pose a significant health risk for the participant[114].

3.2 Genetics of imaging-derived phenotypes

With the advent of these large datasets and through international collaborations such as the ENIGMA consortium[115], it has become feasible to investigate the influence of genetic variants on variation in brain structure derived from brain images (imaging-derived phenotypes[116], or IDP). These morphological features of the brain are shown to be under strong genetic control[117], suggesting that gene-finding studies are likely to be successful when sufficient data are available.

Involvement of genetic factors in individual differences in brain structure is further evidenced by abnormal brain development as a commonly observed symptom of rare monogenic disorders, including a small brain (microcephaly) or an abnormally large brain (macrocephaly). Indeed, large imaging genetics studies of brain structure volume have shown evidence of a large number of common variants that moderate brain volumes[115,118], and implicate several cell-signaling pathways to be involved in brain volume regulation. Interestingly, these SNP associations observed in intracranial volume are correlated with several complex traits, including intelligence[119], educational attainment[79] and neuroticism[120].

3.3 Genetic risk and brain imaging

In vivo assessment of functional and structural characteristics of the brain provides a unique tool to investigate genetic influences on normal variation in brain morphology[121] and the effects of genetic risk variants on neurostructural phenotypes. Involvement of a wide variety of brain structures have repeatedly been reported in the brains of patients suffering from psychiatric disorders, including major depression[122,123], schizophrenia[124] and attention-deficit hyperactivity disorder[125]. Since these psychiatric disorders are highly heritable, neurobiological alterations on brain imaging in patients may suggest a shared genetic etiology between genetic liability for these traits and brain structure. Moreover, these alterations in brain structure have been found to a lesser degree in first-degree relatives of patients compared to controls without relatives with psychiatric disorders[126–128] and implies that genetic risk for the traits is closely related to variations in brain development.

By calculating a PRS in individuals with available genotype and brain imaging data, it has been shown that overall genetic risk for certain psychiatric disorders correlates with brain morphology and function in the general population[129]. The usefulness of these PRSs for finding associations between total genetic risk and variation in imaging-derived phenotypes of the brain is expected to increase as more large well-powered GWAS results are expected to become available in the coming years.

4. Thesis objectives

The main goal of this thesis is to find associations between genes and behavioral outcomes and to identify mechanisms that explain these associations along the 'gene-brain-behavior axis' (**Fig. 2a**). To bridge the gap between genes and behavior, the reported research projects include an extensive analysis pipeline that integrates various sorts of data from several large population-based samples, including genotype, brain imaging and questionnaire data, with bioinformatics data such as functional annotation, functional gene-sets and gene-expression data from tissues and single neuronal cell-types (**Fig. 2b**). GWAS results, the strength of SNP associations with a certain trait, are the most important starting point for all analyses that follow in this thesis, including several different gene-mapping strategies (positional mapping, eQTL mapping, gene-based association tests, chromatin-chro-

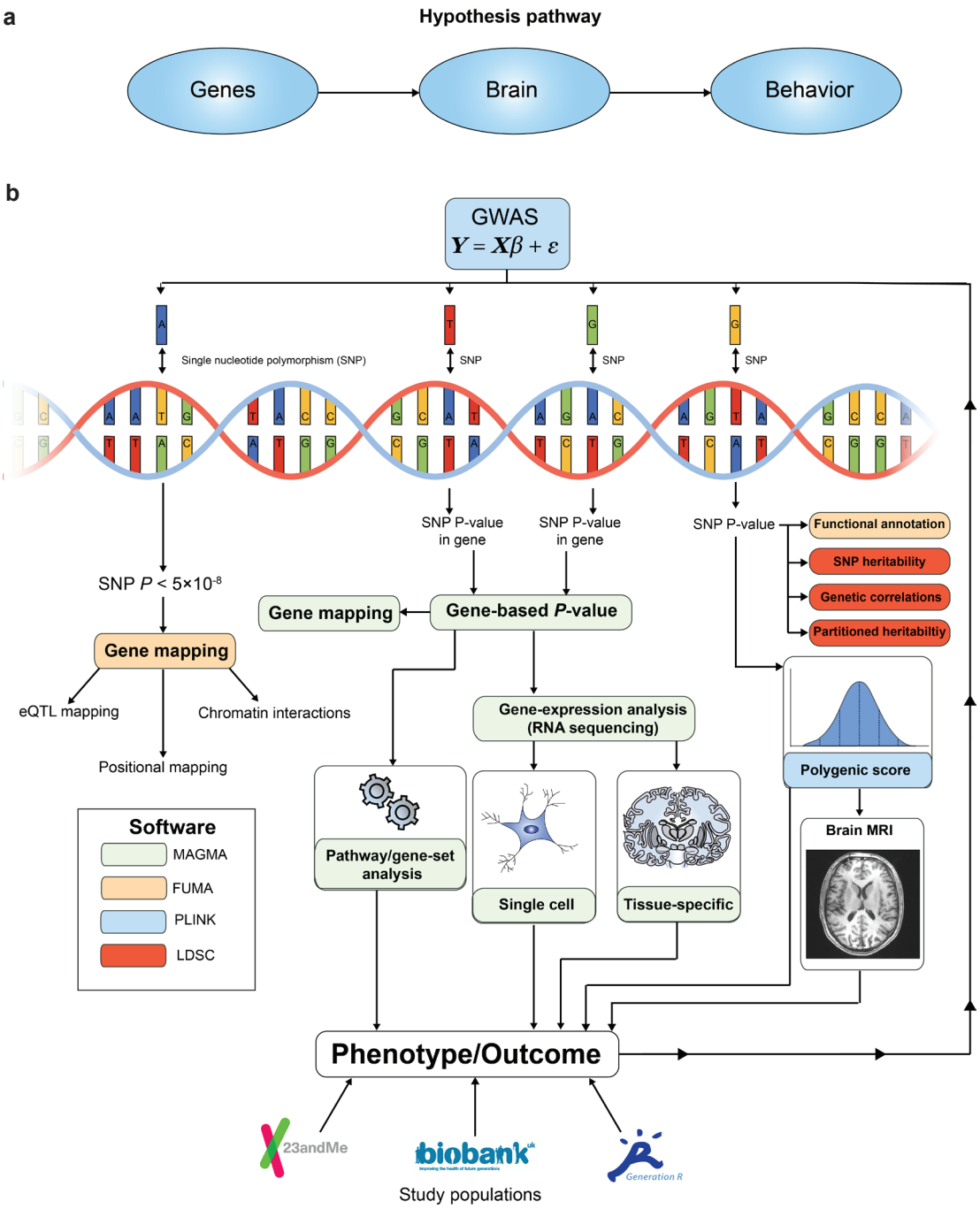


Fig. 2 | Overview of the main goals and analyses that were performed in this thesis. (**a**) The primary aim of this thesis is to find genes and pathways along the gene-brain-behavior axis that explain variation in behavioral outcomes. (**b**) Global analysis pipeline of analyses that were carried out, using GWAS results as the starting point. The rounded squares highlight analysis steps, while the colors indicate the software package that was used.

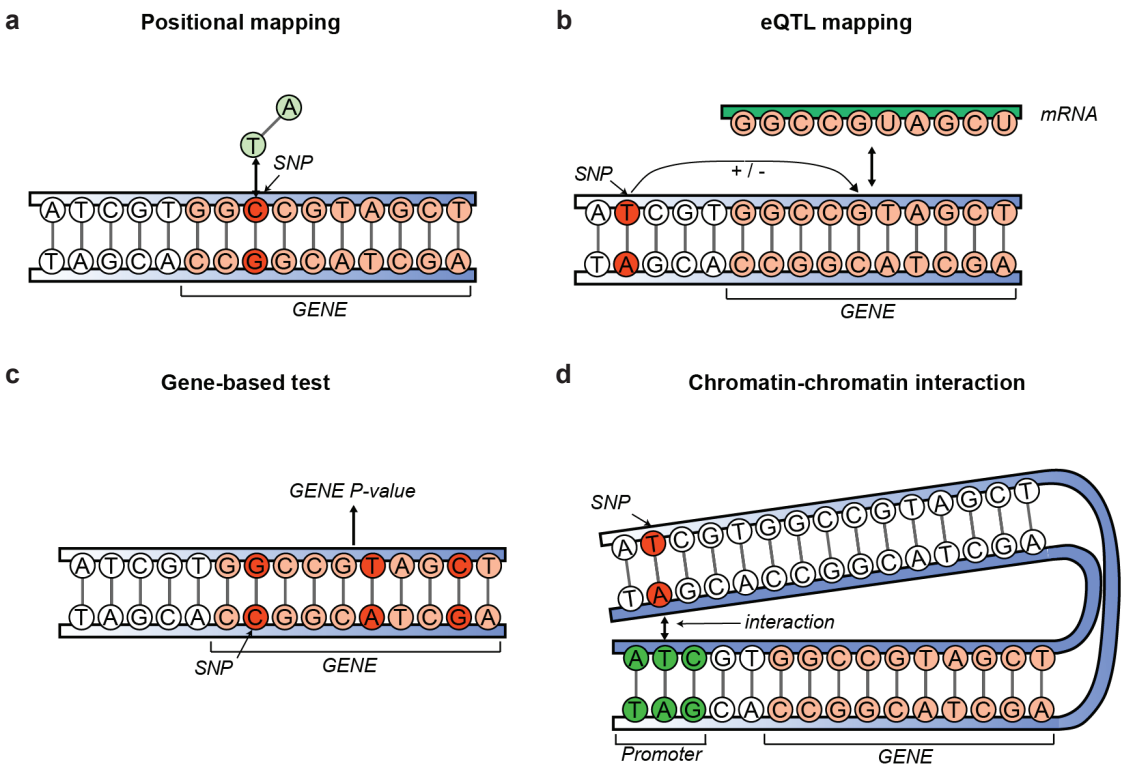


Fig. 3 | Gene-mapping methods used in this thesis. In the reported GWAS studies, we carried out a number of different gene-mapping strategies to link individual genetic variants to genes, including (**a**) positional mapping: mapping variants to genes by their physical position near or within a gene, (**b**) eQTL mapping: variants mapped to a gene through their influence on gene-expression of a gene, (**c**) gene-based association testing: combining variant associations within a gene into a gene-based test statistic, (**d**) chromatin-chromatin interaction: mapping a variant to a gene through a physical interaction with a close or distant region of the genome.

matin interaction gene mapping) that link GWAS results to gene discovery (**Fig. 3**), and polygenic risk score (PRS) analysis.

The objectives of this thesis can be further subdivided into two parts, including genetic studies of behavior, and (genetic) studies that use population-based brain MRI imaging data.

4.1 Genetic studies of behavior

The first aim was to investigate the genetic factors that explain individual differences in complex traits, and to generate hypotheses about functional pathways that are associated with these genetic factors.

In **Chapter 2**, we use polygenic scores for psychiatric disorders and educational attainment in a large population of children in the Generation R Study to investigate whether differences in genetic predisposition for psychiatric disorders are reflected in high levels of behavioral problems already in early childhood from the age of 3 years onwards.

In **Chapter 3,** we increase the scale of genome-wide analysis of insomnia by 10-fold and use genotype and sleep questionnaire data in over 1 million adult individuals in UK Biobank and 23andMe to study the genetics of insomnia complaints and several traits related to sleep in adults, including sleep duration, morningness and daytime napping.

In **Chapter 4**, we carry out the largest genome-wide analysis of neuroticism and depression in adults thus far by comparing results in participants of UK Biobank, 23andMe, the Psychiatric Genomics Consortium (PGC) and the Genetics of Personality Consortium (GPC) and study the biological implications of our findings by performing extensive functional interpretation of the results.

In **Chapter 5**, we aimed to further uncover the genetic architecture of intelligence by combining GWAS analyses of cognitive test scores in even a larger number of individuals from 14 pediatric and adult cohorts and perform extensive novel bioinformatic follow-up analyses of the results, including gene-set and gene-expression analysis.

4.2. Population imaging studies

In **Chapter 6**, we investigate incidental findings in the general pediatric population by reviewing brain scan data in over 4,000 children between the age of 8 and 12 years that participate in large-scale population-based research of the Generation R Study cohort.

In **Chapter 7**, we used GWAS results of brain imaging data collected in thousands of adult individuals in UK Biobank to find genes that explain the observed genetic overlap between brain volume and intelligence[130].

In **Chapter 8**, we used multimodal MRI data of the brain to investigate whether differences in polygenic predisposition based on common variants for psychiatric disorders and cognitive traits are associated with differences in macrostructural morphology of the brain on T1-weighted imaging.

In **Chapter 9**, we studied the association between polygenic scores for psychiatric disorders and cognitive outcomes, and white matter microstructure using diffusion tensor imaging (DTI) data of the brain in over 1,000 children from the Generation R Study between the age of 8 and 12 years.

References

1. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Plomin, R., DeFries, J. C., Knopik, V. S. & Neiderhiser, J. M. Top 10 replicated findings from behavioral genetics. *Perspect. Psychol. Sci.* **11**, 3–23 (2016).
3. Greenspan, R. J. The origins of behavioral genetics. *Curr. Biol.* **18**, R192–R198 (2008).
4. Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. Gt. Britain Irel.* **15**, 246–263 (1886).
5. Galton, F. Hereditary genius: An inquiry into its laws and consequences. 27, (Macmillan, 1869).
6. Galton, F. The history of twins, as a criterion of the relative powers of nature and nurture. *Fraser's Mag.* **12**, 566–576 (1875).
7. Rende, R. D., Plomin, R. & Vandenberg, S. G. Who discovered the twin method? *Behav. Genet.* **20**, 277–285 (1990).
8. Merriman, C. The intellectual resemblance of twins. *Psychol. Monogr.* **33**, i (1924).
9. Siemens, H. W. Die Zwillingspathologie: Ihre Bedeutung, ihre Methodik, ihre bisherigen Ergebnisse (Twin Pathology: Its Importance, Its Methodology, Its Previous Results). Berlin Verlag von Jul. Springer (1924).
10. Turkheimer, E. Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **9**, 160–164 (2000).
11. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
12. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* **3**, 391–397 (2002).
13. Hewitt, J. K. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav. Genet.* **42**, 1–2 (2012).
14. Rietveld, C. A. *et al.* Replicability and robustness of genome-wide-association studies for behavioral traits. *Psychol. Sci.* **25**, 1975–1986 (2014).
15. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
16. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
17. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
18. 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).
19. Sullivan, P. Don't give up on GWAS. *Mol. Psychiatry* **17**, 2–3 (2012).
20. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
21. Wellcome Trust Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
22. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nat. Genet.* **50**, 1112–1121 (2018).
23. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
24. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
25. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).
26. 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).
27. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).
28. Topol, E. J., Murray, S. S. & Frazer, K. A. The genomics gold rush. *JAMA* **298**, 218–221 (2007).
29. Dupuis, J. & O'Donnell, C. J. Interpreting results of large-scale genetic association studies: separating gold from fool's gold. *JAMA* **297**, 529–531 (2007).
30. Tung, J. Y. *et al.* Efficient replication of over 180 genetic associations with self-reported medical data. *PLoS One* **6**, e23473 (2011).
31. Reimers, M. A., Craver, C., Dozmorov, M., Bacanu, S.-A. & Kendler, K. S. The Coherence Problem: Finding Meaning in GWAS Complexity. *Behav. Genet.* **49** 1–9 (2018).
32. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
33. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
34. Leiserson, M. D. M., Eldridge, J. V, Ramachandran, S. & Raphael, B. J. Network analysis of GWAS data. *Curr. Opin. Genet. Dev.* **23**, 602–610 (2013).
35. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**, 441-458 (2015).
36. Plomin, R., Owen, M. J. & McGuffin, P. The genetic basis of complex human behaviors. *Science* **264**, 1733–1739 (1994).
37. Sullivan, P. F. *et al.* Psychiatric genomics: an update and an agenda. *Am. J. Psychiatry* **175**, 15–27 (2017).
38. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).

39. Fromer, M. *et al*. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
40. Colodro-Conde, L. *et al.* Association between population density and genetic risk for schizophrenia. *JAMA psychiatry* **75**, 901–910 (2018).
41. Wray, N. R. *et al*. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
42. Demontis, D. *et al*. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
43. ASD Consortium of the Psychiatric Genetics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 1–17 (2017).
44. Flint, J. Rare genetic variants and schizophrenia. *Nat. Neurosci.* **19**, 525-527 (2016).
45. Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N. & Geschwind, D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* **19**, 1397–1407 (2016).
46. Breen, G. *et al*. Translating genome-wide association findings into new therapeutics for psychiatry. *Nat. Neurosci.* **19**, 1392–1396 (2016).
47. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50**, 322–328 (2018).
48. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117-127 (2017).
49. Bulik-Sullivan, B. K. *et al*. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
50. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
51. Gazal, S. *et al*. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
52. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
53. Maier, R. M., Visscher, P. M., Robinson, M. R. & Wray, N. R. Embracing polygenicity: a review of methods and tools for psychiatric genetics research. *Psychol. Med.* **48**, 1055–1067 (2018).
54. Levsky, J. M., Shenoy, S. M., Pezo, R. C. & Singer, R. H. Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).
55. Macosko, E. Z. *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
56. Skene, N. G. *et al*. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
57. Capurro, A., Bodea, L.-G., Schaefer, P., Luthi-Carter, R. & Perreau, V. M. Computational deconvolution of genome wide expression data from Parkinson's and Huntington's disease brain tissues using population-specific expression analysis. *Front. Neurosci.* **8**, 441 (2015).
58. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **1** 581–590 (2018).
59. Wray, N. R. *et al*. Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
60. Machiela, M. J. *et al*. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epidemiol.* **35**, 506–514 (2011).
61. International Schizophrenia Consortium, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
62. Vilhjálmsson, B. J. *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
63. Maas, P. *et al.* Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302 (2016).
64. Khera, A. V *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 12191224 (2018).
65. Vinkhuyzen, A. A. E. & Wray, N. R. Novel directions for G× E analysis in psychiatry. *Epidemiol. Psychiatr. Sci.* **24**, 12–19 (2015).
66. Power, R. A. *et al*. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955 (2015).
67. Ohayon, M. M. Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med. Rev.* **6**, 97–111 (2002).
68. Hasler, G. *et al*. The association between short sleep duration and obesity in young adults: a 13-year prospective study. *Sleep* **27**, 661–666 (2004).
69. Vgontzas, A. N. *et al*. Insomnia with objective short sleep duration is associated with type 2 diabetes: a population-based study. *Diabetes Care* **32** 1980–1985 (2009).
70. Kao, C.-C., Huang, C.-J., Wang, M.-Y. & Tsai, P.-S. Insomnia: prevalence and its impact on excessive daytime sleepiness and psychological well-being in the adult Taiwanese population. *Qual. Life Res.* **17**, 1073–1080 (2008).
71. Chilcott, L. A. & Shapiro, C. M. The socioeconomic impact of insomnia. *Pharmacoeconomics* **10**, 1–14 (1996).
72. Lane, J. M. *et al.* Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281 (2017).
73. Hammerschlag, A. R. *et al*. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49**, 1584–1592 (2017).
74. Stein, M. B. *et al*. Genome-wide analysis of insomnia disorder. *Mol. Psychiatry* **23**, 2238–2250 (2018).
75. Jones, S. E. *et al*. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. *PLoS Genet.* **12**, e1006125 (2016).
76. Hu, Y. *et al*. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat. Commun.* **7**, 10448 (2016).
77. Bulik-Sullivan, B. K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
78. Lahey, B. B. Public health significance of neuroticism. American *Psychologist* **64**, 241–256 (2009).
79. Okbay, A. *et al*. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).
80. Poropat, A. E. A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* **135**, 322–338 (2009).
81. Costa, P. T. & McCrae, R. R. Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *J. Pers. Soc. Psychol.* **54**, 853–863 (1988).

82. Costa Jr, P. T. & McCrae, R. R. Four ways five factors are basic. *Pers. Individ. Dif.* **13**, 653–665 (1992).
83. Saklofske, D. H., Kelly, I. W. & Janzen, B. L. Neuroticism, depression, and depression proneness. *Pers. Individ. Dif.* **18**, 27–31 (1995).
84. McCrae, R. R. & Costa Jr, P. T. Personality trait structure as a human universal. *Am. Psychol.* **52**, 509516 (1997).
85. Freud, S. in The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIX (1923-1925): The Ego and the Id and Other Works 181–188 (1961).
86. Freud, S. From the history of an infantile neurosis. *Stand. Ed.* **17**, 124 (1918).
87. Freud, S. in The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume X (1909): Two Case Histories ('Little Hans' and the 'Rat Man') 151–318 (1955).
88. Wray, N. R., Birley, A. J., Sullivan, P. F., Visscher, P. M. & Martin, N. G. Genetic and phenotypic stability of measures of neuroticism over 22 years. *Twin Res. Hum. Genet.* **10**, 695–702 (2007).
89. Roberts, B. W. & DelVecchio, W. F. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychol. Bull.* **126**, 3–25 (2000).
90. De Moor, M. H. M. *et al*. Meta-analysis of genome-wide association studies for personality. *Mol. Psychiatry* **17**, 337–349 (2012).
91. Lo, M.-T. *et al*. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2017).
92. Luciano, M. *et al*. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).
93. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & van der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun*. **9**, 905 (2018).
94. Smith, D. J. *et al*. Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Mol. Psychiatry* **21**, 1–9 (2016).
95. Deary, I. J., Penke, L. & Johnson, W. The neuroscience of human intelligence differences. *Nat. Rev. Neurosci.* **11**, 201–211 (2010).
96. Davies, G. *et al*. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
97. Davies, G. *et al*. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N= 112 151). *Mol. Psychiatry* **21**, 758–767 (2016).
98. Sniekers, S. *et al*. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
99. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
100. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
101. Holland, D. *et al*. Beyond SNP Heritability: Polygenicity and Discoverability Estimated for Multiple Phenotypes with a Univariate Gaussian Mixture Model. *bioRxiv* (2018).
102. Green, M. F., Horan, W. P. & Lee, J. Social cognition in schizophrenia. *Nat. Rev. Neurosci.* **16**, 620–631(2015).
103. Demetriou, E. A. *et al.* Autism spectrum disorders: a meta-analysis of executive function. *Mol. Psychiatry* **23**, 1198–1204 (2018).
104. Stern, Y. Cognitive reserve and Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **20**, S69–S74 (2006).
105. White, Y. S., Bell, D. S. & Mellick, R. Sequelae to pneumoencephalography. *J. Neurol. Neurosurg. Psychiatry* **36**, 146–151 (1973).
106. Whittier, J. R. Deaths related to pneumoencephalography during a six year period. *AMA Arch. Neurol. Psychiatry* **65**, 463–471 (1951).
107. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–157 (2001).
108. Le Bihan, D. *et al.* Diffusion tensor imaging: concepts and applications. *J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn. Reson. Med.* **13**, 534–546 (2001).
109. Sporns, O., Tononi, G. & Kötter, R. The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* **1**, e42 (2005).
110. White, T. *et al*. Paediatric population neuroimaging and the Generation R Study: the second wave. *Eur. J. Epidemiol.* **33**, 99–125 (2018).
111. Ikram, M. A. *et al*. The Rotterdam Scan Study: design update 2016 and main findings. *Eur. J. Epidemiol.* **30**, 1299–1315 (2015).
112. Casey, B. J. *et al*. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
113. Miller, K. L. *et al*. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
114. Morris, Z. *et al*. Incidental findings on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* **339**, b3016 (2009).
115. Thompson, P. M. *et al*. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* **8**, 153–182 (2014).
116. Alfaro-Almagro, F. *et al*. UK Biobank Brain Imaging: Automated Processing Pipeline and Quality Control for 100,000 subjects. *Hum. Brain Mapping.* **1877** 400–424 (2016).
117. Jansen, A. G., Mous, S. E., White, T., Posthuma, D. & Polderman, T. J. C. What twin studies tell us about the heritability of brain development, morphology, and function: a review. *Neuropsychol. Rev.* **25**, 27–46 (2015).
118. Adams, H. H. H. *et al*. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19**, 1569–1582 (2016).
119. Savage, J. E. *et al*. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
120. Nagel, M. *et al*. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
121. Hibar, D. P. *et al*. Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224–229 (2015).
122. Schmaal, L. *et al*. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol. Psychiatry* **21**, 806–812 (2016).
123. Schmaal, L. *et al*. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol. Psychiatry* **22**, 900–909 (2017).
124. van Erp, T. G. M. *et al*. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* **21**, 547–553 (2016).
125. Hoogman, M. *et al*. Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in

children and adults: a cross-sectional mega-analysis. *The Lancet Psychiatry* **4**, 310–319 (2017).
126. Cooper, D., Barker, V., Radua, J., Fusar-Poli, P. & Lawrie, S. M. Multimodal voxel-based meta-analysis of structural and functional magnetic resonance imaging studies in those at elevated genetic risk of developing schizophrenia. *Psychiatry Res. Neuroimaging* **221**, 69–77 (2014).
127. McIntosh, A. M. *et al*. Longitudinal volume reductions in people at high genetic risk of schizophrenia as they develop psychosis. *Biol. Psychiatry* **69**, 953–958 (2011).
128. Lawrie, S. M. *et al*. Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. *Lancet* **353**, 30–33 (1999).
129. Dima, D. & Breen, G. Polygenic risk scores in imaging genetics: usefulness and applications. *J. Psychopharmacol*. **29**, 867–871 (2015).
130. Posthuma, D. *et al.* The association between brain volume and intelligence is of genetic origin. *Nat. Neurosci.* **5**, 83–84 (2002).

Part I:
Population-based Genetic Studies of Complex Traits

Chapter 2

Polygenic Scores for Schizophrenia and Educational Attainment are Associated with Behavioural Problems in Early Childhood in the General Population

Philip R. Jansen, Tinca J.C. Polderman, Koen Bolhuis, Jan van der Ende, Vincent W.V. Jaddoe, Frank C. Verhulst, Tonya White, Danielle Posthuma*, Henning Tiemeier*

* = *authors jointly supervised this work*

Published in *Journal of Child Psychology and Psychiatry* 2018 Jan;59(1):39-47

Background: Genome-wide association studies in adults have identified numerous genetic variants related to psychiatric disorders and related traits, such as schizophrenia and educational attainment. However, the effects of these genetic variants on behaviour in the general population remain to be fully understood, particularly in younger populations. We investigated whether polygenic scores of five psychiatric disorders and educational attainment are related to emotional and behaviour problems during early childhood.
Methods: From the Generation R Study, we included participants with available genotype data and behavioural problems measured with the Child Behavior Checklist (CBCL) at the age of 3 (n=1,902), 6 (n=2,202) and 10 years old (n=1,843). Polygenic scores were calculated for five psychiatric disorders and educational attainment. These polygenic scores were tested for an association with the broadband internalizing and externalizing problem scales and the specific CBCL syndrome scale scores.
Results: Analysis of the CBCL broadband scales showed that the schizophrenia polygenic score was associated with significantly higher internalizing scores at 3, 6 and 10 years and higher externalizing scores at age 3 and 6. The educational attainment polygenic score was associated with lower externalizing scores at all time points and lower internalizing scores at age 3. No associations were observed for the polygenic scores of bipolar disorder, major depressive disorder and autism spectrum disorder. Secondary analyses of specific syndrome scores showed that the schizophrenia polygenic score was strongly related to the Thought Problems scores. A negative association was observed between the educational attainment polygenic score and Attention Problems scores across all age groups.
Conclusions: Polygenic scores for adult psychitric disorders and educational attainment are associated with variation in emotional and behavioral problems already at a very early age.

Introduction

Childhood emotional and behavioural problems are common and show moderate stability throughout childhood[1]. Although symptoms may fluctuate over time, early childhood problems have predictive value for psychiatric disorders later in life, as well as academic performance[2] and risk-taking behaviour[3].

Variation in problem behaviour is influenced substantially by genetic factors. Twin study heritability estimates of behavioural problems are fairly constant over the course of childhood, and show that genetic factors explain around 50% of the variance in externalizing (e.g. aggression and oppositional behaviour) and internalizing (e.g. depression and anxiety) behaviours[4]. However, no statistically significant genetic polymorphism has yet been identified specifically for common childhood emotional and behavioural problems[5–7]. In contrast, GWAS studies in adult behaviour-related phenotypes have identified variants for a wide variety of traits and show that the majority of neuropsychiatric traits are genetically complex and determined by many genetic variants, mostly of small effect[8,9].

Educational attainment is an important predictor for a variety of important life outcomes and is closely linked to psychopathology[10]. Recently, proxy phenotype methods have targeted educational attainment (i.e. years of schooling) to detect genetic variants related to psychiatric and personality-related traits, including schizophrenia and neuroticism[11]. It was also shown that in school-aged children, a genetic predisposition to higher educational attainment is associated with higher cognitive performance[12].

Although many genetic variants have been identified for adult psychiatric disorders, it is unclear whether the same genetic variants are also associated with problem behaviour at an early age. Moreover, it is unclear whether children at high risk for psychopathology already show differences in behaviour during childhood. Population-based cohort studies in adolescents have demonstrated that genetic risks for psychopathology, quantified by risk scoring methods, correlate with a diverse array of behavioural outcomes[13]. More specifically, studies using schizophrenia polygenic risk scores reported that this genetic risk is associated with negative symptoms in the general adolescent population[14,15]. However, a recent study investigating the schizophrenia polygenic score in young children, suggests that manifestations in behaviour and neurodevelopment may be present much earlier in life[16]. In addition, most studies in children have tested polygenic risk scores for single traits, without performing comparisons across traits.

The aim of this study was to investigate whether polygenic risk scores of later life outcomes are associated with early childhood behavioural and emotional problems in the general population. We focus on five psychiatric disorders (schizophrenia, bipolar disorders, major depressive disorder, ADHD and autism spectrum disorder) and educational attainment, as these traits have been shown to be associated with early life problem behaviour[17,18]. We

hypothesize that genetic variants, associated with psychiatric disorders and educational attainment, are related to variation in symptoms in internalizing and externalizing domains at early childhood, thus earlier in life than described in most previous studies. This study will provide a better understanding of the association between risk variants and behavioural manifestations early in life and insight into the underlying neurobiology of these traits.

Methods

The Medical Ethics Committee of the Erasmus Medical Center approved all study procedures, and parents of the participants provided written informed consent.

This study was conducted within the Generation R cohort, a large population-based longitudinal cohort focused on child development[19]. Emotional and behavioural problems were assessed prospectively at the approximate age of 3, 6 and 10 years in respectively 4,612, 6,199 and 4,770 children. Of these children, 2,964, 3,926 and 3,058 were subsequently selected based on the availability of genotype data. Of these, 1,902, 2,202 and 1,843 children passed genotype quality control procedures and were included in the analyses.

Child behaviour measures

Behaviour problems were assessed with the Child Behavior Checklist (CBCL), a comprehensive list of items about various child emotional and behavioural problems, to be completed by the primary caregiver[20]. Each CBCL item can be scored as: 0 = 'not true', 1 = 'somewhat or sometimes true', 2 = 'very true or often true'. CBCL items can be scored on two broadband scales: 'Internalizing Problems' and 'Externalizing Problems', and on more specific syndrome scales. During the first and second assessment wave, the preschool CBCL version (CBCL/1½-5) was used, as most children during the assessment were younger than 6 years, and other versions are not appropriate for this age[21]. The CBCL/1½-5 survey consists of 100 problem items. At the third assessment wave, the school-age (CBCL/6-18) version was used, consisting of 120 problem items.

An overview of the CBCL syndrome scales and the number of items within each domain are shown in **Table S1**, available online.

Genotyping and imputation

Genotype calling procedures and subsequent processing for the Generation R Study have been described previously[22]. Briefly, genotype data were either collected from cord blood at birth (Illumina 610K Quad Chip) or via vena puncture (Illumina 660K Quad Chip) during a visit to the research centre.

Additional quality control steps were performed on the genotype data in PLINK[23]. Variants were filtered for minor allele frequency (MAF < 0.01), Hardy-Weinberg disequilibrium ($P < 0.00001$) and missing rate (> 0.05). Individuals from European descent were selected within 4 standard deviations on the first four genetic principal components of the HapMap Phase II Northwestern European (CEU) population. Individuals were additionally filtered on relatedness, sex mismatch and genotype quality (< 0.1).

P-value thresholds (pT) for inclusion of genetic variants in the score varied between pT < 0.01 and pT < 1. **Table S3** shows the number of SNPs that were included in the final polygenic score for each *P*-value threshold. Pearson correlations between the polygenic scores of the six traits are shown in **Figure S1**. Polygenic scores were standardized to a mean of 0 and standard deviation of 1 to increase interpretation of the score.

Statistical analyses

Statistical analyses were performed in R statistical software[24] (version 3.2.1).

Polygenic scores for the six traits were tested individually in a linear regression model for association with CBCL Internalizing and Externalizing Problems, corrected for age, sex and four genetic principal components. Next, we tested the most significant *P*-value threshold of each trait for associations with individual syndrome scales.

To account for varying degrees of skewness in CBCL scores, syndrome scores were transformed at each time point using Box–Cox transformation. This method utilizes maximum likelihood estimation (MLE) to find the optimal transformation parameter to approximate a normal distribution[25].

False-discovery rate (FDR) was applied to correct for multiple comparisons[26]. Based on the total number of statistical tests across polygenic scores, *P*-value thresholds, broadband scales and specific syndrome scales, a corrected *P*-value significance threshold was set to pFDR = 0.0083, and *P*-values below this corrected threshold were considered statistically significant.

Results

Sample Characteristics

Characteristics of the study sample at the three assessment waves are shown in **Table 1**. The three groups had a mean age of 3.0 (s.d. = 0.1), 6.0 (s.d. = 0.4) and 9.7 (s.d. = 0.3) years at the time of the assessment, and sex was equally divided among groups (per cent boys: age 3: 52%, age 6: 50%, age 10: 49%). At all ages, boys scored higher

than girls on externalizing problems (mean difference: age 3: 0.12, $P < 0.001$; age 6: 0.07, $P < .001$; age 10: 0.17, $P < 0.001$); there were no significant sex differences on the internalizing problem scales at age 3, 6 and 10.
The explained variance (R^2) of the broadband internalizing and externalizing problem scales by the six polygenic scores for all *P*-value thresholds is shown in **Figure 1**. **Table S4a–c** show the full regression results for these associations. Here we highlight the *P*-value threshold for each trait that showed the strongest association (i.e. largest increase in R^2) with the outcome across the different age groups

Schizophrenia
Analyses in 3-year olds showed that the SCZ polygenic score was significantly associated with higher levels of internalizing problems (pT < 0.5: β = 0.061, $P = 0.008$) and externalizing problems (pT < 0.5: β = 0.067, $P = 0.004$). At the age of 6 years, an association between the SCZ polygenic score and internalizing scores (pT < 0.5: β = 0.088, $P = < 0.5$: β= 0.070, $P < 0.001$) was also present. At age 10, we observed again the association with internalizing scores (pT < 0.5: β = 0.069, $P = 0.003$), whereas the association with externalizing scores was no longer significant (pT < 0.5: β = 0.039, $P = 0.096$).

ADHD
No significant associations were observed with externalizing or internalizing scores at the age of 3. However, we observed a weak positive association with externalizing scores at age 6 (pT < 0.01: β = 0.042, $P = 0.044$) that was not significant after correction for multiple comparisons. This association became stronger and significant at the age of 10 (pT < 0.01: β = 0.076, $P = 0.001$).

Bipolar disorder, major depressive disorder and autism spectrum disorder
No association was observed between the BP, MDD and ASD polygenic scores and externalizing and internalizing scales in any age group.

Educational attainment
The EA polygenic score was negatively associated with externalizing (pT < 0.5: β = –0.089, $P < 0.001$) and internalizing scores (pT < 0.5: β = –0.067, $P = 0.004$) in 3-year olds. Again at the age of 6 years, the EA polygenic score was associated with lower levels of externalizing problems (pT < 0.5: β = –0.067, $P = 0.001$), but not with internalizing problems. Similarly, EA polygenic scores were associated with lower externalizing scores at the age of 10 years (pT < 0.5: β = –0.051, $P = 0.027$), but the association at this age did not survive multiple comparison correction. Again at this age, no associations with internalizing scores were observed.

Sex interaction
In sensitivity analyses, we tested sex-specific associations of SCZ and EA polygenic scores and the externalizing and internalizing scores; the stratified results are shown in **Figure S2a–f**. Sex showed a weak interaction with the EA polygenic score on internalizing scores ($P = 0.026$) and externalizing scores ($P = 0.024$); however, these results

Table 1 | Sample characteristics.

	Assessment		
	Age 3 n = 1,902	**Age 6 n = 2,202**	**Age 10 n = 1,843**
Characteristics			
Mean age, years	3.04 ± 0.09	5.99 ± 0.37	9.69 ±0.27
Gender, % male	52%	51%	49%
Internalizing problems			
Mean score	4.22 ± 3.86	5.21 ± 5.35	4.57 ± 4.79
Range	0 - 36	0 - 49	0 - 41
Median	3	4	3
Externalizing problems			
Mean score	7.88 ± 5.97	6.87 ± 6.41	3.82 ± 4.73
Score range	0 - 42	0 - 43	0 - 39
Median	7	5	2

were not significant given the number of tests.

Syndrome scales

To test whether specific syndrome scales of the CBCL were driving the associations between polygenic scores and internalizing and externalizing scores, we performed secondary analyses testing associations between the polygenic scores of SCZ, ADHD and EA and the individual CBCL syndrome scales. Only the *P*-value thresholds that showed the strongest association for SCZ (pT < 0.5), ADHD (pT < 0.01) and EA (pT < 0.5) with the externalizing and internalizing scales in the primary analyses were tested in the secondary analyses. A visual representation of the regression coefficients is shown in **Figure 2a–c**, and an overview of the full regression results is available in **Table S5a–c**.

The SCZ polygenic score was mainly associated with higher Emotionally Reactive scores at age 3 (pT < 0.5, β = 0.086, *P* < 0.001). There was a negative association between the SCZ polygenic score and all internalizing subscales at the age 6, with the strongest association being with Withdrawn scores (pT < 0.5, β = 0.072, *P* < 0.001). Interestingly, at age 10, there was a strong positive association with Thought Problems scores of the school-aged CBCL version (pT < 0.5, β = 0.087, *P* < 0.001) (**Figure 2c**). As expected, the association between the ADHD polygenic score and externalizing scores at the age 6 was mainly driven by Attention Problems (pT < 0.01, β = 0.065, *P* = 0.002). At age 10, the ADHD polygenic score was mainly associated with higher levels of aggressive behavior (pT < 0.01, β = 0.083, *P* < 0.001). The previously observed association with Attention Problems was no longer significant (pT < 0.01, β = 0.045, *P* = 0.051). The EA polygenic score showed a strong negative association with Attention Problems scores in 3-year olds (pT < 0.5, β = –0.095, *P* < 0.001). This association was observed again at 6 (pT < 0.5, β = –0.082, *P* < 0.001) and at age 10 (pT < 0.5, β = –0.082, *P* < 0.001) in the school-aged CBCL, in which Attention Problems scores are not part of the externalizing domain.

Discussion

This study presents evidence that in very young children (age 3), a genetic predisposition for psychopathology is associated with more emotional and behavioural problems in the general paediatric population, whereas the polygenic score of EA to lower levels of problem behaviour.

The SCZ polygenic score showed an association with in-

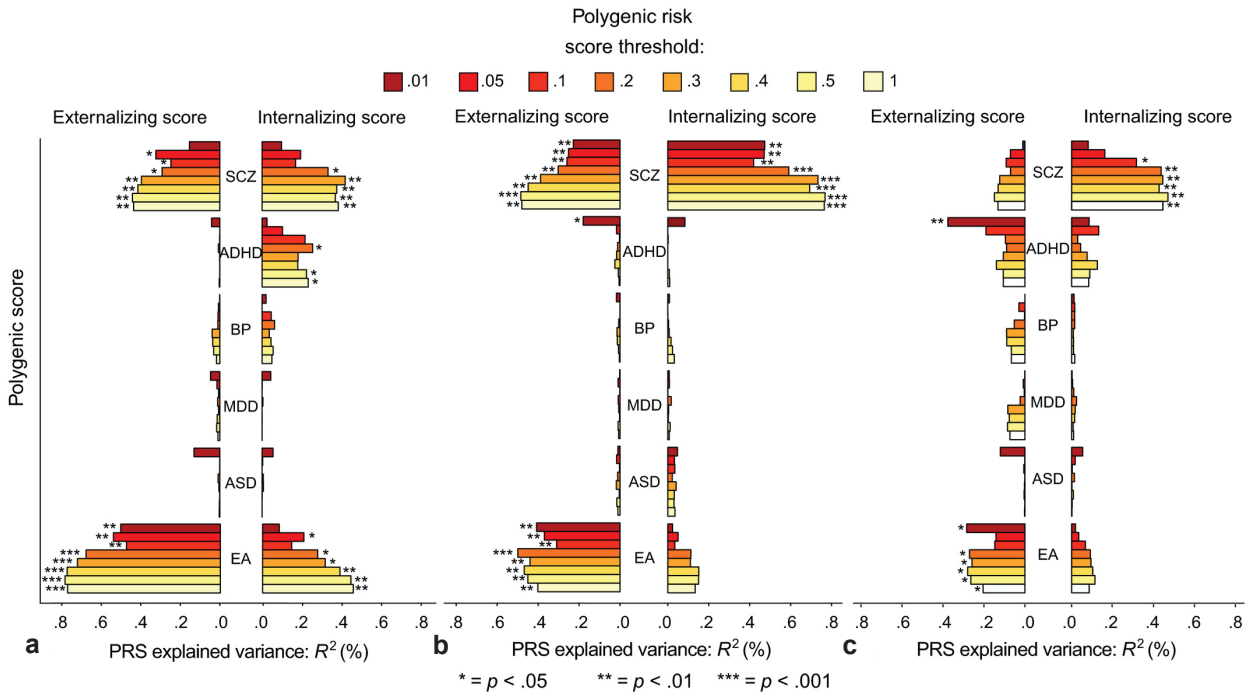


Fig. 1 | Explained variance (%) in externalizing and internalizing CBCL scores by polygenic scores at the age of 3 years (a) 6 years (b) and 10 years (c). Regression results are corrected for age, gender and four principal components. SCZ, schizophrenia; ADHD, attention-deficit hyperactivity disorder; BP, bipolar disorder; MDD, major depressive disorder; ASD, autism spectrum disorder; EA, educational attainment.

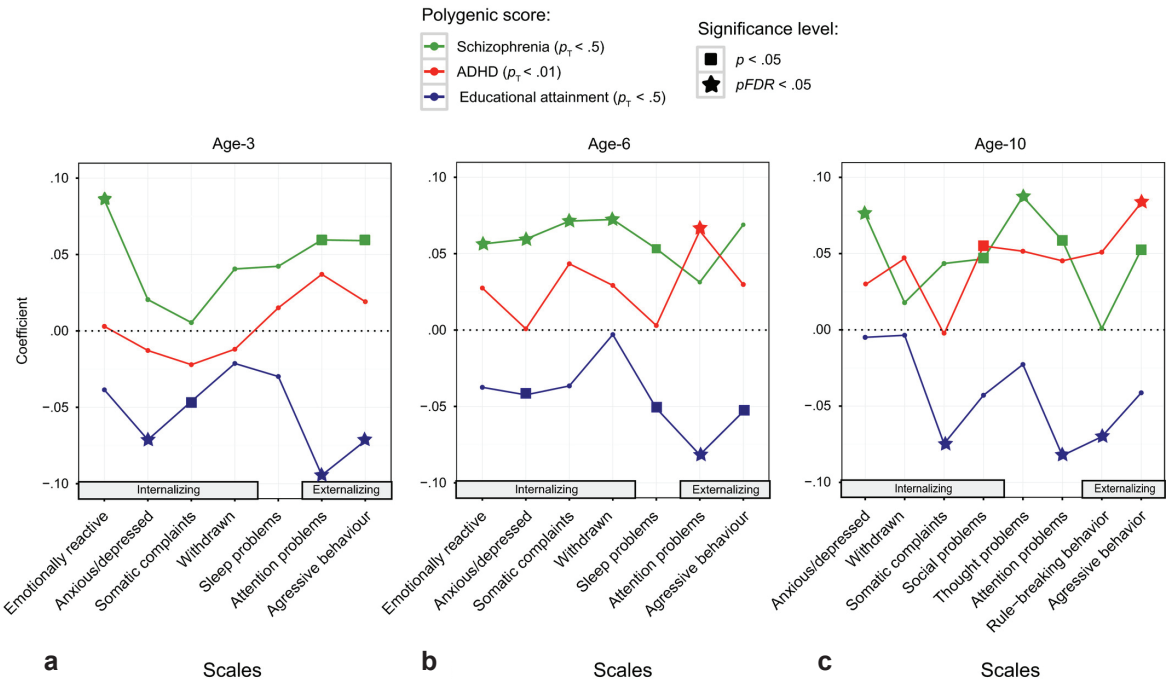


Fig. 2 | Visual representation of associations between polygenic scores and the specific syndrome scores in the age 3 (a), age 6 (b) and age 10 (c) assessment. Coefficients are standardized betas for the association between polygenic score on the individual syndrome scores, corrected for age, sex and four principal components.

ternalizing scores from age 3 onwards and was associated with CBCL Thought Problems scores at age 10. Prior research investigating the SCZ polygenic score and problem behaviour has been performed mainly in adolescents[14,15]. Our study shows that genetic predisposition for schizophrenia is associated with variation in behaviour already at an early age, and possibly as early as behaviour can be reliably assessed. This observation is in line with a recent study that reported associations between the SCZ polygenic score and lower cognitive ability, more social impairments and more behavioural problems in 4- to 9-year-old children[16]. Interestingly, this study showed associations with prosocial behaviour and conduct problems at age 4, but no associations with emotional problems were observed. Although we found similar associations with externalizing behaviours, our results add that the genetic predisposition for schizophrenia is associated with higher levels of emotional reactivity in 3-year olds. This discrepancy may be due to more specific items related to the emotional state of the child in the Emotional Reactivity scale of the CBCL.

Given that the incidence of schizophrenia peaks between adolescence and young adulthood, the prominent effect of schizophrenia polygenic scores on internalizing problems at age 6 compared to age 10 was surprising. This suggests that the associations between schizophrenia polygenic risk scores and behaviour in the general population is best captured by the internalizing syndrome scale within the CBCL/1½-5 internalizing domain (such as Emotionally Reactive scores) rather than the internalizing scales of the CBCL/6-18, which was used at the assessment at age 10. Differences in the subscale items (e.g. more items related to measures of affect regulation in the CBCL/1½-5 Withdrawn scale) may contribute to the different findings between CBCL versions.

The observed association between the SCZ polygenic score and the Thought Problems scale at age 10 is a remarkable new finding and contrasts with two earlier studies in healthy populations that did not find an association between the SCZ polygenic score and positive symptoms[14,15]. The Thought Problems scale, containing items such as 'sees things that aren't there' and 'strange ideas', reflects psychosis-like symptoms and similar behaviour that has previously been found to be a precursor of later life psychosis in prospective studies in similar age groups[27,28]. Where the two previous studies utilized psychotic symptoms as a binary outcome measure (presence/absence of psychotic experiences), our study aimed to measure psychotic symptoms along a continuum, possibly yielding more power to detect the subtle effects of the polygenic

score on psychotic-like experiences in the general population.

We expected to observe similar associations for schizophrenia and bipolar disorder polygenic scores, as the genetic overlap between psychiatric traits is substantial[23,29]. However, these differences are possibly a consequence of differences in sample size of the GWAS for schizophrenia (n = 77,096) compared to bipolar disorder (n = 16,731) and illustrate the importance of a well-powered discovery GWAS for polygenic risk scoring[30]. The fact that no associations were observed for ASD and MDD polygenic scores could also be due to the lack of a well-powered GWAS study. However, despite the smallest sample size of the ADHD discovery GWAS (**Table S2**), we observed evidence for significant associations with Attention Problems and Aggresive Behaviour for one *P*-value threshold. This finding may be explained by a higher prevalence of ADHD and ADHD-related symptoms in the age range of our study. Moreover, the high specificity of specific CBCL scales for measuring ADHD-related symptoms (such as the Attention Problems scale in the Externalizing Problems scale) may further lead to these observed associations. Previous studies showed higher ADHD polygenic scores in children with comorbid aggression[31] and attention problems in the general population[32]. Interestingly, we observed that the ADHD polygenic score was related to attention problems in 3- and 6-year olds, but that this association shifted towards more aggression problems at the age of 10 years. While it is possible that these differences are related to the two CBCL versions used, an alternative possibility is that polygenic scores are not necessarily related to a fixed set of symptoms, but related to dynamics and changes during childhood. The association with both the attention and aggression scales could result from a shared genetic aetiology between ADHD symptoms and oppositional defiant disorder (ODD)-related symptoms, which has been described previously in twin research[33]. Furthermore, distinction between these behavioural scales can be challenging for the parents and may result in classifying attention problems as aggressive behaviour and vice versa.

In our study, genetic variants for educational attainment were negatively associated with externalizing symptoms, which suggests that less externalizing problems early in life may be beneficial for achieving a higher level of education. Indeed, externalizing problems are an important determinant of poor academic performance and have been shown to precede school problems[2]. Our findings imply a shared genetic aetiology for this association and suggest this is partly explained by higher levels of attention problems. This has also been suggested by prospective studies that showed lower academic achievement in children with symptoms of hyperactivity and inattentiveness[34] and impaired cognitive functioning in individuals with ADHD[35]. In addition, prior research reported that cognitive ability and behavioural scores are highly intertwined[36] and suggest that the association between educational attainment polygenic scores and lower externalizing problems could result from better cognitive abilities in these children. Besides a direct association between genetic predisposition and childhood behaviour, the observed associations could partially be explained by the parental polygenic scores: Children with high polygenic scores of psychopathology are more likely to have parents with higher than average polygenic scores. Parental polygenic scores could subsequently lead to differences in environmental factors of the child (e.g. passive gene–environment correlation, including parenting strategies). Studies on this complex interplay suggest that environmental factors such as parenting moderate the associations between polygenic scores and child behaviour[37]. Moreover, studies including genetic data of the mother suggest that the maternal ADHD polygenic score moderates the association between the ADHD polygenic score of the child and educational achievement[38]. Applying polygenic scoring in a family-based setting could provide more insight into the dynamic interaction between the genetic profile of the child and the parents, family upbringing such as parenting, and child behaviour.

Our study suggest that the genetic risk for schizophrenia manifests as more internalizing problems, and to lesser degree as more externalizing problems, in children 3 years of age. Recent developmental studies of schizophrenia onset have focused on early puberty and the occurrence of psychotic experiences such as acoustic hallucinations[39]. However, based on the current results, we carefully speculate that nonspecific symptoms of emotional reactivity and anxiety may further help to tailor prevention programmes for high-risk children, e.g. as defined by family history.

The strength of the study is that behaviour was assessed at multiple time points, providing information about behaviour during different stages of development. Given that childhood behaviour is dynamic, longitudinal studies are important to study the association between genetic predisposition and behaviour at different ages. Our results illustrate this by showing that associations with behaviour problems were found at specific ages that were not present at an earlier age or disappeared at an older age. Future genetic studies should aim to assess behaviour at multiple time points and study whether changes in behaviour are related to the genetic predisposition of the child.

A limitation of this study is that the analyses were restricted to observations from the primary caregiver. Integration of information from different observers could provide more complete information about the child's behaviour. However, given the young age of children in our study, we expect scores reported by the primary caregiver to be the most accurate reflection of the children's behaviour. In addition, due to low power of the discovery GWAS study, the observed associations for the ADHD polygenic score were not as robust as those found for schizophrenia and educational attainment. This is illustrated by the observation that associations were only found for the most stringent *P*-value threshold ($pT < 0.01$), but lacked broader support from other *P*-value thresholds. In contrast, the associations of schizophrenia and educational attainment showed stronger consistency across multiple *P*-value thresholds.

Conclusion

In conclusion, this study shows that genetic predispositions for psychiatric disorders and educational attainment are associated with early behavioural problems. These associations were present throughout early childhood and at an earlier age than described in most previous studies. Children with a high genetic predisposition for psychiatric traits show specific early manifestations of problem behaviour at a young age, which may further aid the early recognition of precursors of psychopathology in high-risk individuals.

References

1. Basten, M. *et al.* The stability of problem behavior across the preschool years: an empirical approach in the general population. *J. Abnorm. Child Psychol.* **44,** 393–404 (2016).
2. Van der Ende, J., Verhulst, F. C. & Tiemeier, H. The bidirectional pathways between internalizing and externalizing problems and academic performance from 6 to 18 years. *Dev. Psychopathol.* **28,** 855–867 (2016).
3. King, S. M., Iacono, W. G. & McGue, M. Childhood externalizing and internalizing psychopathology in the prediction of early substance use. *Addiction* **99,** 1548–1559 (2004).
4. Verhulst, F. C. & Boomsma, D. I. Genetic and environmental contributions to stability and change in children's internalizing and externalizing problems. *J. Am. Acad. Child Adolesc. Psychiatry* **42,** 1212–1220 (2003).
5. Benke, K. S. *et al.* A genome-wide association meta-analysis of preschool internalizing problems. *J. Am. Acad. Child Adolesc. Psychiatry* **53,** 667–676 (2014).
6. Middeldorp, C. M. *et al.* A genome-wide association meta-analysis of attention-deficit/hyperactivity disorder symptoms in population-based pediatric cohorts. *J. Am. Acad. Child Adolesc. Psychiatry* **55,** 896–905 (2016).
7. Pappa, I. *et al.* A genome-wide approach to children's aggressive behavior: The EAGLE consortium. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **171,** 562–572 (2016).
8. Plomin, R., DeFries, J. C., Knopik, V. S. & Neiderhiser, J. M. Top 10 replicated findings from behavioral genetics. *Perspect. Psychol. Sci.* **11,** 3–23 (2016).
9. Sullivan, P. F., Daly, M. J. & O'donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13,** 537–551 (2012).
10. Kessler, R. C., Foster, C. L., Saunders, W. B. & Stang, P. E. Social consequences of psychiatric disorders, I: Educational attainment. *Am. J. Psychiatry* **152,** 1026–1032 (1995).
11. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624–633 (2016).
12. Ward, M. E. *et al.* Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children. *PLoS One* **9,** e100248 (2014).
13. Krapohl, E. *et al.* Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* **21,** 1188–1193 (2016).
14. Derks, E. M. *et al.* Investigation of the genetic association between quantitative measures of psychosis and schizophrenia: a polygenic risk score analysis. *PLoS One* **7,** e37852 (2012).
15. Jones, H. J. *et al.* Phenotypic manifestation of genetic risk for schizophrenia during adolescence in the general population. *JAMA psychiatry* **73,** 221–228 (2016).
16. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The Lancet Psychiatry* **4,** 57–62 (2017).
17. Breslau, J. *et al.* The impact of early behavior disturbances on academic achievement in high school. *Pediatrics* **123,** 1472–1476 (2009).
18. Caspi, A., Moffitt, T. E., Newman, D. L. & Silva, P. A. Behavioral observations at age 3 years predict adult psychiatric disorders: Longitudinal evidence from a birth cohort. *Arch. Gen. Psychiatry* **53,** 1033–1039 (1996).
19. Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31,** 1243–1264 (2016).
20. Achenbach, T. M. & Rescorla, L. A. *Manual for the ASEBA preschool forms and profiles.* **30,** (Burlington, VT: University of Vermont, Research center for children, youth, 2000).
21. Tiemeier, H. *et al.* The Generation R Study: a review of design, findings to date, and a study of the 5-HTTLPR by environmental interaction from fetal life onward. *J. Am. Acad. Child Adolesc. Psychiatry* **51,** 1119–1135 (2012).
22. Medina-Gomez, C. *et al.* Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.* **30,** 317–330 (2015).
23. 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).
24. R Core Team, R: A language and environment for statistical computing. (2013).
25. Sakia, R. M. The Box-Cox transformation technique: a review. *J. R. Stat. Soc. Ser. D. (The Statistician)* **41,** 169–178 (1992).
26. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 289–300 (1995).
27. Poulton, R. *et al.* Children's self-reported psychotic symptoms and adult schizophreniform disorder: a 15-year longitudinal study. *Arch. Gen. Psychiatry* **57,** 1053–1058 (2000).
28. Welham, J. *et al.* Emotional and behavioural antecedents of young adults who screen positive for non-affective psychosis: a 21-year birth cohort study. *Psychol. Med.* **39,** 625–634 (2009).
29. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).

30. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (2013).
31. Hamshere, M. L. *et al.* High loading of polygenic risk for ADHD in children with comorbid aggression. *Am. J. Psychiatry* **170,** 909–916 (2013).
32. Martin, J., Hamshere, M. L., Stergiakouli, E., O'Donovan, M. C. & Thapar, A. Genetic risk for attention-deficit/hyperactivity disorder contributes to neurodevelopmental traits in the general population. *Biol. Psychiatry* **76,** 664–671 (2014).
33. Tuvblad, C., Zheng, M., Raine, A. & Baker, L. A. A common genetic factor explains the covariation among ADHD ODD and CD symptoms in 9–10 year old boys and girls. *J. Abnorm. Child Psychol.* **37,** 153–167 (2009).
34. Polderman, T. J. C., Boomsma, D. I., Bartels, M., Verhulst, F. C. & Huizink, A. C. A systematic review of prospective studies on attention problems and academic achievement. *Acta Psychiatr. Scand.* **122,** 271–284 (2010).
35. Faraone, S. V, Ghirardi, L., Kuja-Halkola, R., Lichtenstein, P. & Larsson, H. The familial co-aggregation of attention-deficit/hyperactivity disorder and intellectual disability: a register-based family study. *J. Am. Acad. Child Adolesc. Psychiatry* **56,** 167–174 (2017).
36. Blanken, L. M. E. *et al.* Cognitive functioning in children with internalising, externalising and dysregulation problems: a population-based study. *Eur. Child Adolesc. Psychiatry* **26,** 445–456 (2017).
37. Salvatore, J. E. *et al.* Polygenic risk for externalizing disorders: Gene-by-development and gene-by-environment effects in adolescents and young adults. *Clin. Psychol. Sci.* **3,** 189–201 (2015).
38. Stergiakouli, E. *et al.* Association between polygenic risk scores for attention-deficit hyperactivity disorder and educational and cognitive outcomes in the general population. *Int. J. Epidemiol.* **46,** 421–428 (2016).
39. Zammit, S. *et al.* Psychotic experiences and psychotic disorders at age 18 in relation to psychotic experiences at age 12 in a longitudinal population-based cohort study. *Am. J. Psychiatry* **170,** 742–750 (2013).

Supplementary information

Supplementary figures (1-2) and tables (1-5) can be found in the online version of the manuscript:

https://onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.12759

Chapter 3

Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways

Philip R. Jansen, Kyoko Watanabe, Sven Stringer, Nathan Skene, Julien Bryois, Anke R. Hammerschlag, Christiaan A. de Leeuw, Jeroen S. Benjamins, Ana B. Muñoz-Manchado, Mats Nagel, Jeanne E. Savage, Henning Tiemeier, Tonya White, The 23andMe Research Team, Joyce Y. Tung, David A. Hinds, Vladimir Vacic, Xin Wang, Patrick F. Sullivan, Sophie van der Sluis, Tinca J.C. Polderman, August B. Smit, Jens Hjerling-Leffler, Eus J.W. Van Someren*, Danielle Posthuma*

* = *authors jointly supervised*

Published in *Nature Genetics* 2019 Mar;51(3):394-403

Insomnia is the second most prevalent mental disorder[1], with no sufficient treatment available. Despite substantial heritability, insight into the associated genes and neurobiological pathways remains limited. Here, we use a large genetic association sample (n = 1,331,010) to detect novel loci and gain insight into the pathways, tissue and cell types involved in insomnia complaints. We identify 202 loci implicating 956 genes through positional, expression quantitative trait loci, and chromatin mapping. The meta-analysis explained 2.6% of the variance. We show gene set enrichments for the axonal part of neurons, cortical and subcortical tissues, and specific cell types, including striatal, hypothalamic, and claustrum neurons. We found considerable genetic correlations with psychiatric traits and sleep duration, and modest correlations with other sleep-related traits. Mendelian randomization identified the causal effects of insomnia on depression, diabetes, and cardiovascular disease, and the protective effects of educational attainment and intracranial volume. Our findings highlight key brain areas and cell types implicated in insomnia, and provide new treatment targets.

Insomnia is the second most prevalent mental disorder. One-third of the general population reports insomnia complaints. The diagnostic criteria for insomnia disorder[2] (that is, difficulties with initiating or maintaining sleep with accompanying daytime complaints at least three times a week for at least three months, which cannot be attributed to inadequate circumstances for sleep[3]) are met by 10% of individuals, and up to one-third of older age individuals[4]. Insomnia contributes significantly to the risk and severity of cardiovascular, metabolic, mood, and neurodegenerative disorders[2] .

Despite evidence of a considerable genetic component (heritability 38–59%[5]), only a small number of genetic loci moderating the risk of insomnia have been identified thus far. Recent genome-wide association studies (GWAS)[6,7] for insomnia complaints ($n = 113{,}006$) demonstrated its polygenic architecture and implicated three genome-wide significant (GWS) loci and seven genes. A prominent role was reported for *MEIS1*, which is associated with insomnia complaints[6,7] and restless legs syndrome (RLS)[8] through pleiotropy and phenotypic overlap; yet, the role of other genes was not unambiguously shown.

We set out to substantially increase the sample size to allow the detection of more genetic risk variants for insomnia complaints, which may aid in understanding its neurobiological mechanisms. By combining data collected in the UK Biobank (UKB) version 2[9] ($n = 386{,}533$) and 23andMe, a privately held personal genomics and biotechnology company[10,11] ($n = 944{,}477$), we obtained an unprecedented sample size of 1,331,010 individuals. Insomnia complaints were measured using questionnaire data; an independent sample (the Netherlands Sleep Register)[12] which gives access to similar question data, as well as clinical interviews assessing insomnia disorder (see **Supplementary Note**), was used to validate the specific questions so that they were good proxies of insomnia disorder.

We found 202 risk loci for insomnia; extensive functional in silico analyses showed the involvement of specific tissue and cell types. Mendelian randomization identified causal effects of insomnia on metabolic and psychiatric traits.

Results

Meta-analysis yields 202 risk loci

The UKB assessed insomnia complaints (hereafter referred to as 'insomnia') with a touchscreen device, whereas 23andMe research participants completed online surveys (**Supplementary Tables 1** and **2**). The assessment of insomnia in both samples shows high accuracy for insomnia disorder in the UKB and somewhat lower accuracy in 23andMe (sensitivity/specificity: UKB = 98/96%; 23andMe = 84/80%) (see **Supplementary Note**). The prevalence of insomnia was 28.3% in the UKB version 2 sample, 30.5% in the 23andMe sample, and 29.9% in the combined sample, which is in keeping with previous estimates for people of advanced age in the UK[4]and elsewhere[13,14]. Older people dominate the UKB (mean age = 56.7, s.d. = 8.0) and 23andMe (two-thirds of the sample older than 45, one-third older than 60 years of age) samples. Prevalence was higher in females (34.6%) than males (24.5%), yielding an odds ratio (OR) of 1.6, which is close to the 1.4 OR reported in a meta-analysis[15]. Quality control was conducted separately per sample, following standardized, stringent protocols (see **Methods).** The GWAS was run separately per sample (UKB: $n = 386{,}533$; 23andMe: $n = 944{,}477$) (**Supplementary Fig. 1**), and then meta-analyzed with METAL[16] by weighing the single nucleotide polymorphism (SNP) effect by sample size (see **Methods**). We first analyzed males and females separately (**Supplementary Fig. 2**) and observed a high genetic correlation between the sexes ($r_g = 0.92$, s.e.m. = 0.02; **Supplementary Table 3**), indicating strong overlap of genetic effects. Owing to the large sample size, the r_g of 0.92 was significantly different from 1 (one-sided Wald test, $P = 2.54 \times 10^{-6}$), suggesting a small role for sex-specific genetic risk factors, consistent with

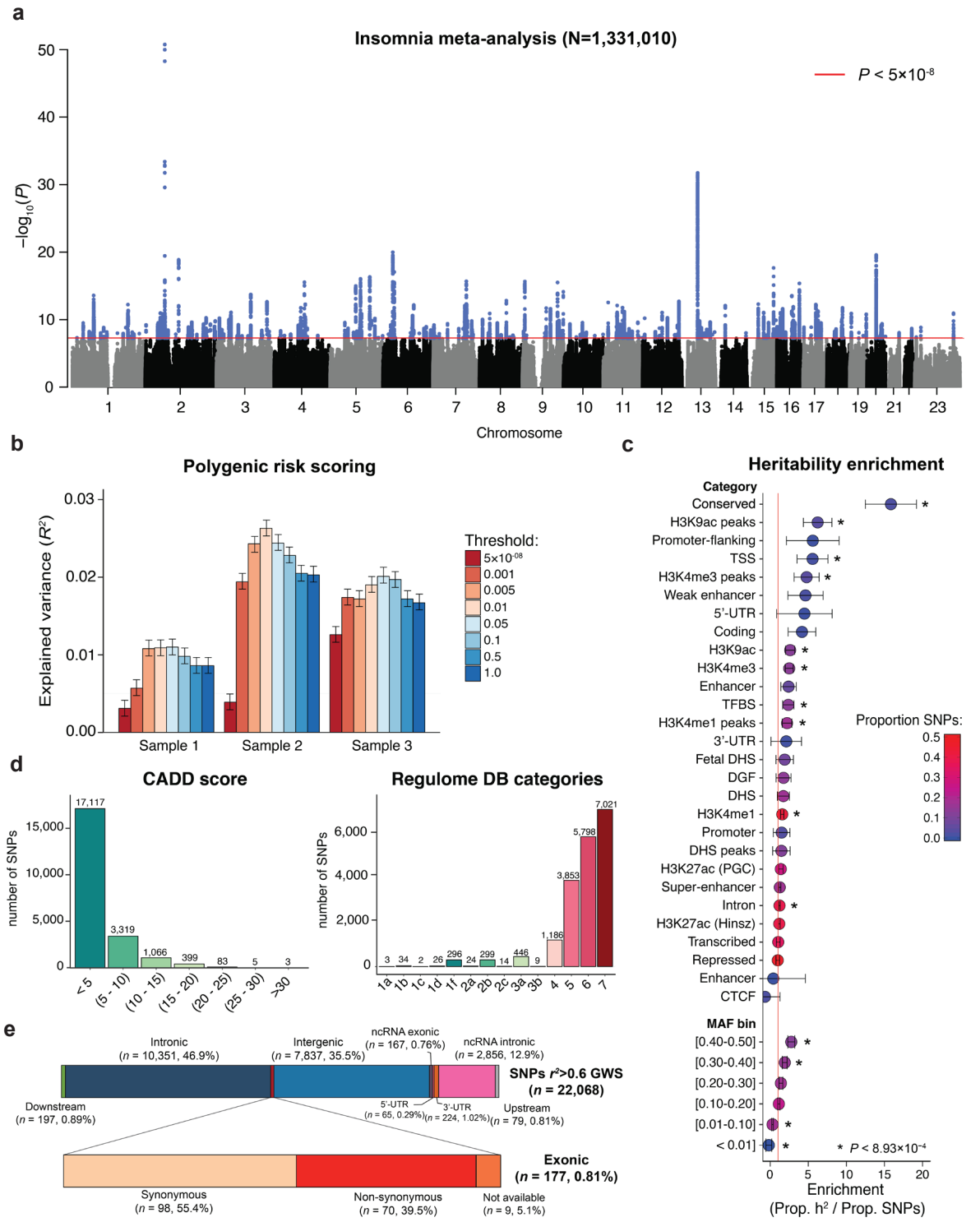


Fig. 1 | SNP-based results from the GWAS meta-analysis on insomnia in 1,331,010 individuals. a, Manhattan plot of the GWAS meta-analysis of insomnia in the UKB and 23andMe cohorts, showing the $-\log_{10}$ transformed *P*-value for each SNP. SNP two-sided *P*-values from a linear model were calculated using METAL, weighting SNP associations by sample size. **b**, PGS prediction in three hold-out samples ($n = 3{,}000$), showing the increase in explained variance in insomnia (Nagelkerke's pseudo R^2)

(continued from previous page) in a logistic regression model and 95% confidence intervals for each *P*-value threshold. All *P*-value thresholds were statistically significant. **c**, Heritability enrichment for functional SNP categories and MAF bins. Enrichment was calculated by dividing the proportion of heritability for each category by the proportion of SNPs in that category. The error bars show the 95% confidence interval around the estimate. Significant enrichments after Bonferroni correction (28 functional categories + 6 MAF bins + 22 chromosomes) are indicated by an asterisk ($P < 0.05/56$ categories $= 8.93 \times 10^{-4}$). TFBS, transcription factor binding site; DHS, DNase I hypersensitive site; DGF, digital genomic footprint; PGC, Psychiatric Genomics Consortium; Hnisz, as reported in Hnisz et al.; CTCF, CCCTC-binding factor. **d**, Distribution of CADD scores and RegulomeDB categories of all annotated SNPs in linkage disequilibrium ($r^2 \geq 0.6$) with one of the GWS SNPs ($n = 22{,}068$). **e**, Functional consequences of these annotated SNPs**.**

our previous study[6] However, since sex-specific effects were relatively small, we focused on identifying genetic effects important in both sexes and continued with the combined sample. (**Supplementary Tables 4** and **5** and the **Supplementary Note** provide sex-specific results.) The genetic correlation of insomnia between the full UKB and 23andMe results was $r_g = 0.69$ (s.e.m. = 0.02).

We observed a significant polygenic signal in the GWAS (lambda inflation factor = 1.808), which could not be ascribed to spurious association (linkage disequilibrium score intercept = 1.075)[17] (**Supplementary Fig. 3a**). Meta-analysis identified 11,990 GWS SNPs ($P < 5 \times 10^{-8}$), represented by 248 independent lead SNPs ($r^2 < 0.1$), located in 202 genomic risk loci (**Fig. 1a, Supplementary Data Set 1**, and **Supplementary Tables 6** and **7**). All lead SNPs showed concordant signs of effect in both samples (**Supplementary Fig. 3b**). We confirmed two (chr2:66,785,180 and chr5:135,393,752) out of six previously reported loci for insomnia[6,7] (**Supplementary Table 8**).

Polygenic score (PGS) prediction in three randomly selected hold-out samples ($n = 3 \times 3{,}000$) estimated the current results to explain up to 2.6% of the variance in insomnia (**Fig. 1b**, **Supplementary Fig. 4**, and **Supplementary Table 9**).

The SNP-based heritability (h^2_{SNP}) was estimated at 7.0% (s.e.m. = 0.002). Partitioning the heritability by functional categories of SNPs (see **Methods**) showed the strongest enrichment of h^2_{SNP} in conserved regions (enrichment = 15.8, $P = 1.57 \times 10^{-14}$). In addition, h^2_{SNP} was enriched in common SNPs (minor allele frequency (MAF) > 0.3) and depleted in rarer SNPs (MAF < 0.01; **Fig. 1c** and **Supplementary Table 10**).

We used FUMA[18] to functionally annotate all SNPs in the risk loci that were in linkage disequilibrium ($r^2 \geq 0.6$) with one of the independent significant SNPs (see **Methods**). The majority of the 22,068 annotated SNPs (76.8%) were in open chromatin regions[19]as indicated by a minimum chromatin state of 1–7 (**Fig. 1d** and **Supplementary Table 11**). In line with findings for other traits[6,20] about half of these SNPs were in intergenic (35.5%) or non-coding RNA (13.0%) regions (**Fig. 1e**); of these, 0.72% were highly likely to have a regulatory function as indicated by a RegulomeDB score < 2 (see **Methods**). However, of these, 51.5% were located inside a protein-coding gene and 0.81% were exonic. Of the 177 exonic SNPs, 71 were exonic non-synonymous (**Supplementary Table 12** and **Supplementary Note**). *WDR90* included four exonic non-synonymous SNPs (rs7190775, rs4984906, rs3752493, and rs3803697) all in high linkage disequilibrium with the same independent significant SNP (rs3184470). There were two exonic non-synonymous SNPs with extremely high combined annotation-dependent depletion (CADD) scores[21], suggesting a strong deleterious effect on protein function: rs13107325 in *SLC39A8* (locus 56, $P = 8.31 \times 10^{-16}$) with the derived allele T (MAF = 0.03), associated with an increased risk of insomnia; and rs35713889 in *LAMB2* (locus 42, $P = 1.77 \times 10^{-7}$), where the derived allele T of rs35713889 (MAF = 0.11) was also associated with an increased risk of insomnia complaints. **Supplementary Table 13** provides a detailed overview of the functional impact of all variants in the genomic risk loci.

Genes implicated in insomnia

To obtain an insight into the (functional) consequences of individual GWS SNPs, we used FUMA[18] to apply three strategies to map associated variants to genes (see Methods). Positional gene mapping aligned SNPs to 412 genes by location. Expression quantitative trait loci (eQTL) gene mapping matched cis-eQTL SNPs to 594 genes whose expression levels they influence. Chromatin interaction mapping annotated SNPs to 159 genes based on three-dimensional DNA–DNA interactions between genomic regions of the GWS SNPs and nearby or distant genes (**Supplementary Data Set 2**, **Supplementary Table 14**, and **Supplementary Note**). Ninety-two genes were mapped by all three strategies (**Supplementary Table 15**), and 336 genes were physically located outside the risk loci but were implicated by eQTL associations (306 genes), chromatin interactions (16 genes), or both (14 genes). Several genes were implicated by GWS SNPs originating from two distinct risk loci on the same chromosome (**Fig. 2a,b**): *MEIS1*, located on chromosome 2 in

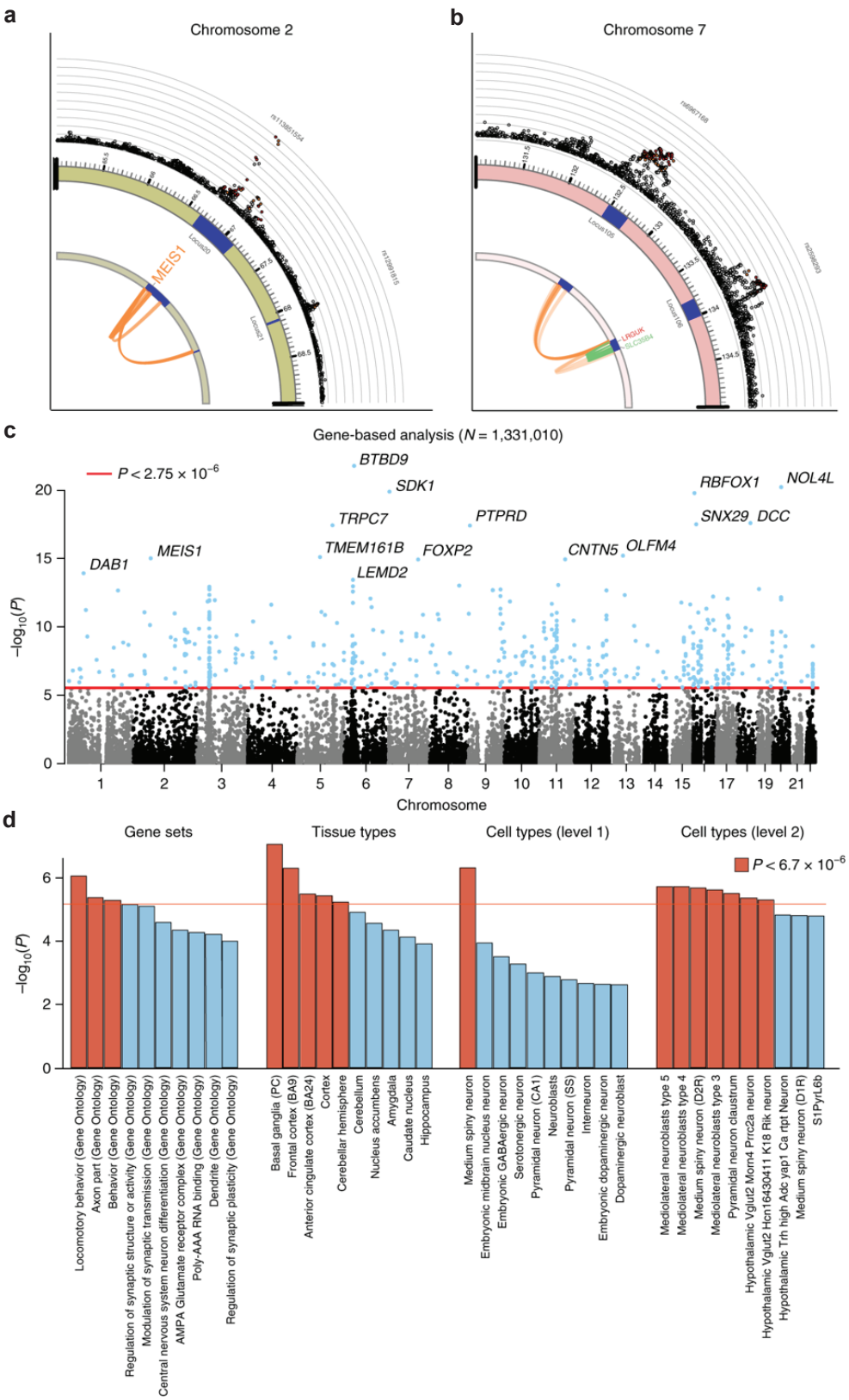

a
Chromosome 2
MEIS1
b
Chromosome 7
c
Gene-based analysis (N = 1,331,010)
P < 2.75 × 10−6
BTBD9
SDK1
TRPC7
PTPRD
TMEM161B
FOXP2
LEMD2
DAB1
MEIS1
CNTN5
OLFM4
RBFOX1
NOL4L
SNX29
DCC
−log10(P)
Chromosome
d
Gene sets
Tissue types
Cell types (level 1)
Cell types (level 2)
P < 6.7 × 10−6
Locomotory behavior (Gene Ontology)
Axon part (Gene Ontology)
Behavior (Gene Ontology)
Regulation of synaptic structure or activity (Gene Ontology)
Modulation of synaptic transmission (Gene Ontology)
Central nervous system neuron differentiation (Gene Ontology)
AMPA Glutamate receptor complex (Gene Ontology)
Poly-AAA RNA binding (Gene Ontology)
Dendrite (Gene Ontology)
Regulation of synaptic plasticity (Gene Ontology)
Basal ganglia (PC)
Frontal cortex (BA9)
Anterior cingulate cortex (BA24)
Cortex
Cerebellar hemisphere
Cerebellum
Nucleus accumbens
Amygdala
Caudate nucleus
Hippocampus
Medium spiny neuron
Embryonic midbrain nucleus neuron
Embryonic GABAergic neuron
Serotonergic neuron
Pyramidal neuron (CA1)
Neuroblasts
Pyramidal neuron (SS)
Interneuron
Embryonic dopaminergic neuron
Dopaminergic neuroblast
Mediolateral neuroblasts type 5
Mediolateral neuroblasts type 4
Medium spiny neuron (D2R)
Mediolateral neuroblasts type 3
Pyramidal neuron claustrum
Hypothalamic Vglut2 Morn4 Prrc2a neuron
Hypothalamic Vglut2 Hcn16430411 K18 Rik neuron
Hypothalamic Trh high Adc yap1 Ca rtpt Neuron
Medium spiny neuron (D1R)
S1PyrL6b

Fig. 2 (previous page) | Gene-based and gene set analyses of insomnia in 1,331,010 individuals. a,b, Zoomed-in circos plots showing the genes implicated by two genomic risk loci on chromosome 2 (**a**) and chromosome 7 (**b**), with the genomic risk loci indicated as blue areas, eQTL associations in green, and chromatin interactions in orange. Genes mapped by both eQTL and chromatin interactions are red. The outer layer shows a Manhattan plot containing the $-\log_{10}$ transformed *P*-value of each SNP in the GWAS meta-analysis of insomnia in the UKB and 23andMe cohorts. Full circos plots of all autosomal chromosomes are provided in **Supplementary Data Set 2. c**, Genome-wide gene-based analysis (GWAS) of 18,185 genes that were tested for association with insomnia in MAGMA. The y axis shows the $-\log_{10}$ transformed two-sided *P*-value of the gene-based test, and the x axis shows the starting position on the chromosome. Gene-based two-sided *P*-values were calculated with MAGMA. The red line indicates the Bonferroni-corrected threshold for genome-wide significance ($P = 0.05/18{,}185$ genes $= 2.75 \times 10^{-6}$). The top 15 most significant genes are highlighted. **d**, Gene set analysis of the top 10 for each of the MSigDB pathways, tissue expression of GTEx tissue types, and cell types from single-cell RNA sequencing. Gene set analyses were performed with MAGMA. The red line shows the Bonferroni significance threshold ($P < 0.05/7{,}473$ gene sets $= 6.7 \times 10^{-6}$), correcting for the total number of tested gene sets. The red bars indicate the significant gene sets.

the strongest associated locus (locus 20), was positionally mapped by 51 SNPs and mapped by chromatin interactions in 10 tissue types, including cross-loci interactions from locus 21, and is a known gene involved in insomnia[6]; and *LRGUK*, located on chromosome 7 in locus 106, was positionally mapped by 22 SNPs and chromatin interactions in 3 tissue types, including cross-loci interactions from locus 105. *LRGUK* was also implicated by eQTL associations of 125 SNPs in 14 general tissue types. *LRGUK* was previously implicated in type 2 diabetes[22] and autism spectrum disorder[23] (disorders with prominent insomnia). However, it is not yet directly implicated in sleep-related phenotype, and is the most likely candidate to explain the observed association at loci 105 and 106.

Apart from linking individually associated genetic variants to genes, we conducted a genome-wide gene association analysis (GWGAS) using MAGMA[24]. GWGAS provides aggregate association *P*-values based on all variants located in a gene, and complements the three FUMA mapping strategies (see **Methods**). GWGAS identified 517 associated genes (**Fig. 2c** and **Supplementary Table 16**). The top gene *BTBD9* ($P = 8.51 \times 10^{-23}$) on chromosome 6 in locus 81 was also mapped using positional and eQTL mapping (tissue type: left ventricle of the heart), and is part of a pathway that regulates circadian rhythms. *BTBD9* has been associated with RLS, periodic limb movement disorder[25,26] , and Tourette syndrome[27]. Involvement in sleep regulation was shown in *Drosophila*[28]; mouse mutants show fragmented sleep[29] and increased levels of dynamin 1[30], a protein that mediates the increased sleep onset latency that follows presleep arousal[31].

Of the 517 MAGMA-based associated genes, 222 were outside of the GWAS risk loci, and 309 were also mapped by FUMA. In total, 956 unique genes were mapped by at least one of the three FUMA gene mapping strategies or by MAGMA (**Supplementary Fig. 5**). Of these, *MEIS1*, *MED27*, *IPO7*, and *ACBD4* confirmed previous results[6,7] (**Supplementary Table 17**). Sixty-two genes were implicated by all four mapping strategies, indicating that, apart from a GWS gene-based *P*-value, there were: (1) GWS SNPs located in proximity of or inside these genes; (2) GWS SNPs associated with differential expression of these genes; and (3) GWS SNPs involved in genomic regions interacting with these genes. We note that genes that were indicated by positional mapping and GWS gene-based *P*-values, but not via eQTL or chromatin interaction mapping ($n = 54$ genes), may be of equal importance; yet, they are more likely to exert their influence on insomnia via structural changes in gene products (that is, at the protein level) and not via quantitative changes in the availability of gene products.

Implicated pathways, tissues, and cell types

To test whether GWS genes converged in functional gene sets and pathways, we conducted gene-set analyses using MAGMA (see **Methods**). We tested the associations of 7,473 gene sets: 7,246 sets derived from the MSigDB-[32]gene expression values from 54 tissues from the GTEx database[33]; and cell-specific gene expression in 173 types of brain cells (**Fig. 2d** and **Supplementary Table 18**). Competitive testing was used and a Bonferroni-corrected threshold of $P < 6.7 \times 10^{-6}$ (0.05/7,473) to correct for multiple testing. Of the MSigDB gene sets, three Gene Ontology gene sets survived multiple testing: Gene Ontology:locomotory behavior ($P = 8.95 \times 10^{-7}$); Gene Ontology:behavior ($P = 5.23 \times 10^{-6}$); and Gene Ontology:axon part ($P = 4.25 \times 10^{-6}$). This set includes 16 GWS genes: *KIF3B, SNCA, GRIA1, CDH8, ROBO2, DNM1, RANGAP1, GABBR1, P2RX3, NRG1, POLG, DAG, NFASC,* and *CALB2*).

Tissue specific gene-set analyses showed strong enrichment of genetic signal in genes expressed in the brain. Correcting for overall expression, four specific brain tissues reached the threshold for significance: the overall cerebral cortex ($P = 3.68 \times 10^{-6}$); Brodmann area 9 of the frontal cortex ($P = 5.04 \times 10^{-7}$); BA24 of the anteri-

or cingulate cortex ($P = 3.25 \times 10^{-6}$); and the cerebellar hemisphere ($P = 5.93 \times 10^{-6}$). Several other brain tissues also showed strong enrichment just below the threshold, including three striatal basal ganglia structures (nucleus accumbens, caudate nucleus, putamen). To test whether genes expressed in all three basal ganglia structures together would show significant enrichment of low *P*-values, we used the first principal component (BG_{PC}) of these basal ganglia structures (**Methods**) and found significant enrichment ($P = 8.33 \times 10^{-8}$).

When conditioning the three top cortical structures on the BG_{PC}, they were no longer significantly associated after multiple testing correction (minimum $P = 0.03$), which was expected given that the BG_{PC} correlated strongly with gene expression in cortical (and other) areas ($r^2 > 0.96$). Similar results were obtained vice versa; that is, using the first principal component of all cortical areas and conditioning the three basal ganglia structures on this resulted in no evidence of enrichment of low *P*-values for basal ganglia structures (minimum $P = 0.53$). These results show that (1) genes expressed in the brain are important in insomnia, (2) genes expressed in cortical areas are more strongly associated than genes expressed in basal ganglia, and (3) there is a strong correlation between gene expression patterns across brain tissues, which suggests involvement of general cellular signatures rather than specific brain tissue structures.

Brain cell type-specific gene-set analyses were first carried out on 24 broad, cell-type categories. Cell type-specific gene expression was quantified using single-cell RNA sequencing of disassociated cells from the somatosensory cortex, hippocampus, hypothalamus, striatum, and midbrain from the mouse (see **Methods**), which closely resembles gene expression in humans[39]. Results indicat-

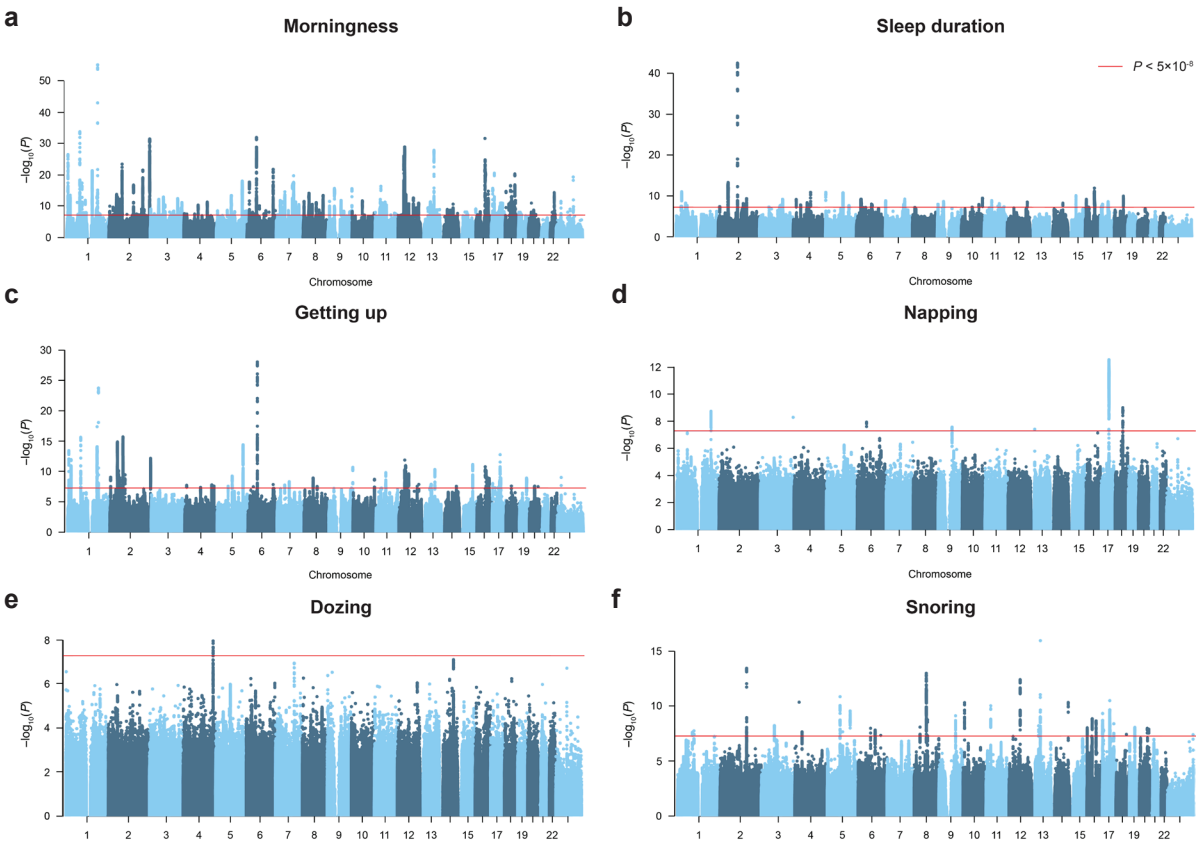


Fig. 3 | Genome-wide analyses of six sleep-related traits. a–f, Manhattan plots of the genome-wide analyses of **(a)** morningness (UKB and 23andMe cohorts, $n = 434{,}835$), **(b)** sleep duration (UKB, $n = 384{,}317$), **(c)** ease of getting up (UKB, $n = 385{,}949$), **(d)** napping (UKB, $n = 386{,}577$), **(e)** daytime dozing (UKB, $n = 386{,}548$), and **(f)** snoring (UKB, $n = 359{,}916$). The y axis shows the $-\log_{10}$ transformed SNP two-sided *P*-value from a linear or logistic regression model, and the x axis the base-pair position of the SNPs on each chromosome. The red line indicates the Bonferroni-corrected significance threshold ($P < 5 \times 10^{-8}$).

ed that genes expressed specifically in the medium spiny neurons (MSNs, $P = 4.83 \times 10^{-7}$) were associated with insomnia; no other broad, cell type-specific gene set survived our strict threshold of $P < 6.7 \times 10^{-6}$. MSNs represent 95% of neurons within the human striatum, which is one of the four major nuclei of the subcortical basal ganglia. Specifically, the striatum consists of the ventral (nucleus accumbens and olfactory tubercle) and dorsal (caudate nucleus and putamen) subdivisions. The association with MSNs thus likely explains the observed association of the basal ganglia striatal structures (nucleus accumbens, caudate nucleus, putamen).

Using broad cell classes risks not detecting associations that are specific to distinctive yet rare cell types. To account for this, we then tested 149 specific brain cell-type categories and found significant associations with 7 specific cell types: mediolateral neuroblasts type 3, 4, and 5 ($P = 2.36 \times 10^{-6}$, $P = 1.88 \times 10^{-6}$, and $P = 1.87 \times 10^{-6}$, respectively); D2-type MSNs ($P = 2.12 \times 10^{-6}$); claustrum pyramidal neurons ($P = 3.09 \times 10^{-6}$); hypothalamic Vglut2 Morn4 Prrc2a neurons ($P = 4.36 \times 10^{-6}$); and hypothalamic Vglut2 Hcn16430411 K18 Rik neurons ($P = 4.98 \times 10^{-6}$). The hypothalamus contains multiple nuclei that are key to the control of sleep and arousal, including the suprachiasmatic nucleus, which accommodates the biological clock of the brain[40]. These results suggest a role of distinct mature and developing cell types in the midbrain and hypothalamus.

Modest genetic overlap with sleep traits

Other sleep-related traits may easily be confounded with specific symptoms of insomnia, like early morning awakening, and difficulties maintaining sleep. The most recent genome-wide studies for other sleep-related traits included 59,128–128,266 individuals and assessed genetic effects on morningness[41–43] (that is, being a morning person), sleep duration[7,43], and daytime sleepiness/dozing[7]. Using increased sample sizes for each of these sleep-related traits (maximum $n = 434{,}835$), we investigated to what extent insomnia and other sleep-related traits are genetically distinct or overlapping. We performed GWAS analyses for the following six sleep-related traits: morningness; sleep duration; ease of getting up in the morning; taking naps during the day; daytime dozing; and snoring (**Supplementary Note** and **Supplementary Figs. 6** and **7**). Of the 202 risk loci for insomnia, 39 were also GWS in at least one of the other sleep-related traits (**Fig. 3** and **Supplementary Table 20**). The strongest overlap in loci was found with sleep duration; 14 out of 49 sleep duration loci overlapped with insomnia. Insomnia showed the highest genetic correlation with sleep duration (−0.47, s.e.m. = 0.02; **Supplementary Table 21**) compared to other sleep-related traits; this was not surprising given that insomnia also shared the largest number of risk loci with sleep duration (see further discussion of results for sleep phenotypes in the **Supplementary Note**)

Gene mapping of SNP associations of sleep-related traits

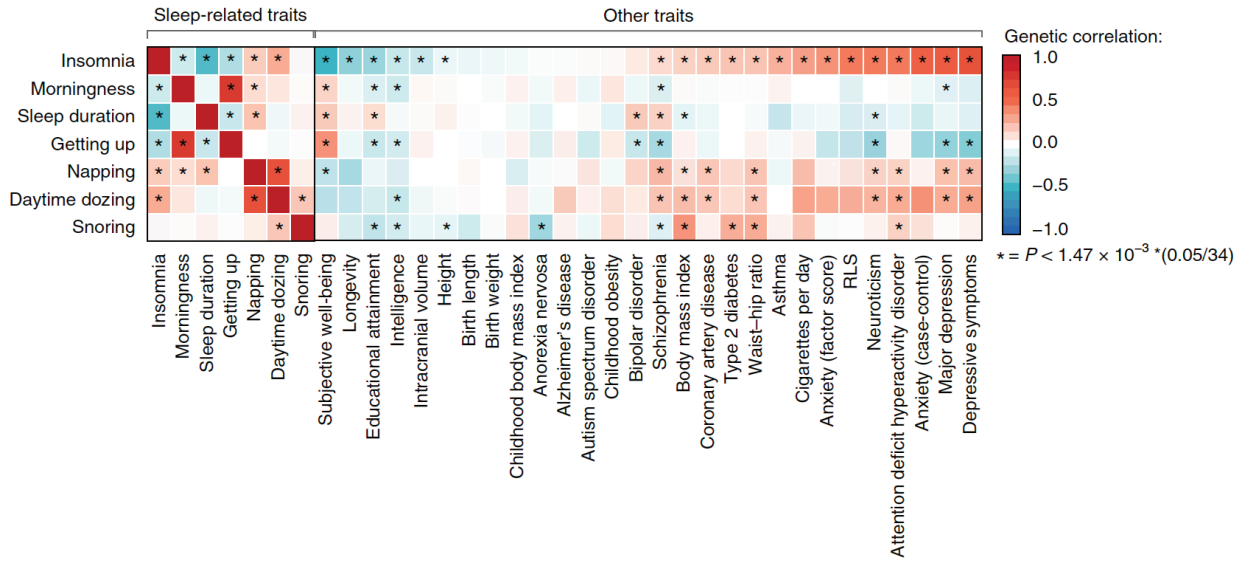


Fig. 4 | Genetic overlap of insomnia with other sleep-related traits and psychiatric and metabolic traits. Heatmap of genetic correlations between the insomnia GWAS meta-analysis, sleep-related phenotypes, and neuropsychiatric and metabolic traits studies. Genetic correlations and two-sided *P*-values were calculated using linkage disequilibrium score regression. Red indicates a positive r_g, whereas green indicates a negative r_g. Correlations that were significant after Bonferroni correction ($P < 0.05/34$ traits $= 1.47 \times 10^{-3}$) are indicated with an asterisk (see also **Supplementary Tables 21 and 29**).

resulted in 973 unique genes (**Supplementary Fig. 8** and **Supplementary Tables 22–26**). Gene-based analysis showed that, of the 517 GWS genes for insomnia, 120 were GWS in at least one of the other sleep-related traits, and one gene (*RBFOX1*) was GWS in all traits except napping and daytime dozing (**Supplementary Table 27**). The largest proportion of overlap in GWS genes for insomnia was again with sleep duration, with 37 of the 134 (27%) GWS genes for sleep duration being GWS for insomnia also.

There was overlap in tissue enrichment in cortical structures and basal ganglia between insomnia and both morningness and sleep duration. At the single-cell level, MSNs were also implicated for morningness and sleep duration, but not for the other sleep-related traits (**Supplementary Table 28**). Taken together, these results suggest that, at a genetic level, insomnia shows considerable genetic overlap with sleep duration, and modest overlap with other sleep-related traits.

Strong overlap between insomnia and psychiatric traits

We confirm previously reported genetic correlations between insomnia and neuropsychiatric and metabolic traits, including type 2 diabetes, waist–hip ratio, and body mass index[6,41] (**Supplementary Table 29**), and also identify several GWS SNPs for insomnia that have previously been associated with these traits (**Supplementary Table 30**). The strongest correlations were with depressive symptoms ($r_g = 0.64$, s.e.m. = 0.04, $P = 1.21 \times 10^{-71}$), followed by anxiety disorder ($r_g = 0.56$, s.e.m. = 0.11, $P = 1.40 \times 10^{-7}$), subjective well-being ($r_g = -0.51$, s.e.m. = 0.03, $P = 4.93 \times 10^{-52}$), major depression ($r_g = 0.50$, s.e.m. = 0.07, $P = 8.08 \times 10^{-12}$), and neuroticism ($r_g = 0.48$, s.e.m. = 0.02, $P = 8.72 \times 10^{-80}$). Genetic correlations with metabolic traits ranged between 0.09 and 0.20.

Notably, we observed a positive correlation with RLS ($r_g = 0.44$, s.e.m. = 0.07, $P = 4.36 \times 10^{-10}$), a trait that shares phenotypic characteristics with insomnia[6]. This suggests a partial genetic overlap, which we discuss in more detail in the **Supplementary Note** and **Supplementary Tables 31** and **32**. In this study, we show that although insomnia lead SNPs are enriched in RLS, there is only a partial genome-wide overlap between insomnia and RLS, in line with previous analyses[6]. The genetic correlations between insomnia and anxiety and depression-related traits (anxiety, neuroticism, major depression, and depressive symptoms) were also stronger than the correlations between insomnia and the other sleep-related traits (Mann–Whitney *U*-test *Z* score = −2.56, $P = 0.01$). Since a similar high reliability has been reported for both sleep and psychiatric phenotypes, the findings suggest that genetically insomnia more closely resembles neuropsychiatric traits than other sleep-related traits (**Fig. 4**). These genetic correlations were consistent within the two meta-analyzed samples separately (Pearson's $r^2 = 0.98$; **Supplementary Fig. 9**).

To infer directional associations between insomnia and these correlated traits, we performed bidirectional multi-SNP Mendelian randomization analysis[44] (see **Methods**). The results support a direct risk effect of insomnia on metabolic syndrome phenotypes including body mass index ($b_{xy} = 0.36$, s.e.m. = 0.05, $P = 1.25 \times 10^{-12}$), type 2 diabetes ($b_{xy} = 0.62$, s.e.m. = 0.11, $P = 2.29 \times 10^{-8}$), and coronary artery disease ($b_{xy} = 0.61$, s.e.m. = 0.09, $P = 2.88 \times 10^{-12}$). We also found risk effects of insomnia on several psychiatric traits, including major depression ($b_{xy} = 1.57$, s.e.m. = 0.07, $P = 1.73 \times 10^{-111}$), schizophrenia ($b_{xy} = 0.68$, s.e.m. = 0.10, $P = 5.12 \times 10^{-11}$), attention-deficit hyperactivity disorder ($b_{xy} = 0.77$, s.e.m. = 0.06, $P = 2.50 \times 10^{-45}$), neuroticism ($b_{xy} = 0.45$, s.e.m. = 0.02, $P = 3.56 \times 10^{-92}$), and anxiety disorder ($b_{xy} = 0.47$, s.e.m. = 0.10, $P = 4.11 \times 10^{-6}$), with evidence of a reverse risk effect of major depression ($b_{xy} = 0.06$, s.e.m. = 0.003, $P = 6.93 \times 10^{-99}$) and neuroticism ($b_{xy} = 0.24$, s.e.m. = 0.01, $P = 7.90 \times 10^{-157}$) on insomnia. In addition, insomnia was bidirectionally associated with educational attainment ($b_{xy} = -0.32$, s.e.m. = 0.02, $P = 4.12 \times 10^{-45}$) and vice versa ($b_{xy} = -0.10$, s.e.m. = 0.01, $P = 2.27 \times 10^{-23}$); the same bidirectional pattern was observed for intelligence. Unidirectional protective effects were only observed for height ($b_{xy} = -0.03$, s.e.m. = 0.02, $P = 1.68 \times 10^{-77}$) and intracranial volume ($b_{xy} = -0.03$, s.e.m. = 0.01, $P = 3.72 \times 10^{-16}$).

Using GWAS results from RLS in the 23andMe cohort, we observed patterns of bidirectional effects of insomnia on RLS ($b_{xy} = 0.35$, s.e.m. = 0.05, $P = 2.53 \times 10^{-12}$) and vice versa ($b_{xy} = 0.12$, s.e.m. = 0.01, $P = 1.21 \times 10^{-35}$). Overall, only a small proportion of SNPs showed pleiotropy between insomnia and other traits (**Supplementary Table 33** and **Supplementary Note**).

Discussion

In the largest GWAS study to date of 1,331,010 participants, we identified 202 genomic risk loci for insomnia. Using extensive functional annotation of associated genetic variants, we demonstrated that the genetic component of insomnia points toward a role of genes enriched in locomotory behavior, and enriched in specific cell types from the claustrum, hypothalamus, and striatum, specifically in MSNs (**Fig. 5**).

MSNs are γ-aminobutyric acid (GABA)ergic inhibitory cells and represent 95% of neurons in the human striatum, one of the four major nuclei of the basal gan-

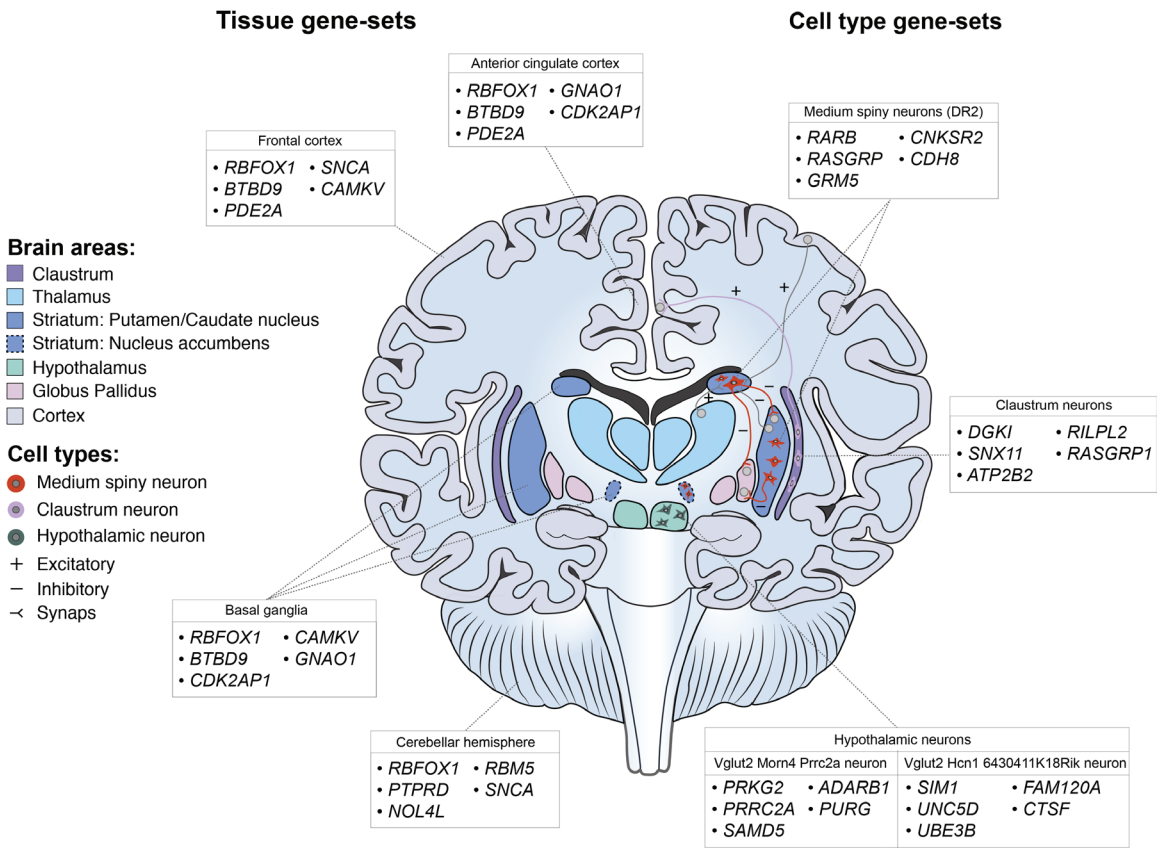


Fig. 5 | Overview of brain tissues and cell types associated with insomnia based on the GWAS results in 1,331,010 individuals. For each associated gene set, the top five genes driving the association are reported for each brain area and cell type. The results for the GTEx brain tissue type gene expression are shown on the left side, whereas the results from the level 2 single-cell gene expression are shown on the right.

glia (for reviews, see Vetrivelan et al.[45], Lazarus et al.[46], and Swardfager et al.[47]). MSNs were the first neurons in which the up and down states characteristic of slow-wave sleep were described[48]. Cell body-specific striatal lesions of the rostral striatum induce profound sleep fragmentation, which is highly characteristic of insomnia[45,49]. As discussed more extensively in the **Supplementary Note,** fragmented REM sleep is highly characteristic of insomnia and related to the ongoing thought-like mental content that makes patients with insomnia underestimate sleep duration[50–52]. Consistently short objective sleep across nights occurs only in a minority of patients with insomnia[53].

A role for the basal ganglia in sleep regulation is also suggested by the high prevalence of insomnia in neurodegenerative disorders, such as Parkinson's disease and Huntington's disease, where the basal ganglia are affected. Vetrivelan et al.[45] proposed a cortex-striatum-globus pallidus$_{external}$-cortex network involved in the control of sleep–wake behavior and cortical activation, where midbrain dopamine disinhibits the globus pallidus$_{external}$ and promotes sleep through the activation of D2 receptors in this network. Furthermore, brain imaging studies have suggested that the caudate nucleus of the striatum is a key node in the neuronal network imbalance of insomnia[54]; they also reported abnormal function in the cortical areas we found to be most enriched (BA9[55], BA24[56]). Our results support the involvement of the striato-cortical network in insomnia, by showing enrichment of risk genes for insomnia in cortical areas as well as the striatum, and specifically in MSNs. We recently showed that, along with several other cell types, MSNs mediate the risk for mood disorders[57]and schizophrenia[39]. MSNs are strongly implicated in reward processing; future work should address whether the genetic overlap between insomnia and mood disorders is medi-

ated by gene function in MSNs.

Our results also showed enrichment of insomnia genes in the pyramidal neurons of the claustrum. This subcortical brain region is structurally closely associated with the amygdala and has been implicated in salience coding of incoming stimuli and binding of multisensory information into conscious percepts[58]. These functions are highly relevant to insomnia because the disorder is characterized by increased processing of incoming stimuli[59]. Claustrum activity during REM sleep is moreover key to activation of the anterior cingulate cortex that was also enriched for insomnia gene expression[60]. We found enrichment of insomnia genes in mediolateral neuroblasts from the embryonic midbrain and in two hypothalamic cell types. The role of the mediolateral neuroblasts is less clear; although they were obtained from the embryonic midbrain, at present it is unknown what type of mature neurons they differentiate into. We note that the midbrain is similar on a bulk transcriptomic level to the pons[61], and lacking cells from that region we cannot conclusively say that midbrain cell types are enriched.

The current findings provide an insight into the causal mechanism of insomnia, showing enrichment in specific cell types, brain areas, and biological functions. These findings are starting points for the development of new therapeutic targets for insomnia and may also provide valuable insights into other genetically related disorders.

Online Methods

Meta-analysis

A meta-analysis of the GWAS results of insomnia and morningness in the UKB and 23andMe cohorts was performed using fixed-effects meta-analysis METAL[16], using SNP *P*-values weighted by sample size. To investigate sex-specific genetic effects, we ran the meta-analysis between the UKB and 23andMe datasets for males and females separately

Genomic risk loci definition

We used FUMA[18] version 1.2.4, an online platform for functional mapping and annotation of genetic variants, to define genomic risk loci and obtain functional information of the relevant SNPs in these loci. FUMA provides comprehensive annotation information by combining several external data sources. We first identified independent significant SNPs that had a GWS *P*-value ($< 5 \times 10^{-8}$) and were independent from each other at $r^2 < 0.6$. These SNPs were further represented by lead SNPs, a subset of the independent significant SNPs that were in approximate linkage equilibrium with each other at $r^2 < 0.1$. We then defined independent genomic risk loci by identifying physical regions in linkage disequilibrium with these lead SNPs that were > 250 kilobases (kb) apart from each other. The borders of the genomic risk loci were defined by identifying all SNPs in linkage disequilibrium ($r^2 \geq 0.6$) with one of the independent significant SNPs in the locus; the region containing all these candidate SNPs was considered to be a single independent genomic risk locus. Linkage disequilibrium information was calculated using the UKB genotype data as a reference. Risk loci were defined based on evidence from independent significant SNPs available in both 23andMe and UKB datasets.

SNPs that were GWS but only available in the 23andMe dataset were not included when defining genomic risk loci and were not included in any follow-up annotations or analyses because there was no external replication in the UKB sample. If such SNPs were located in a risk locus, they are displayed in LocusZoom plots (gray, as there is no linkage disequilibrium information in the UKB). When risk loci contained GWS SNPs based solely on the 23andMe dataset, we did not count that risk locus because there were no other SNPs available in both samples that supported these GWS SNPs.

Gene-based analysis

SNP-based *P*-values from the meta-analysis were used as input for the GWGAS; 18,182–18,185 protein-coding genes (each containing at least one SNP in the GWAS, the total number of tested genes can thus be slightly different across phenotypes) from the NCBI 37.3 gene definitions were used as the basis for the GWGAS in MAGMA[24]. Bonferroni correction was applied to correct for multiple testing ($P < 2.75 \times 10^{-6}$).

Gene-set analysis

Results from the GWGAS analyses were used to test for association in three types of 7,473 predefined gene sets:

1. 7,246 curated gene sets representing known biological and metabolic pathways derived from 9 data resources, cataloged by and obtained from the MSigDB version 6.0 (ref.[62]).
2. Gene expression values from 53 tissues obtained from GTEx[63] $\log_2$-transformed with pseudocount 1 after winsorization at 50 and averaged per tissue (+1 combined gene expression in the basal ganglia by taking the first principal component from principal component analysis of gene expression in three basal ganglia structures). We caution that only a limited set of brain tissues were included; thus, we cannot rule out associations with many important areas such as

the pons, midbrain, or thalamus based on this analysis.

3. Cell type-specific expression in 173 types of brain cells (24 broad categories of cell types 'level 1', and 149 specific categories of cell types 'level 2'), which were calculated following the method described by Skene et al.[39]. Briefly, brain cell-type expression data was drawn from single-cell RNA sequencing data from mouse brains. For each gene, the value for each cell type was calculated by dividing the mean unique molecular identifier counts for the given cell type by the summed mean unique molecular identifier counts across all cell types. Single-cell gene sets were derived by grouping genes into 40 equal bins based on specificity of expression. Mouse cell gene expression was shown to closely approximate gene expression in postmortem human tissue[39]

These gene sets were tested using MAGMA. MAGMA uses a continuous measure of association (gene-based *P*-value) of all genes that could be mapped by at least one SNP in the gene-based test and can perform gene-set analysis based on dichotomous gene sets (genes present in a gene set or not) or continuous values of gene expression in tissues and cells. We computed competitive *P*-values, which represent the test of association for a specific gene set compared with genes not in the gene set to correct for the baseline level of genetic association in the data[64]. The Bonferroni-corrected significance threshold was $P = 0.05/7{,}473$ gene sets $= 6.7 \times 10^{-6}$. Conditional analyses were performed as a follow-up using MAGMA to test whether each significant association observed was independent of all others. The association between each gene set in each of the three categories was tested conditional on the most strongly associated set, and then, if any substantial ($P < 0.05$/number of gene sets) associations remained, by conditioning on the first and second most strongly associated set, and so on until no associations remained. Gene sets that retained their association after correcting for other sets were considered to represent independent signals. We note that this is not a test of association per se, but rather a strategy to identify, among gene sets with known significant associations and overlap in genes, which set(s) are responsible for driving the observed association.

SNP-based heritability and genetic correlation

Linkage disequilibrium score regression[17] was used to estimate genomic inflation and SNP-based heritability of the phenotypes, and to estimate the cross-cohort genetic correlations. Precalculated linkage disequilibrium scores from the 1000 Genomes European reference population were obtained from https://data.broadinstitute.org/alkes-group/LDSCORE/.

Genetic correlation

Genetic correlations between sleep-related traits, and between sleep-related traits and previously published GWAS studies of sufficient sample size were calculated using linkage disequilibrium score regression on HapMap 3 SNPs only. Genetic correlations were corrected for multiple testing based on the total number of correlations (between 6 sleep-related phenotypes and 28 previous GWAS studies) by applying a Bonferroni-corrected threshold of $P < 0.05/34 = 1.47 \times 10^{-3}$.

Stratified heritability

To test whether specific categories of SNP annotations were enriched for heritability, we partitioned SNP heritability for binary annotations using stratified linkage disequilibrium score regression[65]. Heritability enrichment was calculated as the proportion of heritability explained by an SNP category divided by the proportion of SNPs that are in that category. Partitioned heritability was computed by 28 functional annotation categories, by MAF in six percentile bins, and by 22 chromosomes. Annotations for binary categories of functional genomic characteristics (for example, coding or regulatory regions) were obtained from the LD Score website. The Bonferroni-corrected significance threshold for 56 annotations was set at $P < 0.05/56 = 8.93 \times 10^{-4}$.

Functional annotation

Functional annotation of SNPs implicated in the meta-analysis was performed using FUMA[18]. We selected all candidate SNPs in genomic risk loci having an $r^2 \geq 0.6$ with one of the independent significant SNPs (see above), a *P*-value ($P < 1 \times 10^{-5}$), a MAF > 0.0001 for annotations, and availability in both UKB and 23andMe datasets. The functional consequences for these SNPs were obtained by matching each SNP's chromosome location, base-pair position, reference, and alternate alleles to databases containing known functional annotations, including ANNOVAR[66] categories, CADD scores, RegulomeDB[21]scores, and chromatin state[67]. ANNOVAR categories identify the SNP's genic position (for example, intron, exon, intergenic) and associated function. CADD scores predict how deleterious the effect of an SNP is likely to be for a protein structure/function, with higher scores representing higher deleteriousness. A CADD score > 12.37 is potentially pathogenic. The RegulomeDB score is a categorical score based on information from eQTLs and chromatin marks,

which ranges from 1a to 7 with lower scores indicating an increased likelihood of having a regulatory function. Scores are as follows: 1a = eQTL + transcription factor binding + matched transcription factor motif + matched DNase footprint + DNase peak; 1b = eQTL + transcription factor binding + any motif + DNase footprint + DNase peak; 1c = eQTL + transcription factor binding + matched transcription factor motif + DNase peak; 1d = eQTL + transcription factor binding + any motif + DNase peak; 1e = eQTL + transcription factor binding + matched transcription factor motif; 1f = eQTL + transcription factor binding/DNase peak; 2a = transcription factor binding + matched transcription factor motif + matched DNase footprint + DNase peak; 2b = transcription factor binding + any motif + DNase footprint + DNase peak; 2c = transcription factor binding + matched transcription factor motif + DNase peak; 3a = transcription factor binding + any motif + DNase peak; 3b = transcription factor binding + matched transcription factor motif; 4 = transcription factor binding + DNase peak; 5 = transcription factor binding or DNase peak; 6 = other; 7 = not available. The chromatin state represents the accessibility of genomic regions (every 200 base pairs (bp)) with 15 categorical states predicted by a hidden Markov model based on 5 chromatin marks for 127 epigenomes in the Roadmap Epigenomics Project[68]. A lower state indicates higher accessibility, with states 1–7 referring to open chromatin states. We annotated the minimum chromatin state across tissues to SNPs. The 15 core chromatin states as suggested by the Roadmap Epigenomics Project are as follows: 1 = active transcription start site (TSS); 2 = flanking active TSS; 3 = transcription at gene 5' and 3'; 4 = strong transcription; 5 = weak transcription; 6 = genic enhancers; 7 = enhancers; 8 = zinc finger genes and repeats; 9 = heterochromatic; 10 = bivalent/poised TSS; 11 = flanking bivalent/poised TSS/enhancer; 12 = bivalent enhancer; 13 = repressed polycomb; 14 = weak repressed polycomb; 15 = quiescent/low.

Gene-mapping

GWS loci obtained by GWAS were mapped to genes in FUMA[18] using three strategies:

1. Positional mapping maps SNPs to genes based on physical distance (within a 10-kb window) from known protein-coding genes in the human reference assembly (GRCh37/hg19).
2. eQTL mapping maps SNPs to genes with which they show a significant eQTL association (that is, allelic variation at the SNP is associated with the expression level of that gene). eQTL mapping uses information from 45 tissue types in 3 data repositories (GTEx[33], Blood eQTL browser[68], BIOS QTL browser[69]), and is based on cis-eQTLs that can map SNPs to genes up to 1 megabase apart. We used a false discovery rate of 0.05 to define significant eQTL associations.
3. Chromatin interaction mapping was performed to map SNPs to genes when there is a three-dimensional DNA–DNA interaction between the SNP region and another gene region. Chromatin interaction mapping can involve long-range interactions since it does not have a distance boundary. FUMA currently contains Hi-C data of 14 tissue types from the study of Schmitt et al.[70]. Since chromatin interactions are often defined in a certain resolution, such as 40 kb, an interacting region can span multiple genes. If an SNP is located in a region that interacts with a region containing multiple genes, it will be mapped to each of those genes. To further prioritize candidate genes, we selected only interaction-mapped genes where one region involved in the interaction overlaps with a predicted enhancer region in any of the 111 tissue/cell types from the Roadmap Epigenomics Project[68], and the other region is located in a gene promoter region (250 bp upstream and 500 bp downstream of the TSS and also predicted by the Roadmap Epigenomics Project to be a promoter region). This method reduces the number of genes mapped but increases the likelihood that those identified will indeed have a plausible biological function. We used a false discovery rate of $< 1 \times 10^{-5}$ to define significant interactions, based on previous recommendations[70] and modified to account for the differences in the cell lines used in this study.

GWAS catalogue look-up

We used FUMA to identify SNPs with previously reported ($P < 5 \times 10^{-5}$) phenotypic associations in published GWAS listed in the NHGRI-EBI catalog[71] which matched with SNPs in linkage disequilibrium with one of the independent significant SNPs identified in the meta-analysis.

Polygenic risk scoring

To calculate the explained variance in insomnia by our GWAS results, we calculated PGS based on the SNP effect sizes in the meta-analysis. The PGS were calculated using two methods: LDpred[72] and PRSice[73] a script for calculating *P*-value thresholded PGS in PLINK. PGS were calculated using a leave-one-out method, where summary statistics were recalculated each time with one sample of $n = 3{,}000$ from the UKB dataset excluded from the analysis. This sample was then used as a target sample for estimating the explained variance in insomnia by the PGS.

Mendelian randomization

To investigate causal associations between insomnia and genetically correlated traits, we analyzed the direction of effects using generalized summary-data-based Mendelian randomization[44]. This method uses effect sizes from GWAS summary statistics (standardized betas or log-transformed ORs) to infer causality of effects between two traits based on GWS SNPs. Built-in HEIDI outlier detection was applied to remove SNPs with pleiotropic effects on both traits, since these may bias the results. We tested for causal associations between insomnia and traits that were significantly genetically correlated with insomnia (b_{zx}). In addition, we tested for bidirectional associations by using other traits as the determinant and insomnia as the outcome (b_{zy}). We selected independent ($r^2 < 0.1$) lead SNPs with a GWS $P < 5 \times 10^{-8}$ as instrumental variables in the analyses. For traits with < 10 lead SNPs (that is, the minimum number of SNPs on which generalized summary-data-based Mendelian randomization can perform a reliable analysis) we selected independent SNPs ($r^2 < 0.1$), with a $P < 1 \times 10^{-5}$. If the outcome trait is binary, the estimated b_{zx} and b_{zy} are approximately equal to the natural log of the OR. An OR of 2 can be interpreted as a doubled risk compared to the population prevalence of a binary trait for every s.d. increase in the exposure trait. For quantitative traits, b_{zx} and b_{zy} can be interpreted as a 1 s.d. increase explained in the outcome trait for every s.d. increase in the exposure trait.

References

1. Wittchen, H.-U. *et al*. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur. *Neuropsychopharmacol.* **21**, 655–679 (2011).
2. Morin, C. M. *et al*. Insomnia disorder. *Nat. Rev. Dis. Prim.* **3**, 15026 (2015).
3. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). (American Psychiatric Pub, 2013).
4. Morphy, H., Dunn, K. M., Lewis, M., Boardman, H. F. & Croft, P. R. Epidemiology of insomnia: a longitudinal study in a UK population. *Sleep* **30**, 274–280 (2007).
5. Lind, M. J., Aggen, S. H., Kirkpatrick, R. M., Kendler, K. S. & Amstadter, A. B. A longitudinal twin study of insomnia symptoms in adults. *Sleep* **38**, 1423–1430 (2015).
6. Hammerschlag, A. R. *et al*. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49**, 1584–1592 (2017).
7. Lane, J. M. *et al*. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281 (2017).
8. Schormair, B. *et al*. Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. *Lancet Neurol.* **16**, 898–907 (2017).
9. 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).
10. Eriksson, N. *et al*. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, e1000993 (2010).
11. Tung, J. Y. *et al*. Efficient replication of over 180 genetic associations with self-reported medical data. *PLoS One* **6**, e23473 (2011).
12. Benjamins, J. S. *et al*. Insomnia heterogeneity: Characteristics to consider for data-driven multivariate subtyping. *Sleep Med. Rev.* **36**, 71–81 (2017).
13. Paparrigopoulos, T. *et al*. Insomnia and its correlates in a representative sample of the Greek population. *BMC Public Health* **10**, 531 (2010).
14. Cho, Y. W. *et al*. Epidemiology of insomnia in Korean adults: prevalence and associated factors. *J. Clin. Neurol.* **5**, 20–23 (2009).
15. Zhang, B. & Wing, Y.-K. Sex differences in insomnia: a meta-analysis. *Sleep* **29**, 85–93 (2006).
16. Willer, C. J., Li, Y., Abecasis, G. R. & Overall, P. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
17. Bulik-Sullivan, B. K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
19. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
20. Sniekers, S. *et al*. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
21. Kircher, M. *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
22. Laramie, J. M. *et al*. Polymorphisms near EXOC4 and LRGUK on chromosome 7q32 are associated with Type 2 Diabetes and fasting glucose; the NHLBI Family Heart Study. *BMC Med. Genet.* **9**, 46 (2008).
23. Butler, M. G., Rafi, S. K. & Manzardo, A. M. High-resolution chromosome ideogram representation of currently recognized genes for autism spectrum disorders. *Int. J. Mol. Sci.* **16**, 6464–6495 (2015).
24. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
25. Kripke, D. F. *et al*. Genetic variants associated with sleep disorders. *Sleep Med.* **16**, 217–224 (2015).
26. Stefansson, H. *et al*. A genetic risk factor for periodic limb movements in sleep. *N. Engl. J. Med.* **357**, 639–647 (2007).
27. Janik, P., Berdyński, M., Safranow, K. & Żekanowski, C. The BTBD9 gene polymorphisms in Polish patients with Gilles de la Tourette syndrome. *Acta Neurobiol Exp* **74**, 218–226 (2014).
28. Freeman, A. *et al*. Sleep fragmentation and motor restlessness in a Drosophila model of Restless Legs Syndrome. *Curr. Biol.* **22**, 1142–1148 (2012).
29. DeAndrade, M. P. *et al*. Motor restlessness, sleep disturbances, thermal sensory alterations and elevated serum iron levels in Btbd9 mutant mice. *Hum. Mol. Genet.* **21**, 3984–3992 (2012).
30. DeAndrade, M. P. *et al*. Enhanced hippocampal long-term potentiation and fear memory in Btbd9 mutant mice. *PLoS One* **7**, e35518 (2012).
31. Suzuki, A., Sinton, C. M., Greene, R. W. & Yanagisawa, M. Behavioral and biochemical dissociation of arousal and homeostatic sleep need influenced by prior wakeful experience in

mice. *Proc. Natl. Acad. Sci.* **110**, 10288–10293 (2013).
32. Liberzon, A. *et al*. The molecular signatures database hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
33. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
34. Toffoli, M. *et al*. SNCA 3' UTR genetic variants in patients with Parkinson's disease and REM sleep behavior disorder. *Neurol. Sci.* **38**, 1233–1240 (2017).
35. Edwards, T. L. *et al*. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Ann. Hum. Genet.* **74**, 97–109 (2010).
36. McDowell, K. A., Shin, D., Roos, K. P. & Chesselet, M.-F. Sleep dysfunction and EEG alterations in mice overexpressing alpha-synuclein. *J. Parkinsons. Dis.* **4,** 531–539 (2014).
37. Colombo, M. A. *et al*. Wake high-density electroencephalographic spatiospectral signatures of insomnia. *Sleep* **39**, 1015–1027 (2016).
38. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
39. Skene, N. G. *et al*. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
40. Saper, C. B., Scammell, T. E. & Lu, J. Hypothalamic regulation of sleep and circadian rhythms. *Nature* **437**, 1257-1263 (2005).
41. Lane, J. M. *et al*. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nat. Commun.* **7**, 10889 (2016).
42. Hu, Y. *et al*. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat. Commun.* **7**, 10448 (2016).
43. Jones, S. E. *et al*. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. *PLoS Genet.* **12**, e1006125 (2016).
44. Zhu, Z. *et al*. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
45. Vetrivelan, R., Qiu, M.-H., Chang, C. & Lu, J. Role of basal ganglia in sleep–wake regulation: neural circuitry and clinical significance. *Front. Neuroanat.* **4**, 145 (2010).
46. Lazarus, M., Huang, Z.-L., Lu, J., Urade, Y. & Chen, J.-F. How do the basal ganglia regulate sleep–wake behavior? *Trends Neurosci.* **35**, 723–732 (2012).
47. Swardfager, W., Rosenblat, J. D., Benlamri, M. & McIntyre, R. S. Mapping inflammation onto mood: Inflammatory mediators of anhedonia. *Neurosci. Biobehav. Rev*. **64**, 148–166 (2016).
48. Wilson, C. J. & Groves, P. M. Spontaneous firing patterns of identified spiny neurons in the rat neostriatum. *Brain Res.* **220**, 67–80 (1981).
49. Qiu, M., Vetrivelan, R., Fuller, P. M. & Lu, J. Basal ganglia control of sleep–wake behavior and cortical activation. *Eur. J. Neurosci.* **31**, 499–507 (2010).
50. Wassing, R. *et al*. Slow dissolving of emotional distress contributes to hyperarousal. *Proc. Natl. Acad. Sci.* **113**, 2538–2543 (2016).
51. Feige, B. *et al*. Insomnia—perchance a dream? Results from a NREM/REM sleep awakening study in good sleepers and patients with insomnia. *Sleep* **41**, zsy032 (2018).
52. Krystal, A. D., Edinger, J. D., Wohlgemuth, W. K. & Marsh, G. R. NREM sleep EEG frequency spectral correlates of sleep complaints in primary insomnia subtypes. *Sleep* **25**, 626–636 (2002).
53. Johann, A. F. *et al*. Insomnia with objective short sleep duration is associated with longer duration of insomnia in the Freiburg Insomnia Cohort compared to insomnia with normal sleep duration, but not with hypertension. *PLoS One* **12**, e0180339 (2017).
54. Stoffers, D. *et al.* The caudate: a key node in the neuronal network imbalance of insomnia? *Brain* **137**, 610–620 (2013).
55. Altena, E. *et al*. Prefrontal hypoactivation and recovery in insomnia. *Sleep* **31**, 1271–1276 (2008).
56. Dai, X.-J. *et al*. Altered intrinsic regional brain spontaneous activity and subjective sleep quality in patients with chronic primary insomnia: a resting-state fMRI study. *Neuropsychiatr. Dis. Treat.* **10**, 2163-2175 (2014).
57. Nagel, M. *et al*. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet*. **50**, 920-927 (2018).
58. Mathur, B. N. The claustrum in review. *Front. Syst. Neurosci*. **8**, 48 (2014).
59. Wei, Y. *et al*. I keep a close watch on this heart of mine: increased interoception in insomnia. *Sleep* **39**, 2113–2124 (2016).
60. Renouard, L. *et al*. The supramammillary nucleus and the claustrum activate the cortex during REM sleep. *Sci. Adv.* **1**, e1400177 (2015).
61. Hawrylycz, M. *et al*. Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18**, 1832 (2015).
62. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
63. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
64. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
65. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
66. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
67. Westra, H.-J. *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
68. Roadmap Epigenomics Consortium *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
69. Zhernakova, D. V *et al*. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
70. Schmitt, A. D. *et al*. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
71. MacArthur, J. *et al*. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).
72. Vilhjálmsson, B. J. *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
73. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31,** 1466–1468 (2014).

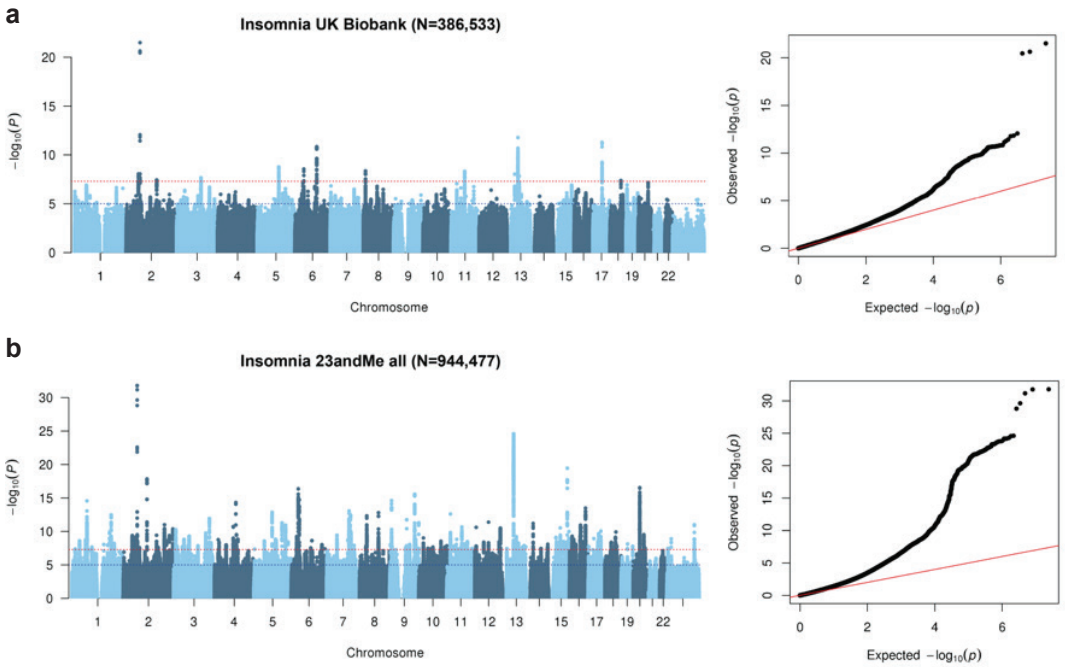
Supplementary information

Supplementary information, figures (1-11) and tables (1-33) can be found in the online manuscript:

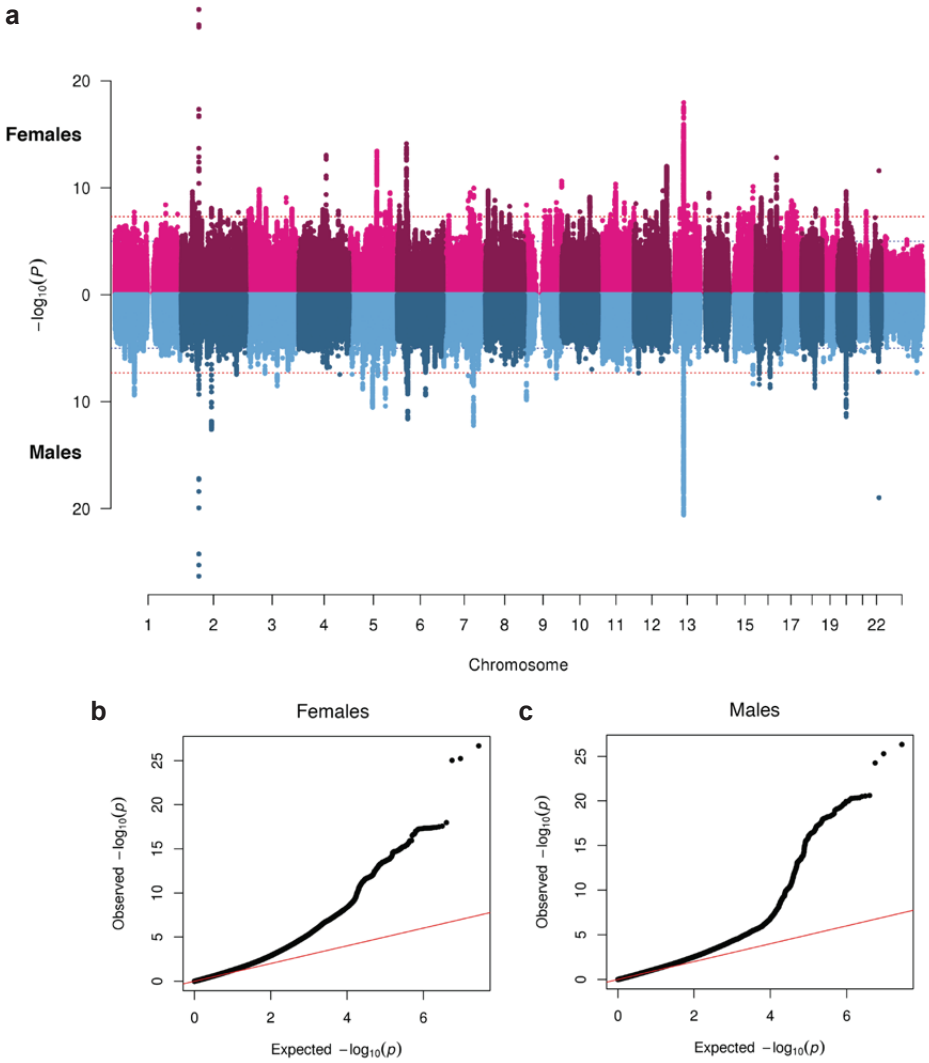
https://www.nature.com/articles/s41588-018-0333-3

Supplementary information

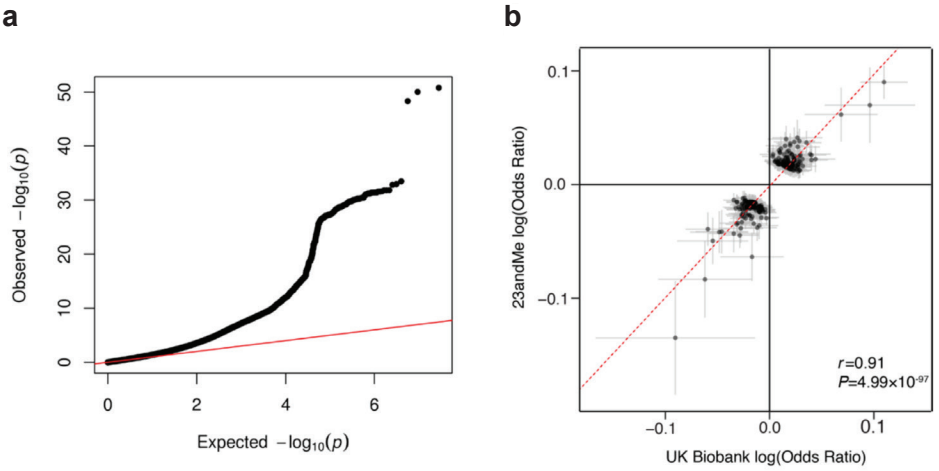
Supplementary Fig. 1 | Manhattan and Q-Q plots of the genome-wide analysis of insomnia in 1,331,010 individuals. Results are shown for the genome-wide analysis in (a) UK Biobank (*n* = 386,533 individuals) and (**b**) 23andMe (*n* = 944,477 individuals). The Manhattan plot shows the $-\log_{10}$-transformed *P*-value on the y-axis and the chromosomal position on the x-axis. Inflation in observed median *P*-value in the Q-Q plots were 1.307 (UKB) and 1.699 (23andMe). The LD Score intercepts were 1.014 (UKB) and 1.075, indicating that the inflation in both analyses was largely explained by a highly polygenic trait.



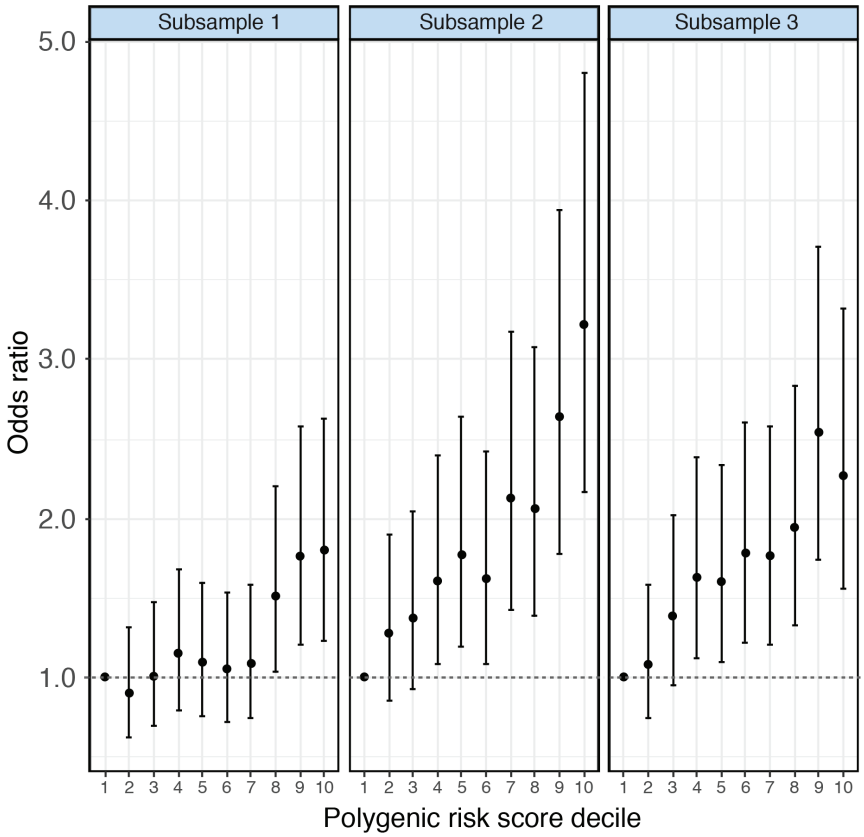
Supplementary Fig. 2 | Sex-specific Manhattan plot and Q-Q plot of the insomnia meta-analysis in males and females (UK Biobank + 23andMe). (**a**) Miami plot showing sex-specific $-\log_{10}$ transformed SNP *P*-values for females on the upper side (n = 709,986 females) and males (n = 621,024 males) on the lower side. Two-sided SNP *P*-values were calculated using METAL. (**b**) Q-Q plot in females, and (**c**) in males.



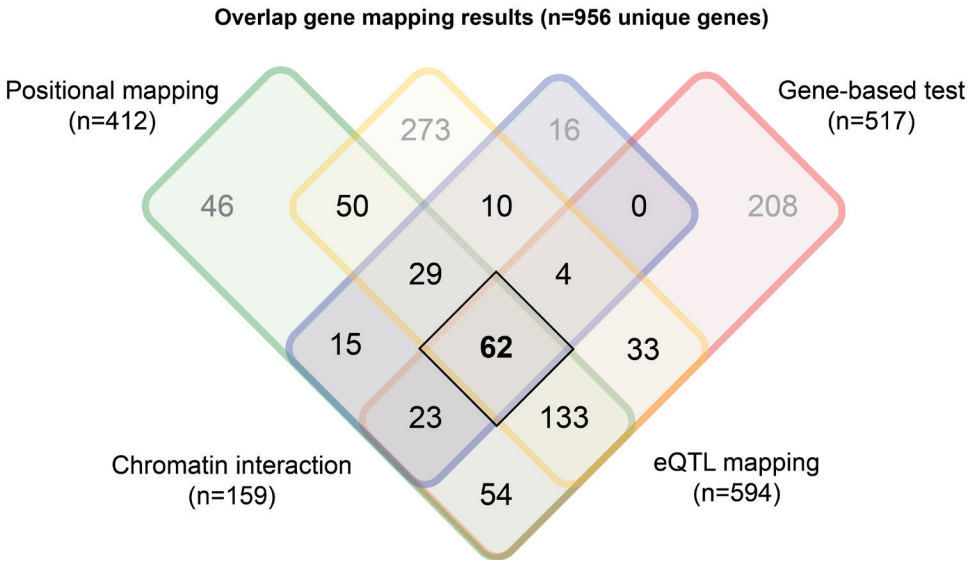
Supplementary Fig. 3 | Q-Q plot and lead SNPs of the GWAS meta-analysis for insomnia in 1,331,010 individuals. (**a**) QQ-plot of the insomnia meta-analysis showing the expected $-\log_{10}$ transformed *P*-value distribution on the x-axis, and the observed negative log10-transformed *P*-value on the y-axis, (**b**) effect size plot of the 248 lead SNP of the insomnia meta-analysis. The dots represent the SNP log-transformed odds ratio in UK Biobank on the x-axis, and in 23andMe on the y-axis. The error bars represent the 95% confidence intervals of the log-transformed odds ratios in both samples.



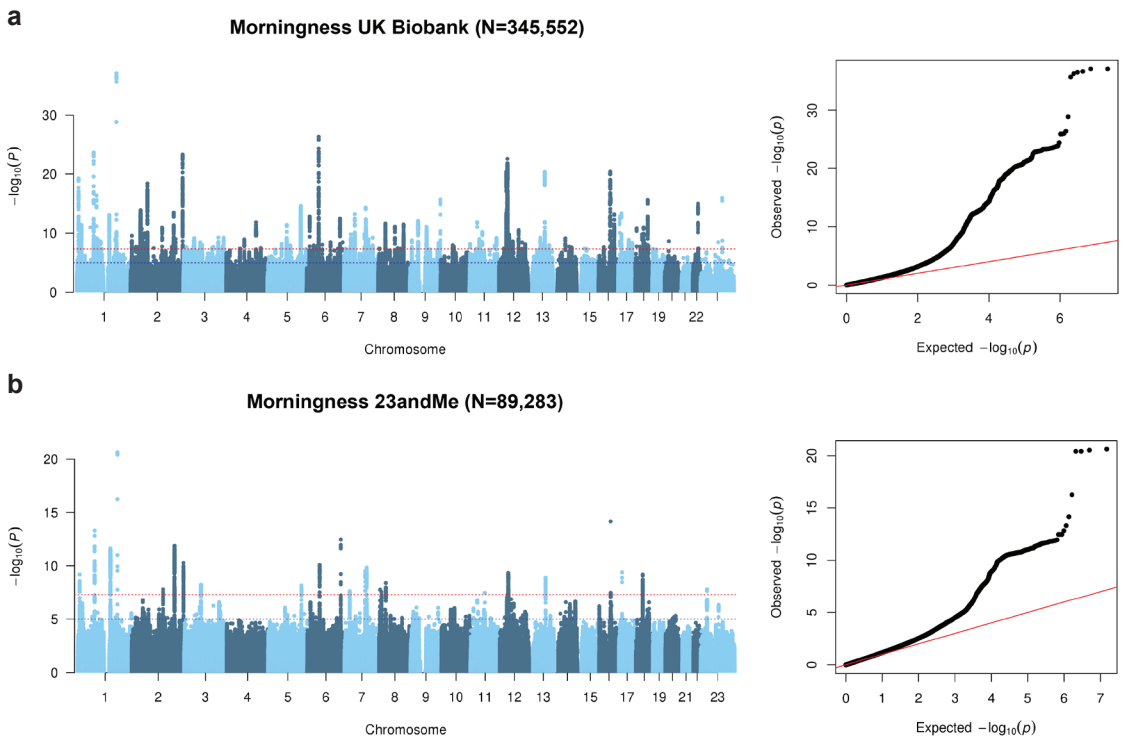
Supplementary Fig. 5 | Risk of insomnia per polygenic risk score decile in three independent holdout samples (n=3×3,000 individuals). Odds ratios for deciles in polygenic risk score were calculated based on a logistic regression model, using the lowest polygenic risk score decile as the reference. The estimate represents the odds ratio of insomnia in each decile compared to the lowest polygenic risk score decile. The error bars represent the 95% confidence interval (CI) around the odds ratio. Polygenic risk scores were calculated based on GWAS after exclusion of the holdout sample (*n* = 1,228,010 individuals).



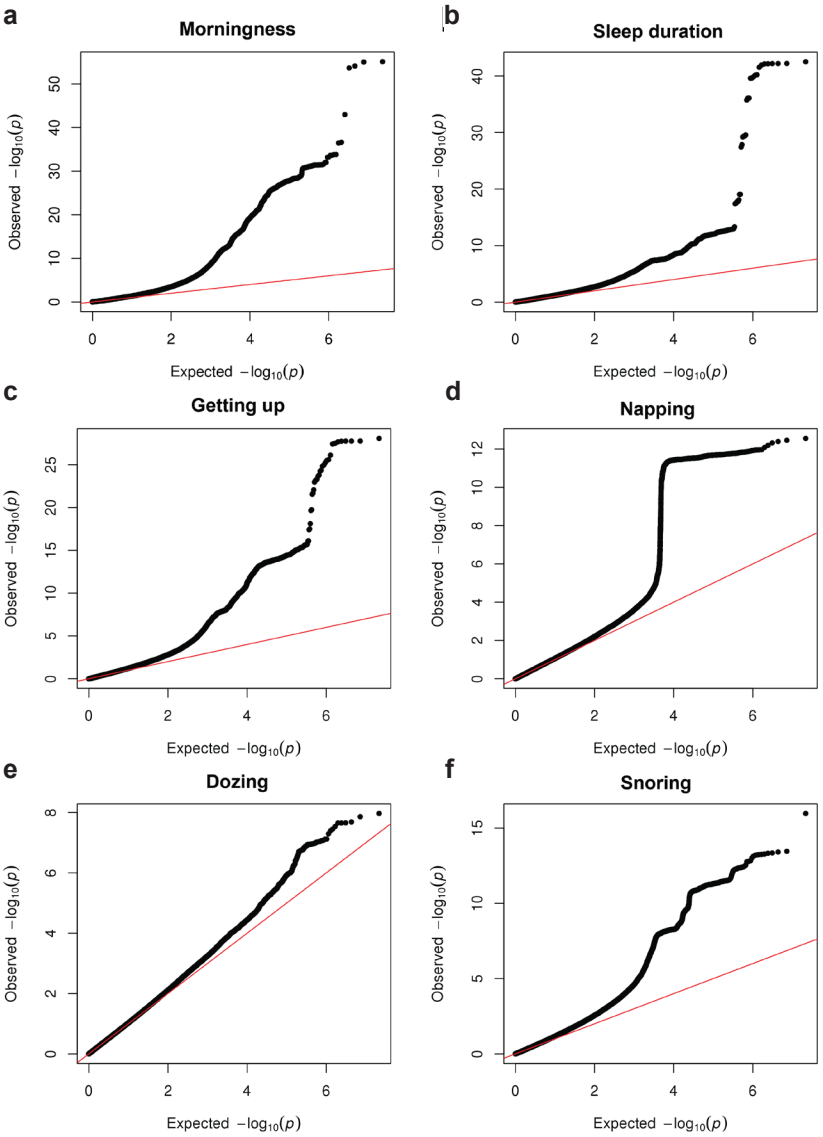
Supplementary Fig. 7 | Venn diagram showing the number of genes that were mapped by four gene-mapping strategies. Each square shows the number of overlapping genes (n represents the total number of genes mapped by that strategy) between three gene-mapping methods in FUMA (positional mapping, eQTL mapping and chromatin interaction mapping) and significant genes in gene-based tests in MAGMA. The number of genes in bold highlights the number of genes that were implicated by all four methods.



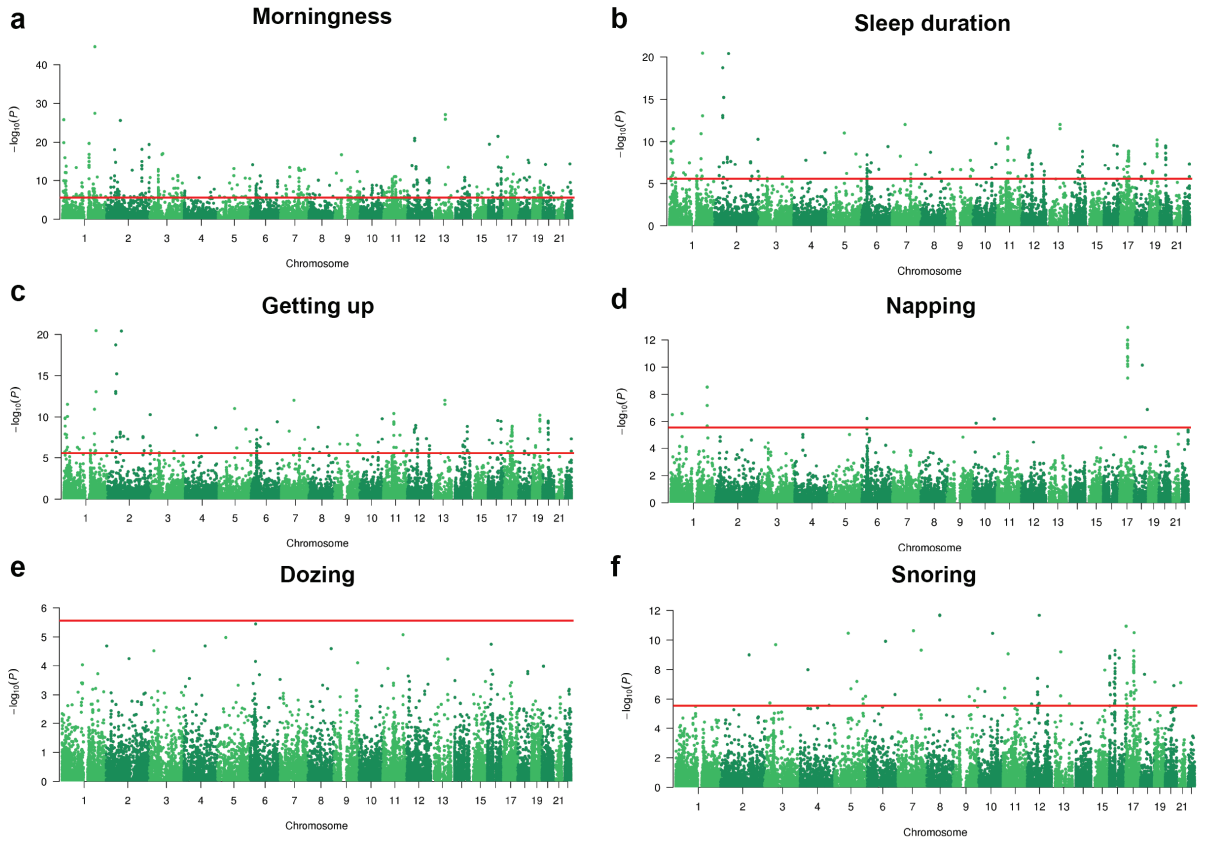
Supplementary Fig. 8 | Manhattan plot and Q-Q plot of the genome-wide analysis of morningness in UK Biobank and 23andMe. Results are shown for (**a**) UK Biobank (n = 345,552 individuals) and (**b**) 23andMe (n =89,283 individuals).



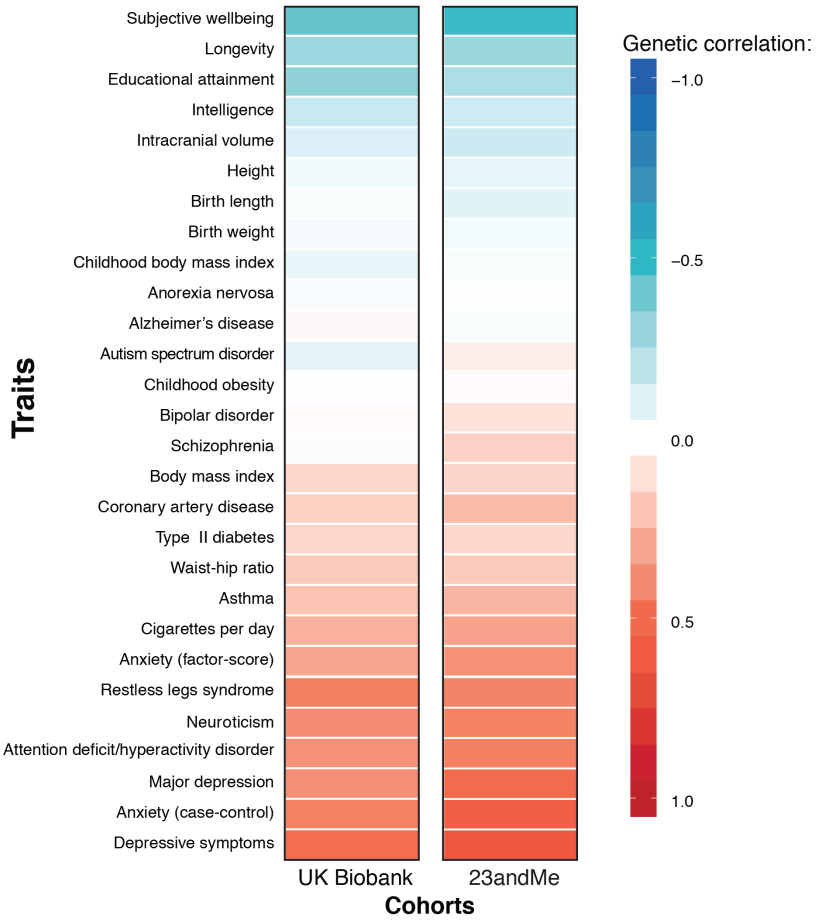
Supplementary Fig. 9 | Q-Q plots of the genome-wide analysis of six sleep related traits. (a) morningness (including UKB and 23andMe, *n* = 434,835 individuals), **(b)** sleep duration (*n* = 384,317 individuals), **(c)** ease of getting up (*n* = 385,949 individuals) **(d)** daytime napping (*n* = 386,577 individuals), **(e)** daytime dozing (*n* = 385,333 individuals), **(f)** snoring (*n* = 359,916 individuals). Manhattan plots of the genome-wide analyses are shown in **Fig. 3**.



Supplementary Fig. 10 | Genome-wide gene-based association analysis of six sleep-related phenotypes. Manhattan plots of the genome-wide gene-based analysis (GWGAS) results for **(a)** morningness (n = 434,835 individuals), **(b)** sleep duration (n = 384,317 individuals), **(c)** ease of getting up (n = 385,949 individuals), **(d)** daytime napping (n = 386,577 individuals), **(e)** daytime dozing (n = 385,333 individuals), **(f)** snoring (n = 359,916 individuals). Two-sided gene-based P-values were calculated using MAGMA. The analysis of morningness was based on GWAS meta-analysis of UKB and 23andMe, while other sleep-related phenotypes were analysed in UKB. The red line indicates Bonferroni corrected significance threshold depending on the number of genes tested.



Supplementary Fig. 11 | Genetic correlations between previous GWAS studies and insomnia in UK Biobank and 23andMe separately. Heatmap showing genetic correlations as estimated by LD Score regression with insomnia in UK Biobank and 23andMe. The Pearson correlation between the genetic correlations within each sample was 0.98.



Chapter 4

Meta-analysis of Genome-wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways

Mats Nagel*, Philip R. Jansen*, Sven Stringer, Kyoko Watanabe, Christiaan A. de Leeuw, Julien Bryois, Jeanne E. Savage, Anke R. Hammerschlag, Nathan G. Skene, ana B. Munoz-Manchado, 23andMe Research Team, Tonya White, Henning Tiemeier, Sten Linnarson, Jens Hjerling-Leffler, Tinca J.C. Polderman, Patrick F. Sullivan, Sophie van der Sluis[#], Danielle Posthuma[#]

* = *authors contributed equally*

[#] = *authors jointly supervised this work*

Published in *Nature Genetics* 2018 Jul;50(7):920-927

Neuroticism is an important risk factor for psychiatric traits including depression[1], anxiety[2,3], and schizophrenia[4–6]. Previous genome-wide association studies[7–12] (GWAS) reported 16 genomic loci[10–12]. Here, we report the largest neuroticism GWAS meta-analysis to date (n = 449,484), and identify 136 independent genome-wide significant loci (124 novel at the time of analysis), implicating 599 genes. Extensive functional follow-up analyses show enrichment in several brain regions and involvement of specific cell types, including dopaminergic neuroblasts ($P = 3.49 \times 10^{-8}$), medium spiny neurons ($P = 4.23 \times 10^{-8}$) and serotonergic neurons ($P = 1.37 \times 10^{-7}$). Gene-set analyses implicate three specific pathways: neurogenesis ($P = 4.43 \times 10^{-9}$), behavioural response to cocaine processes ($P = 1.84 \times 10^{-7}$), and axon part ($P = 5.26 \times 10^{-8}$). We show that neuroticism's genetic signal partly originates in two genetically distinguishable subclusters[13] (depressed affect and worry, the former being genetically strongly related to depression, rg = 0.84), suggesting distinct causal mechanisms for subtypes of individuals. Mendelian randomization showed uni- and bidirectional effects between neuroticism and multiple psychiatric traits. These results vastly enhance our neurobiological understanding of neuroticism, and provide specific leads for functional follow-up experiments.

The neuroticism meta-analysis comprised data from the UK Biobank Study (UKB, full release[14]; *n* = 372,903; **Online Methods; Supplementary Fig. 1-2**), 23andMe, Inc.[15] (*n* = 59,206), and the Genetics of Personality Consortium (GPC1[9]; *n* = 17,375; **Online Methods**, *n* = 449,484 in total). In all samples, neuroticism was measured through (digital) questionnaires (**Online Methods; Supplementary Note**). To achieve optimal power, SNP associations were meta-analyzed using METAL[16], weighted by sample size (**Online Methods**). We choose to meta-analyze the available samples, rather than use a two-stage discovery-replication strategy, because Skol et al.[17] showed that this is almost always more powerful, even though less correcting for multiple testing is required in the replication stage.

The quantile-quantile (Q-Q) plot of the genome-wide meta-analysis on 449,484 subjects and 14,978,477 SNPs showed inflation (LD Score regression (LDSC)[18]: λ_{GC}= 1.65, mean χ^2 statistic = 1.91; **Fig. 1a**; **Supplementary Table 1**), yet the LDSC intercept (1.02; SE=0.01) and ratio (2.1%) both indicated that the inflation was largely explained by true polygenicity and large sample size[19]. The λ_{GC}= 1.65 is in line with values observed in recent large-sample GWAS studies (i.e. *n* > 100,000) for diverse and polygenic traits (see **Supplementary Note**). The LDSC SNP-based heritability (h^2_{SNP}) of neuroticism was 0.100 (SE = 0.003). The GWAS meta-analysis identified 9,745 genome-wide significant (GWS) SNPs ($P < 5 \times 10^8$), of which 157 and 2,414 were located in known associated inversions on chromosomes 8 and 17[10–12], respectively (see **Supplementary Table 2** for cohort-specific information; **Fig. 1b**; **Supplementary Fig. 3**).

FUMA[20], a tool to functionally map and annotate GWAS results (**Online Methods**), extracted 170 independent lead SNPs (158 novel; see **Online Methods** for definition of lead SNPs), which mapped to 136 independent genomic loci (124 novel at the time of analysis; **Online Methods**; **Supplementary Note; Supplementary Tables 3-8**). Of all lead SNPs, 4 were in exonic, 88 in intronic, and 52 in intergenic regions. Of the 17,794 SNPs in high LD with one of the independent significant SNPs (see **Online Methods** for definition), most were intronic (9,147: 51,4%) or intergenic (5,460: 30,7%), and 3.8% was annotated as potentially having a functional impact, with 0.9% (155 SNPs) being exonic (**Fig. 1c**, **Supplementary Table 9**; see **Supplementary Tables 10-11** for an overview of chromatin state and regulatory functions of these SNPs). Of these 155, 70 were exonic non-synonymous (ExNS) (**Table 1, Supplementary Table 12**). The ExNS SNP with the highest CADD (Combined Annotation Dependent Depletion[21]; indicating likelihood of being deleterious) score (34; see **Online Methods**) was rs17651549, located on chromosome 17 in exon 6 of *MAPT*, with a GWAS *P*-value of 1.11×10^{-28}, in high LD with the lead SNP in that region ($r^2 = 0.97$). rs17651549 is a missense mutation leading to an arginine-to-tryptophan change with allele frequencies matching the inversion in that region. The ancestral allele C is associated with a lower neuroticism score (see **Table 1** and **Supplementary Table 12** for a detailed overview of all functional variants in genomic risk loci). Stratified LDSC[22] (**Online Methods**), showed significant enrichment for h^2 of SNPs located in conserved regions (enrichment = 13.79, $P = 5.14 \times 10^{-16}$), intronic regions (enrichment = 1.24 $P = 1.27 \times 10^{-6}$), and in H3K4me3 (enrichment = 2.14, $P = 1.02 \times 10^{-5}$) and H3K9ac regions (enrichment = 2.17, $P = 3.06 \times 10^{-4}$) (**Fig. 1d**; **Supplementary Table 13).**

Polygenic scores (PGS) calculated using PRSice[23] (clumping followed by *P*-value thresholding) and LDpred[24] in three randomly drawn hold-out samples (UKB only, *n* = 3,000 each; **Online Methods**), explained up to 4.2%

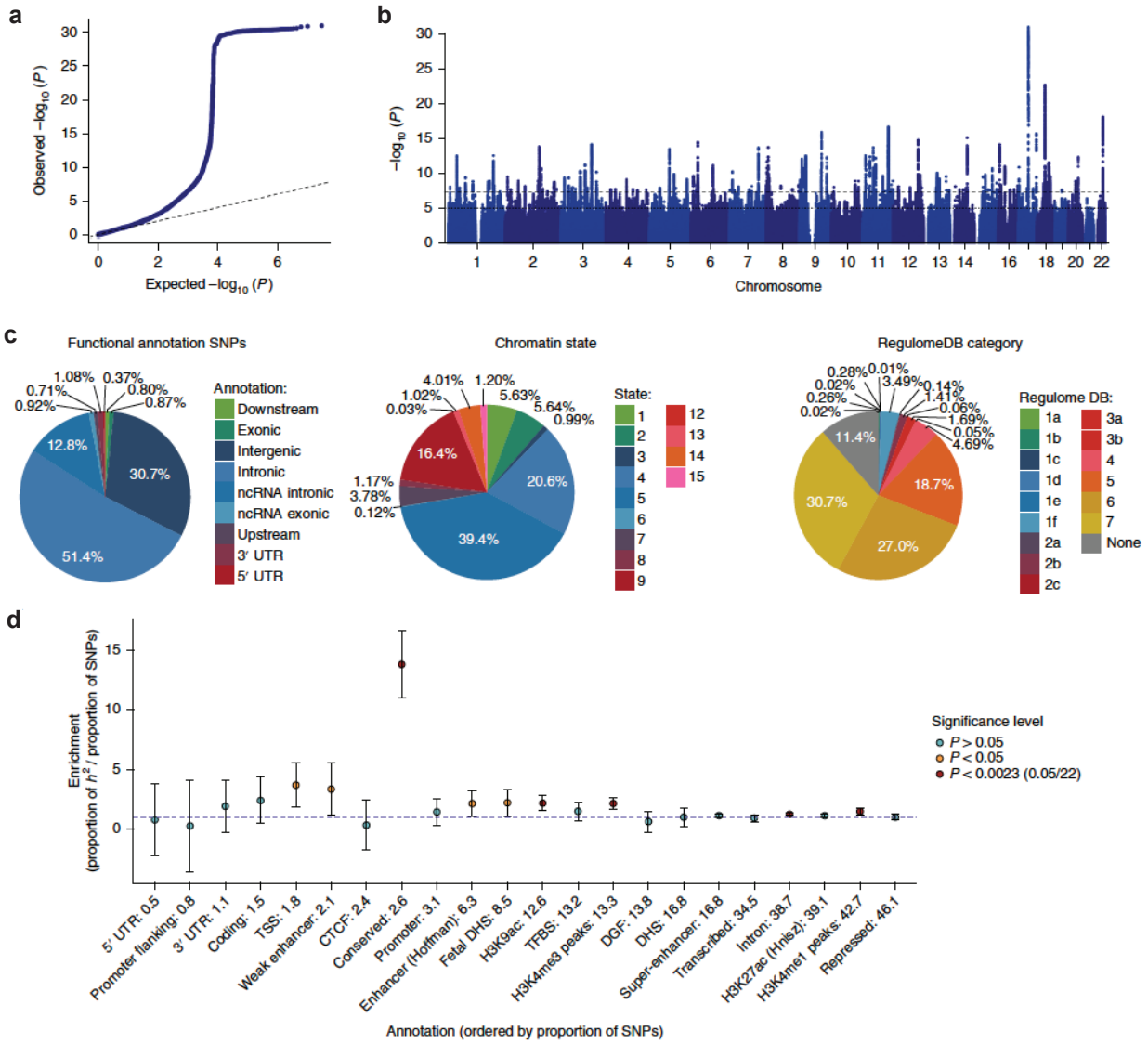


Fig. 1 | SNP-based associations with neuroticism in the GWAS meta-analysis.(a) Quantile-quantile plot of the SNP-based associations with neuroticism (n = 449,484 individuals). SNP P-values were computed in METAL using a two-sided, weighted z-score method. **(b)** Manhattan plot showing the $-\log_{10}$ transformed P-value of each SNP on the y-axis and base pair positions along the chromosomes on the x-axis (n = 449,484 individuals). SNP P-values were computed in METAL using a two-sided, weighted z-score method. The dashed line indicates genome-wide significance ($P < 5 \times 10^{-8}$), the dotted line the threshold for suggestive associations ($P < 1 \times 10^{-5}$). **(c)** Pie charts showing the distribution of functional consequences of SNPs in linkage disequilibrium (LD) with genome-wide significant lead SNPs in the meta-analysis, the minimum chromatin state across 127 tissue and cell types and the distribution of regulome DB score, a categorical score between 1a and 7, indicating biological evidence of a SNP being a regulatory element, with a low score denoting a higher likelihood of being regulatory. **(d)** Heritability enrichment of 22 functional SNP annotations calculated with stratified LD Score regression (summary statistics of the meta-analysis of neuroticism were used as input for this analysis). The dots signify the estimated enrichment, whereas the dashed line indicates Enrichment = 1. Error bars represent 95% confidence intervals.

Table 1 | Exonic non-synonymous (ExNS) variants in the genomic loci associated with neuroticism and in LD (r^2>0.6) with one of the ndependent GWS SNPs.

rsID	Exon	Gene	A1	MAF	gwas P	Z-score	r^2	Indep. Sign. SNP	Locus	CADD	RDB	Min. Chromatin state
rs41266050	14	*RABGAP1L*	T	0.25	5.65E-06	-4.54	0.84	rs7536102	6	2.85	7	5
rs34605051	10	*KDM3A*	T	0.16	1.64E-07	-5.24	0.99	rs11127043	14	13.85	4	4
rs2073498	3	*RASSF1*	A	0.11	2.71E-08	5.56	0.98	rs6776145	25	19.43	7	3
rs4434138	62	*STAB1*	A	0.46	3.10E-07	5.12	0.93	rs2015971	26	5.48	5	4
rs66782572	1	*NT5DC2*	A	0.46	2.65E-06	4.70	0.67	rs2015971	26	0.00	4	1
rs11177	3	*GNL3*	A	0.38	2.49E-07	-5.16	0.75	rs2015971	26	22.90	1d	1
rs2289247	11	*GNL3*	A	0.41	2.28E-06	-4.73	0.65	rs2015971	26	12.82	1f	3
rs6617	1	*SPCS1*	C	0.41	1.96E-06	4.76	0.65	rs2015971	26	0.00	1f	1
rs1029871	5	*NEK4*	C	0.38	2.29E-07	-5.17	0.75	rs2015971	26	24.10	1f	2
rs678	12	*ITIH1*	A	0.36	1.44E-06	4.82	0.65	rs2015971	26	25.90	1f	2
rs1042779	12	*ITIH1*	A	0.37	6.09E-06	4.52	0.63	rs2015971	26	0.15	1f	2
rs198844	1	*HIST1H1T*	C	0.47	3.07E-08	-5.54	0.98	rs198825	45	0.10	1f	5
rs200484	1	*HIST1H2BL*	A	0.13	1.53E-07	5.25	0.61	rs200965	46	2.68	1f	1
rs240780	39	*ASCC3*	C	0.43	3.01E-08	5.54	0.96	rs240769	49	19.95	7	4
rs3173615	6	*TMEM106B*	C	0.41	2.08E-08	-5.61	0.67	rs11509880	51	21.40	6	4
rs11765552	11	*LMTK2*	A	0.46	7.68E-08	5.38	0.98	rs34320230	55	12.24	6	4
rs10821128	3	*FAM120AOS*	T	0.33	2.73E-09	5.95	0.99	rs10821129	71	0.05	4	4
rs41274386	2	*FAM120AOS*	T	0.08	1.10E-07	5.31	0.66	rs78046549	71	2.36	4	1
rs1055710	1	*FAM120AOS*	A	0.33	1.11E-09	-6.09	0.99	rs10821129	71	0.05	4	1
rs3816614	33	*LRP4*	T	0.23	5.69E-07	5.00	0.90	rs7940441	84	22.70	NA	4
rs2030166	5	*NDUFS3*	T	0.35	2.02E-10	-6.36	0.93	rs11039389	84	3.13	6	4
rs1064608	13	*MTCH2*	C	0.35	1.15E-10	-6.45	0.93	rs11039389	84	25.40	6	4
rs12286721	13	*AGBL2*	A	0.45	7.81E-08	-5.37	0.78	rs7107356	84	14.22	1f	5
rs3816605	5	*NUP160*	T	0.46	5.68E-08	-5.43	0.75	rs7107356	84	6.61	6	4
rs4926	7	*SERPING1*	A	0.27	6.12E-07	4.99	0.86	rs73480560	85	23.50	5	4
rs11604671	6	*ANKK1*	A	0.49	2.57E-10	-6.32	0.64	rs2186800	88	1.39	5	4
rs2734849	8	*ANKK1*	A	0.49	8.43E-10	6.14	0.64	rs2186800	88	0.00	3a	4
rs1800497	8	*ANKK1*	A	0.20	8.45E-06	4.45	0.69	rs11214607	88	0.81	4	4
rs3825393	30	*MYO1H*	T	0.36	2.95E-07	-5.13	0.79	rs2111216	94	10.93	1f	4
rs2058804	2	*KCTD10*	A	0.48	1.82E-10	-6.38	0.76	rs2111216	94	2.11	6	4
rs7298565	12	*UBE3B*	A	0.48	2.24E-10	6.34	0.76	rs2111216	94	22.70	6	4

Table 1 | Exonic non-synonymous (ExNS) variants in the genomic loci associated with neuroticism and in LD (r^2>0.6) with one of the ndependent GWS SNPs.

rsID	Exon	Gene	A1	MAF	gwas P	Z-score	r^2	Indep. Sign. SNP	Locus	CADD	RDB	Min. Chromatin state
rs9593	9	*MMAB*	A	0.48	8.54E-10	-6.14	0.76	rs2111216	94	0.53	1f	4
rs8007859	7	*EXD2*	T	0.39	2.28E-08	5.59	0.80	rs1275411	108	3.95	5	4
rs2286913	4	*RPS6KL1*	A	0.37	1.46E-07	5.26	0.89	rs3213716	110	12.96	5	2
rs7156590	3	*RPS6KL1*	T	0.37	2.79E-07	5.14	0.86	rs3213716	110	19.46	5	4
rs35755513	NA	*CSNK1G1*	T	0.07	2.87E-08	5.55	1.00	rs35755513	114	23.90	4	1
rs12443627	1	*ENSG00000268863*	C	0.37	1.28E-10	6.43	0.77	rs3751855	119	3.58	2b	1
rs9938550	7	*HSD3B7*	A	0.37	1.48E-10	-6.41	0.79	rs3751855	119	0.04	1d	3
rs35713203	1	*ZNF646*	C	0.38	3.67E-11	-6.62	0.98	rs3751855	119	0.05	2b	3
rs7196726	1	*ZNF646*	A	0.38	1.29E-11	-6.77	1.00	rs3751855	119	0.00	2b	3
rs7199949	8	*PRSS53*	C	0.38	1.32E-11	-6.77	1.00	rs3751855	119	0.00	2b	2
rs3803704	3	*CMTR2*	T	0.25	7.14E-08	-5.39	0.97	rs1424144	121	0.07	6	4
rs3748400	12	*ZCCHC14*	T	0.23	8.83E-09	-5.75	0.98	rs2042395	122	24.00	5	4
rs12949256	1	*ARHGAP27*	T	0.19	1.47E-23	10.00	0.73	rs77804065	126	11.97	4	1
rs16940674	6	*CRHR1*	T	0.23	5.24E-29	11.18	0.97	rs77804065	126	12.86	1f	5
rs16940681	13	*CRHR1*	C	0.23	2.18E-30	11.46	0.97	rs77804065	126	1.76	4	5
rs62621252	1	*SPPL2C*	T	0.23	9.05E-31	-11.53	0.97	rs77804065	126	0.00	5	5
rs242944	1	*SPPL2C*	A	0.44	2.88E-12	-6.98	1.00	rs242947	126	0.00	5	5
rs62054815	1	*SPPL2C*	A	0.23	1.74E-30	11.48	0.97	rs77804065	126	0.00	5	5
rs12185233	1	*SPPL2C*	C	0.23	6.76E-29	11.16	0.96	rs77804065	126	25.60	1f	5
rs12185268	1	*SPPL2C*	A	0.23	4.08E-30	-11.40	0.97	rs77804065	126	0.00	1f	5
rs12373123	1	*SPPL2C*	T	0.23	7.80E-29	-11.14	0.97	rs77804065	126	22.70	1f	5
rs12373139	1	*SPPL2C*	A	0.23	2.19E-30	11.46	0.97	rs77804065	126	0.53	1f	5
rs12373142	1	*SPPL2C*	C	0.22	2.60E-28	-11.04	0.97	rs77804065	126	0.12	1f	5
rs754512	1	*MAPT*	A	0.23	1.17E-28	-11.11	0.97	rs77804065	126	2.39	1d	4
rs63750417	6	*MAPT*	T	0.23	4.89E-30	11.39	0.97	rs77804065	126	8.68	5	4
rs62063786	6	*MAPT*	A	0.23	1.05E-29	11.32	0.97	rs77804065	126	7.65	5	4
rs62063787	6	*MAPT*	T	0.23	4.57E-30	-11.39	0.97	rs77804065	126	0.00	5	4
rs17651549	6	*MAPT*	T	0.23	1.11E-28	11.11	0.97	rs77804065	126	34.00	1f	4
rs10445337	8	*MAPT*	T	0.23	1.41E-28	-11.09	0.96	rs77804065	126	9.93	1f	4
rs62063857	1	*STH*	A	0.23	3.71E-30	-11.41	0.97	rs77804065	126	0.00	7	4
rs34579536	15	*KANSL1*	A	0.23	1.92E-30	-11.47	0.96	rs77804065	126	8.02	3a	3
rs34043286	8	*KANSL1*	A	0.23	3.14E-30	-11.43	0.97	rs77804065	126	15.71	4	4
rs4969391	14	*BAIAP2*	A	0.16	1.69E-14	7.67	0.90	rs56084168	128	12.58	4	4
rs2282632	11	*ASXL3*	A	0.50	3.37E-08	-5.52	0.73	rs10460051	129	1.54	6	4

Table 1 | Exonic non-synonymous (ExNS) variants in the genomic loci associated with neuroticism and in LD (r^2>0.6) with one of the ndependent GWS SNPs.

rsID	Exon	Gene	A1	MAF	gwas P	Z-score	r^2	Indep. Sign. SNP	Locus	CADD	RDB	Min. Chromatin state
rs7232237	12	*ASXL3*	A	0.50	1.59E-08	-5.65	0.84	rs10460051	129	0.00	5	4
rs17522826	1	*TCF4*	A	0.18	2.17E-10	6.35	0.60	rs10503002	133	14.22	5	1
rs20551	15	*EP300*	A	0.29	3.44E-18	-8.70	0.98	rs9611519	138	3.23	5	4
rs139431	2	*L3MBTL2*	T	0.37	9.45E-07	-4.90	0.63	rs7289932	138	10.26	7	4
rs739134	2	*C22orf46*	T	0.19	4.55E-07	5.04	0.61	rs761366	138	22.60	1f	4

SNP *P*-values and *z*-scores were computed in METAL by a weighted *z*-score method (two-sided test). Per-SNP *N* are reported in **Supplementary Table 2** (for genome-wide significant SNPs) and in the publicly available summary statistics. rsID = rs number of the ExNS SNP; Exon = exon in which the SNP is located; Gene = nearest gene; A1 = effect allele; MAF = minor allele frequency; gwas *P* = SNP *P*-value in the GWAS meta-analysis; *Z*-score = *z*-score from the GWAS meta-analysis in METAL; r^2 = maximum r^2 of the SNP with one of the independent significant SNPs; Locus = index of the genomic risk locus; CADD = CADD score; RDB = regulome DB score; Min. Chromatin state = Minimum chromatin state of the SNP. Results are reported on hg19 coordinates (NCBI b37). Genes containing multiple ExNS are annotated in red.

($P = 1.39 \times 10^{-30}$) of the variance in neuroticism **(Supplementary Fig. 4**; **Supplementary Table 14**; **Supplementary Note**). Although the current sample size is considered large for GWAS and PGS scores can be calculated with relatively low standard errors, the variance explained by all SNPs combined in the PGS is still relatively small, although this is not unexpected given the h^2_{SNP} of 10%. Our current results thus have little predictive power in independent samples, mostly due to the low average effect sizes of contributing SNPs, and indicate that the genetic architecture of neuroticism is extremely polygenic. We do note that our current meta-analysis did not include possible genetic interactions (as even with the current sample sizes, power would be limited) but that adding these in the future may increase the predictive value of PGS for neuroticism.

We used four strategies to link our SNP results to genes: positional, eQTL, and chromatin interaction mapping (**Online Methods**), and genome-wide gene-based association study (GWGAS; MAGMA[25]). GWGAS evaluates the joint association effect of all SNPs within a gene yielding a gene-based *P*-value. Based on our meta-analytic results, 283 genes were implicated through positional mapping, 369 through eQTL-mapping, and 119 through chromatin interaction-mapping (**Fig. 2a; Supplementary Table 15**). GWGAS identified 336 GWS genes ($P < 2.75 \times 10^{-6}$, **Figs. 2b-c**; **Supplementary Table 16**; **Supplementary Note**), of which 203 overlapped with genes implicated by FUMA, resulting in 599 unique neuroticism-related genes. Of these, 50 were implicated by all four methods, of which 49 had chromatin interaction and eQTL associations in the same tissue/cell type (**Fig. 2a**; **Supplementary Table 15**). 19 of the 119 genes implicated through chromatin interaction mapping are especially interesting as they are implicated via interactions between two independent GWS genomic risk loci. There are several chromatin interactions in 7 tissue types (aorta, hippocampus, left ventricle, right ventricle, liver, spleen, pancreas) across two risk loci on chromosome 6 (**Fig. 3a**). Two genes are located in locus 45 and are mapped by chromatin interactions from risk locus 46 (*HFE* and *HIST1H4C*), and 16 genes encode histones in locus 46 and are mapped by interactions from locus 45 (**Supplementary Table 15**). One gene, *XKR6*, is located on chromosome 8 in risk locus 61, and is implicated by chromatin interactions in 5 tissue types (aorta, left ventricle, liver, pancreas and spleen) including cross-locus interactions from locus 60 (**Fig. 3b**; **Supplementary Table 15**). This gene is also mapped by eQTLs in blood and transformed fibroblasts. This gene is also mapped by eQTLs in blood and transformed fibroblasts. Out of

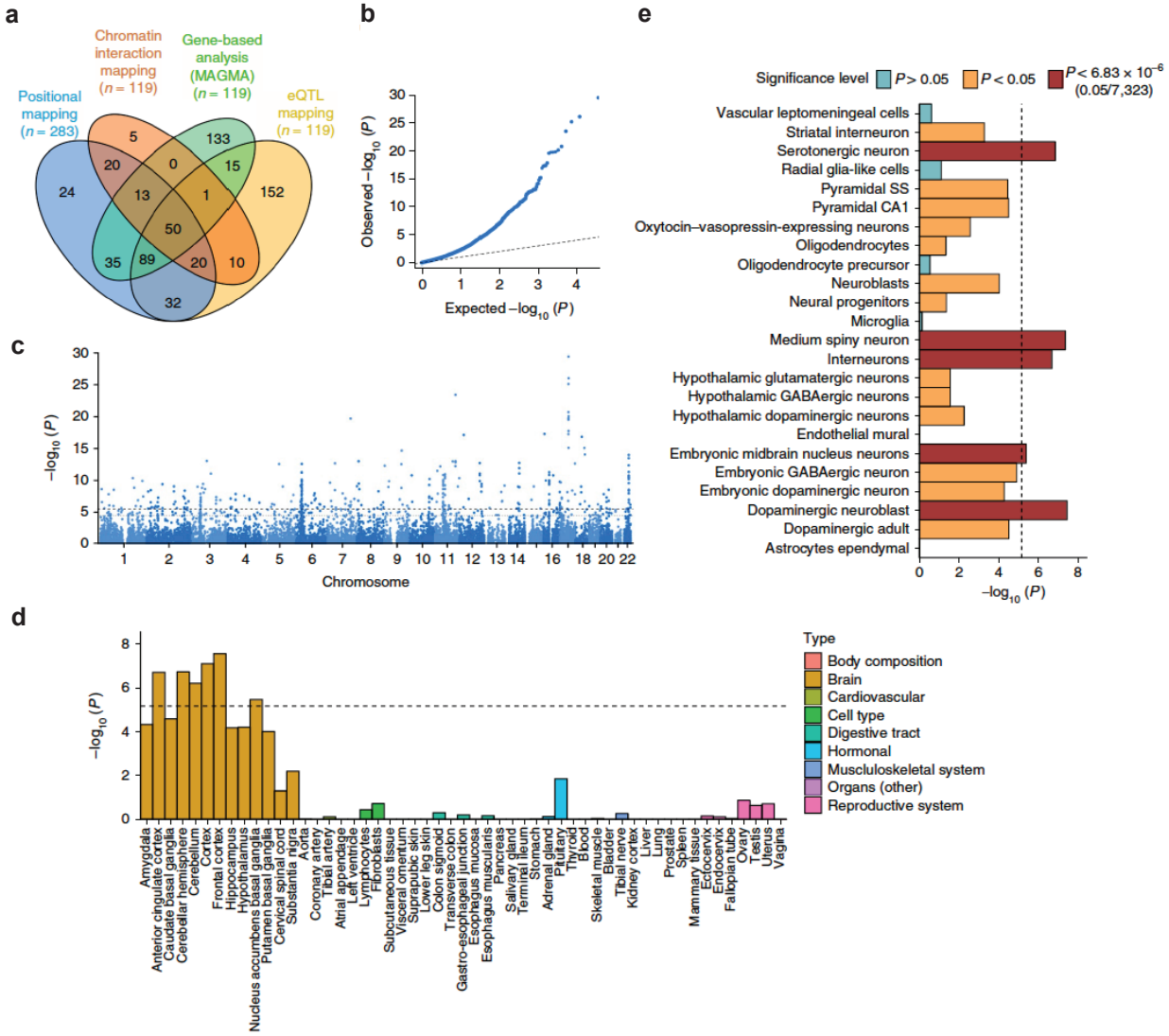


Fig. 2 | Mapping of genes and tissue- and cell expression profiles. (a) Venn diagram showing overlap of genes implicated by positional mapping, eQTL mapping, chromatin interaction mapping, and genome-wide gene-based association (GWGAS). **(b)** Quantile-quantile plot of the GWGAS (n = 449,484 individuals). Gene P-values were computed using MAGMA's gene-based test. **(c)** Manhattan plot of the GWGAS on neuroticism (n = 449,484 individuals). Gene P-values were computed using MAGMA's gene-based test. The y-axis shows the $-\log_{10}$ transformed P-value of each gene, and the chromosomal position (start position) on the x-axis. The dashed line indicates the threshold for genome-wide significance of the gene-based test ($P < 2.76 \times 10^{-6}$; 0.05/18,128), and the dotted line indicates the suggestive threshold ($P < 2.76 \times 10^{-5}$; 0.5/18,128). **(d)** Gene expression profiles of identified genes for 53 tissue types. Expression data were extracted from the Genotype-Tissue Expression (GTEx) database. Expression values (RPKM) were $\log_2$ transformed with pseudocount 1 after winsorization at 50 and averaged per tissue. Gene-set tests for tissue expressions were calculated using MAGMA **(Online Methods)**. **(e)** Enrichment of genetic signal for neuroticism in 24 cell types derived from mouse brain. The dashed line indicates the Bonferroni-corrected significance threshold (P = 0.05/7,323 = 6.83×10^{-6}).

the 19 genes mapped by two loci, 4 are located outside of the risk loci (*HIST1 H2AI*, *HIST1H3H*, *HIST1H2AK* and *HIST1H4L*), and 7 are also implicated by eQTLs in several tissue types (HFE in adipose subcutaneous, aorta, esophagus muscularis, lung, tibial nerve, sub-exposed skin and thyroid; *HIST1H4J* in blood and adrenal gland; and *HIST1H4K*, *HIST1H2AK*, *HIST1H2BO* and *XKR6* in blood).

Gene-based P-values were used for gene-set analysis in MAGMA[25], testing 7,246 pre-defined gene sets derived from MsigDB[26], gene expression profiles in 53 tissue types obtained from the GTEx Project[27], and 24 cell type

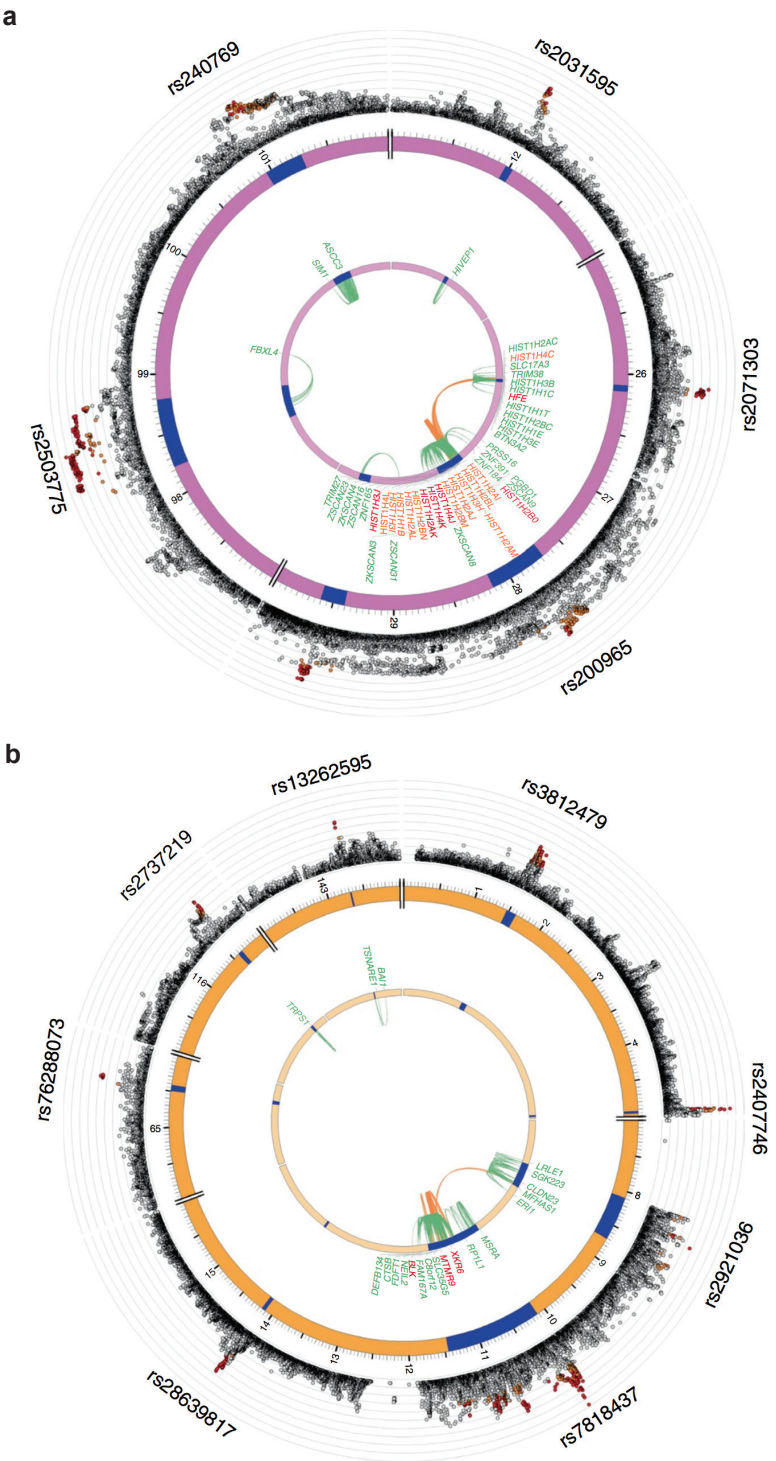


Fig. 3 | Genomic risk loci, eQTL associations and chromatin interaction for chromosome 6 and 8, containing cross-locus interactions. Circos plot showing genes on **(a)** chromosome 6 and **(b)** chromosome 8 that were implicated by the genomic risk (blue areas) loci by chromatin interaction (CTI; orange), eQTL (green) or implicated by both eQTL and CTI mapping (red). The outer layer shows a Manhattan plot containing the $-\log_{10}$ transformed *P*-value of each SNP in the GWAS meta-analysis of neuroticism (n = 449,484 individuals). Empty regions in the Manhattan plot layer indicate regions where no SNPs with $P < 0.05$ are situated.

specific expression profiles using RNAseq information[28] **(Online Methods)**. Neuroticism was significantly associated with genes predominantly expressed in 6 brain tissue types (**Fig. 2d**; **Supplementary Table 17-18)** and with 7 gene ontology (GO) gene sets, with the strongest association for neurogenesis ($P = 4.43 \times 10^{-9}$) and neuron differentiation ($P = 3.12 \times 10^{-8}$) (**Supplementary Table 17)**. Conditional gene-set analyses (**Online Methods**) suggested that 3 of the 7 gene sets (neurogenesis, $P = 4.43 \times 10^{-9}$; behavioral response to cocaine, $P = 1.84 \times 10^{-7}$; axon part, $P = 5.26 \times 10^{-8}$) had largely independent associations, implying a role in neuroticism (**Supplementary Table 19**). Conditional analyses of the tissue-specific expression ascertained general involvement of (frontal) cortex expressed genes (**Supplementary Table 20**; **Supplementary Fig. 5**).

Cell-type-specific gene-set analysis showed significant association with genes expressed in multiple mice-derived brain cell types **(Online Methods**; **Fig. 2e**; **Supplementary Table 21**), with dopaminergic neuroblasts ($P= 3.49 \times 10^{-8}$), medium spiny neurons ($P = 4.23 \times 10^{-8}$), and serotonergic neurons ($P = 1.37 \times 10^{-7}$) showing the strongest associations. Conditional analysis indicated that these three cell types were also independently associated with neuroticism.

Aiming to further specify neuroticism's neurobiological interpretation, we compared the genetic signal of the full neuroticism trait to that of two genetically distinguishable neuroticism subclusters *depressed affect* and *worry* (**Online Methods**), which we previously established through hierarchical clustering of the genetic correlations between the 12 neuroticism items[13]. As a validation of the *depressed affect* dimension, we also compared the genetic signal of neuroticism and the two subclusters to that of depression. GWA analyses of the subclusters were conducted on the UKB-data only (dictated by item-level data availability; **Online Methods**; *depressed affect*, *n* = 357,957; *worry*, *n* = 348,219). For depression, our meta-analysis comprised data from the UKB[14] (*n* = 362,696; **Supplementary Fig. 6**), 23andMe[15] (*n* = 307,354), and the Psychiatric Genetics Consortium (PGC[29]; *n* = 18,759) (total *n* = 688,809, not previously published, largest N for depression to date; r_g between samples: 0.61-0.80; **Online Methods**; **Supplementary Table 22,** see **Supplementary Note** for details on the depression GWAS results).
Genetic correlations of neuroticism with all three phenotypes were considerable (depression: $r_g = 0.79$; *depressed affect*: r_g=0.88, *worry*: $r_g = 0.87$; **Supplementary Table 23**). The positive genetic correlations between neuroticism and depression might in part be due to overlap in item content between the instruments used to gauge these phenotypes, reducing their operational distinctness[13].
The subclusters showed notable differences in genetic signal (e.g., exclusive GWS associations on chromosomes 2 and 19 for *depressed affect,* and chromosomes 3 and 22 for *worry*; **Supplementary Figs. 7-13**; **Supplementary Tables 24-26**). Of the 136 genetic loci associated with neuroticism, 32 were also GWS for *depressed affect* (7 shared with depression) but not for *worry*, and 26 were also GWS for *worry* (3 shared with depression) but not for *depressed affect* (**Supplementary Table 27**; **Supplementary Fig. 13**). These results were mirrored by gene-based analyses (**Supplementary Note**; **Supplementary Tables 28-30**; **Supplementary Fig. 14**), suggesting that part of neuroticism's genetic signal originates specifically in one of the two subclusters, possibly implicating different causal genetic mechanisms.
To further verify the biological distinctness of the two clusters, cluster-specific functional annotation was conducted, demonstrating that with respect to those SNPs that are highly likely to have functional consequences (ExNS), the clusters are 1) distinct and 2) adding information to the results of neuroticism sum-score analysis (**Supplementary Note**; **Supplementary Tables 31-34**; **Supplementary Fig. 15**).
To test whether the signal of the gene-sets implicated in neuroticism rather originated from one of the specific subclusters, we conducted conditional analyses correcting neuroticism for *depressed affect*, and *worry* scores, respectively (**Supplementary Table 35**; **Supplementary Fig. 16**). The association with 'axon-part' was markedly lower after correction for *worry* scores (uncorrected $P = 5.26 \times 10^{-8}$; corrected for *depressed affect* $P = 2.42 \times 10^{-6}$; corrected for *worry* $P = 0.0013$), suggesting that the involvement of 'axon-part' in neuroticism originates predominantly from the *worry*-component.
To examine the genetic correlational pattern of neuroticism, and to compare it to the patterns observed for depression, *depressed affect* and *worry*, we used LDSC[18,30] to calculate genetic correlations with 35 traits for which large-scale GWAS summary statistics were available (**Supplementary Table 36**; **Online Methods**). We observed 11 Bonferroni-corrected significant genetic correlations between neuroticism and other traits ($\alpha=0.05/(4 \times 35)$; $P < 3.6 \times 10^{-4}$) (**Fig. 4; Supplementary Table 37)**, covering previously reported psychiatric traits (r_g range: 0.20-0.82) and subjective well-being ($r_g = -0.68$). These correlations were supported by enrichment of neuroticism genes in sets of genes previously implicated in psychiatric traits (**Supplementary Table 38**). The r_g's of depression and *de-*

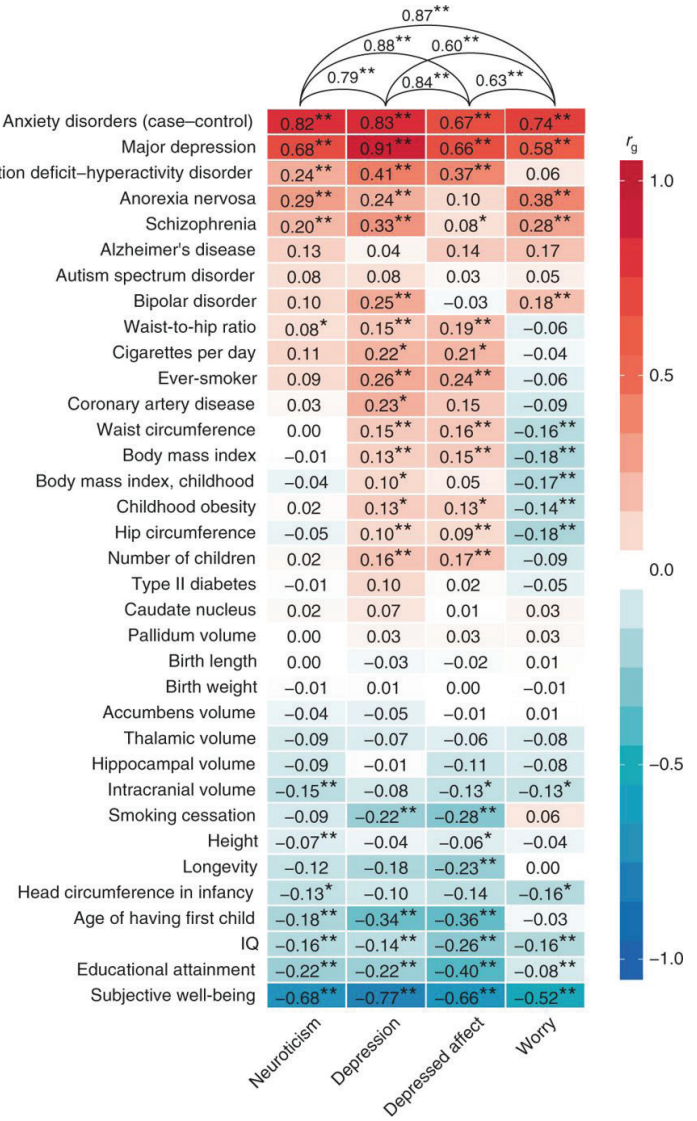


Fig. 4 | Genetic correlations between neuroticism and other traits. Genetic correlations of neuroticism, depression, *depressed* affect and *worry* with various traits and diseases. LD Score regression **(Online Methods)** tested genome-wide SNP associations for the neuroticism score against previously published results for 35 neuropsychiatric outcomes, antropometric and health-related traits, and brain morphology (**Supplementary Table 36-37**). Genetic correlations among neuroticism, depression, *depressed* affect and *worry* are displayed in the top part of this figure. Red and blue indicate positive and negative genetic correlations, whereas hue indicates the strength of the genetic correlations. Sample sizes for the traits in this figure are presented in **Supplementary Table 36**). *P<0.01; **Bonferroni corrected P-value threshold ($P < 3.6 \times 10^{-4}$).

pressed affect strongly mirrored eachother (correlation between their r_g's is $r = 0.98$; **Supplementary Note**), validating the *depressed affect* cluster. The correlational patterns for *depressed affect* and *worry* were markedly different (e.g., anorexia nervosa, schizophrenia, ever smoker) and sometimes in opposite directions (e.g., BMI). The genetic correlations of the full neuroticism trait appeared a mix of the genetic signal of both clusters, with neuroticism's r_g's generally falling in between the cluster-specific r_g's.

To investigate whether these genetic correlations reflect directional effects, we performed Mendelian randomization (MR) using the GSMR package[31] (**Online Methods**). Among things, we observed unidirectional effects of BMI on depression and *depressed affect* ($b_{xy} = 0.061$, $P = 4.96 \times 10^{-12}$ and $b_{xy} = 0.049$, $P = 5.35 \times 10^{-6}$, respectively), and bidirectional associations between neuroticism and depression, and between all four main traits and subjective well-being, cognition and several psychiatric disorders

(**Supplementary Table 39**, **Supplementary Note**).

We aimed to identify gene-drug interactions (DGIdb[32,33]; **Online Methods**), of genes identified for each of the four traits, and observed a large number of potential targets for pharmatherapeutic intervention that were either shared between traits or distinct for each phenotype (**Supplementary Note**; **Supplementary Tables 40-41**; **Supplementary Fig. 17**).

In conclusion, we identified 124 novel genetic loci for neuroticism (73 taking into account a simultaneously conducted study by Luciano et al.[34] (see **Supplementary Note**; **Supplementary Table 42**). Extensive functional annotations highlighted several genes being implicated through multiple routes. We demonstrated the involvement of specific neuronal cell types and three independently associated genetic pathways, and established the genetic multidimensionality of the neuroticism phenotype, and its link with depression. The current study provides new leads, and testable functional hypotheses for unraveling the neurobiology of neuroticism, its subtypes, and genetically associated traits.

Online Methods

Samples

UK Biobank: The UK Biobank (UKB) Study is a major data resource, containing genetic and a wide range of phenotypic data of ~500,000 participants aged 39-73 at recruitment[14]. We used data released in July 2017, and selection (discussed below) resulted in final sample sizes of n = 372,903 and n = 362,696 individuals for neuroticism and depression, respectively (**Supplementary Note)**. The UKB received ethical approval from the National Research Ethics Service Committee North West–Haydock (reference 11/NW/0382), and all study procedures were performed in accordance with the World Medical Association for medical research. The current study was conducted under UKB application number 16406.

23andMe: 23andMe, Inc. is a large personal genomics company that provides genotype and health-related information to customers. For the neuroticism and depression meta-analyses, we used neuroticism and depression GWAS summary statistics, respectively, from a subset of 23andMe research participants (neuroticism: n = 59,206; depression: n = 307,354), described in more detail elsewhere[10,35]. All included participants provided informed consent and were of European ancestry, and related individuals were excluded. Online data collection procedures were approved by the Ethical & Independent Review Services (E&I Review), an AAHRPP-accredited private institutional review board.

Genetics of Personality Consortium: The Genetics of Personality Consortium (GCP) is a large body of cooperation concerning GWAS on personality. We used summary statistics of neuroticism from the first GPC personality meta-analysis (GPC1)[9], on 10 discovery cohorts (SardiNIA, NTR/NESDA, ERF, SAGE, HBCS, NAG, IRPG, QIMR, LBC1936, BLSA, EGPUT), including in total n =17,375 participants of European descent. All included studies were approved by local ethic committees, and informed consent was obtained from all participants.

Psychiatric Genetics Consortium: The Psychiatric Genetics Consortium (PGC) unites investigators worldwide to conduct genetic meta- and mega-analyses for psychiatric disorders. We used summary statistics from the latest published PGC meta-analysis on depression[29], which included data from 8 cohorts (Bonn/Mannheim, GAIN, GenRED, GSK, MDD2000, MPIP, RADIANT, STAR*D), covering n = 18,759 participants of European descent. All included studies were approved by local ethic committees, and informed concent was obtained from all participants.

Phenotype assessment – Neuroticism

UK Biobank: Neuroticism was measured with 12 dichotomous (yes/no) items of the Eysenck Personality Questionnaire Revised Short Form (EPQ-RS[36], using a touchscreen-questionnaire at the UKB assessment centers (**Supplementary Note**). Participants with valid responses to <10 items were excluded from analyses. A weighted neuroticism sum-score was calculated by adding up individual valid item responses, and dividing that sum by the total number of valid responses. In addition, we constructed two scores based on subsets of genetically homogeneous neuroticism items, as established previously[13] through hierarchical clustering analysis of the genetic correlations between the 12 neuroticism items (see **Supplementary Note**). Specifically, scores on 4 EPQ-RS items (i.e., "Do you often feel lonely?", "Do you ever feel 'just miserable' for no reason?", "Does your mood often go up and down?", and "Do you often feel 'fed-up'?") were summed to obtain scores for the cluster *depressed affect*. Similarly, scores on 4 other EPQ-RS items (i.e., "Are you a worried?", "Do you suffer from nerves?", "Would you call yourself a nervous person?", and "Would you call yourself tense or highly strung") were summed to obtain scores for the cluster *worry*. In the item-cluster analyses, only participants with complete scores on all 4 items were included, resulting in n = 357,957 and n = 348,219 for *depressed affect* and *worry*, respectively.

23andMe: Neuroticism was operationalized as of the sum of 8 neuroticism items (5-point Likert scale; 'Disagree

strongly' to 'Agree strongly') from the Big Five Inventory (BFI[37,38]), as obtained in an online survey. Only participants with valid responses to all items were included in the analyses (**Supplementary Note**).

Genetic Personality Consortium: All 10 cohorts included in the first meta-analysis of the GPC used sums of the scores on 12 items (5-point Likert scale; 'Strongly disagree' to 'Strongly agree') of the NEO-FFI[39] to measure neuroticism. If <4 item scores were missing, data on invalid items were imputed by taking an individual's average score on valid items. Participants were excluded from analyses if they had invalid scores on >3 items[9] (**Supplementary Note**).

Phenotype assessment - Depression

UK Biobank: Depression was operationalized by adding up the scores on two continuous items ("Over the past two weeks, how often have you felt down, depressed or hopeless?", "Over the past two weeks, how often have you had little interest or pleasure in doing things?"; both evaluated on a 4-point Likert scale; 'Not at all' to 'Nearly every day'), resulting in a continuous depression score (as used previously[12]). Only participants with scores on both items were included in the analyses, resulting in n = 362,696 (**Supplementary Note**).

23andMe: This concerns a case-control sample. Four self-report survey items were used to determine case-control status. Cases were defined as replying affirmatively to at least one of these questions, and not replying negatively to previous ones. Controls replied negatively to at least one of the questions, and did not report being diagnosed with depression on previous ones (**Supplementary Note**).

Psychiatric Genetics Consortium: This concerns a case-control sample. Cases had a DSM-IV lifetime (sometimes (early onset) recurrent) major depressive disorder (MDD) diagnosis, either established through structured diagnostic interviews or clinician-administered DSM-IV checklists. Most cases were ascertained from clinical sources, while controls were randomly selected from population resources and screened for lifetime history of MDD[29] (**Supplementary Note**).

Genotyping and imputation

UK Biobank - Neuroticism: We used genotype data released by the UKB in July 2017. The genotype data collection and processing are described in detail by the responsible UKB group[14]. In short, 489,212 individuals were genotyped on two customized SNP arrays (the UK BiLEVE Axiom array (n = 50,520) and UK Biobank Axiom array (n = 438,692)), covering 812,428 unique genetic markers (95% overlap in SNP content). After quality control procedures[14], 488,377 individuals and 805,426 genotypes remained. Genotypes were phased and imputed by the coordinating team to approximately 96 million genotypes using a combined refence panel including the Haplotype Reference Consortium and the UK10K haplotype panel. Imputed and quality controlled genotype data was available for 487,422 individuals and 92,693,895 genetic variants. As recommended by the UKB team, variants imputed from the UK10K reference panel were removed from the analyses due to technical errors in the imputation process.

In our analyses, only individuals from European descent (based on genetic principal components) were included. Therefore, principal components from the 1000 Genomes reference populations[40] were projected onto the called genotypes available in UK Biobank. Subjects were identified as European if their projected principal component score was closest (based on the Mahalanobis distance) to the average score of the European 1000 Genomes sample[41]. European subjects with a Mahalanobis distance > 6 S.D. were excluded. In addition, participants were excluded based on withdrawn consent, UKB provided relatedness (subjects with most inferred relatives, 3rd degree or closer, were removed until no related subjects were present), discordant sex, sex aneuploidy. After selecting individuals based on available neuroticism sum-score and active consent for participation, 372,903 individuals remained for the analyses.

To correct for population-stratification, 30 principal components were calculated on the subset of QC-ed unrelated European subjects based on 145,432 independent (r^2<0.1) SNPs with MAF>0.01 and INFO=1 using FlashPC2[42]. Subsequently, imputed variants were converted to hard call using a certainty threshold of 0.9. Multi-allelic SNPs, indels, and SNPs without unique rs id were excluded, as well as SNPs with a low imputation score (INFO score <0.9), low minor allele frequency (MAF<0.0001) and high missingness (>0.05). This resulted in a total of 10,847,151 SNPs used for downstream analysis.

UK Biobank - Depression: Similar genotyping / imputation/filtering procedures as described above for the UKB neuroticism GWAS were followed for the UKB depression GWAS, resulting in n = 362,696.

Genome-wide association analyses

UK Biobank - Neuroticism: Genome-wide association analyses were performed in PLINK[43,44], using a linear regression model of additive allelic effects with age, sex, townsend deprivation index, genotype array, and 10 genetic European-based principal components as covariates (**Supplementary Note**).

UK Biobank - Depression, depressed affect, worry: The settings, covariates, and exclusion criteria for the UKB depression, UKB *depressed affect*, and UKB *worry* GWAS were the same as described above for UKB neuroticism GWAS, with 10,847,151 SNPs remaining after all exclusion steps (**Supplementary Note**).

Other samples: Summary statistics were used for 23andMe, GPC and PGC. Details on the genome-wide association analyses of these samples can be found elsewhere (23andMe neuroticism[10]; 23andMe depression[35]; GPC neuroticism[9]; PGC depression[29]).

Meta-analysis

In order to maximize the statistical power to detect associated genetic variants of small effect, we conducted meta-analyses for both neuroticism and depression[17] (**Supplementary Note**). All meta-analyses were carried out in METAL[16].

Neuroticism: The meta-analysis of the neuroticism GWAS in UKB, 23andMe, and GPC was performed on the *P*-value of each SNP using a sample size-weighted fixed-effects analysis. Bonferroni correction was applied to correct for multiple testing. The genetic signal correlated strongly between the three samples (r_g range: 0.83 – 1.07; **Supplementary Table 1**), supporting the decision to meta-analyze.

Depression: As the UKB GWAS concerned a continuous operationalization of the depression phenotype, while 23andMe and PGC used case-control phenotypes, the odds ratio from the 23andMe and PGC summary statistics were converted to log odds, reflecting the direction of the effect. The meta-analysis was then performed on the *P*-value of each SNP using a sample size weighted fixed-effects analysis. Bonferroni correction was applied to correct for multiple testing. Genetic correlations between the three samples were moderate to strong (r_g range: 0.61 – 0.80; **Supplementary Table 22**).

Genomic risk loci and functional annotation

Functional annotation was performed using FUMA[17], an online platform for functional mapping of genetic variants. We first defined *independent significant SNPs* which have a genome-wide significant *P*-value (5×10^{-8}) and are independent at $r^2<0.6$. A subset of these *independent significant SNPs*, that were independent from each other at $r^2<0.1$, was marked as lead SNPs (based on LD information from UK Biobank genotypes; see **Supplementary Note** for a more detailed explanation). Subsequently, genomic risk loci were defined by merging lead SNPs that physically overlapped or for which LD blocks were less than 250 kb apart. Note that when analyzing multiple phenotypes, as in the current study, a locus may be discovered for different phenotypes, whilst different lead SNPs are identified.

All SNPs in the meta-analysis results that were in LD ($r^2 > 0.6$) with one of the *independent significant SNPs*, had a *P*-value lower than 1.0×10^{-5} and minor allele frequency (MAF) > 0.0001 were selected for annotation. The rationale behind this inclusive approach is that the most significant SNP in the locus is not necessarily the causal SNP, but may be in LD with the causal SNP. We thus annotated all SNPs in LD with the most significant SNP to get insight into the possible biological reasons for observing a statistical association. We note that liberalizing the r^2 and *P*-value thresholds can dilute the functional annotation results, while more stringent thresholds may result in exclusion of possibly interesting functional variants. Functional consequences for these SNPs were obtained by performing ANNOVAR[45] gene-based annotation using Ensembl genes. In addition, CADD scores (indicating the deleteriousness of a SNP, with scores >12.37 seen as likely deleterious[21]) and RegulomeDB scores[46] (where a higher probability of having a regulatory function is indicated by lower scores) were annotated to SNPs by matching chromosome, position, reference and alternative alleles. CADD scores integrate a number of diverse annotations into a single measure that correlates with pathogenicity, disease severity and experimentally measured regulatory effects and complex trait associations[21].

Gene-mapping

SNPs in genomic risk loci that were GWS or were in LD ($r^2 > 0.6$) with one of the *independent significant SNPs* were mapped to genes in FUMA[20] using either of three strategies.

First, positional mapping uses the physical distances (i.e., within 10kb window) from known protein coding genes in the human reference assembly (GRCh37/hg19) to map SNPs to genes. The second strategy, eQTL mapping, uses information from 3 data repositories (GTEx, Blood eQTL browser BIOS QTL browser), and maps SNPs to genes based on a significant eQTL association (i.e. the expression of that gene is associated with allelic variation at the SNP). eQTL mapping is based on cis-eQTLs which can map SNPs to genes up to 1Mb apart. FUMA applied a false discovery rate (FDR) of 0.05 to define significant eQTL associations. Thirdly, chromatin interaction mapping mapped SNPs to genes based on a significant chromatin interaction between a genomic region in a risk locus and promoter regions of genes (250bp up- and 500bp downstream of transcription start site (TSS)). This type of mapping does not have a distance boundary (as in eQTL

mapping), and may therefore involve long-range interactions. Currently, FUMA contains Hi-C data of 14 tissue types from the study of Schmitt et al. (2016)[47]. Importantly, as chromatin interactions are usually defined in a certain resolution (in the current study; 40kb), an interacting region may span several genes. Hence, this method would map all SNPs within these regions to genes in the corresponding interaction region. By integrating predicted enhancers and promoters in 111 tissue/cell types from the Roadmap Epigenomics Project[48] we aimed to prioritize candidate genes from chromatin interaction mapping. Using this information FUMA selected chromatin interactions for which one region involved in the interaction overlaps with predicted enhancers and the other with predicted promoters in 250bp up- and 500bp downstream of TSS site of a gene. Like with the eQTL mapping, we used a FDR of 1×10^{-5} to define significant interactions.

Gene-based analysis

A genome-wide gene-based association analysis (GWGAS) can identify genes in which multiple SNPs show moderate association to the phenotype of interest without reaching the stringent genome-wide significance level. At the same time, as a GWGAS takes all SNPs within a gene into account, a gene harbouring a genome-wide significant SNP may not be implicated by a GWGAS analyses when multiple other SNPs within that gene show only very weak association signal. The *P*-values from the SNP-based GWAS meta-analyses for neuroticism and depression, and the GWAS for *depressed affect* and *worry*, were used as input for the genome-wide gene-based association analysis (GWGAS) in MAGMA[25], and all 19,427 protein-coding genes from the NCBI 37.3 gene definitions were used. We annotated all SNPs in our GWA (meta-) analyses to these genes, resulting in 18,187, 18,187, 18,182, and 18,182 genes that were represented by at least one SNP in the neuroticism meta-analysis, the depression meta-analysis, the *depressed affect* GWAS, and the *worry* GWAS, respectively. We included a window around each gene of 2 kb before the transcription start site and 1 kb after the transcription stop site. Gene association tests were performed taking into account the LD between SNPs, and a stringent Bonferroni correction was applied to correct for multiple testing (0.05/number of genes tested: $P < 2.75 \times 10^{-6}$).

Gene-set analysis

We used MAGMA[25] to test for association of predefined gene sets with neuroticism, depression, *depressed affect*, and *worry*. A total of 7,246 gene sets were derived from several resources, including BioCarta, KEGG, Reactome[49] and GO. All gene sets were obtained from the MsigDB version 6.0). Additionally, we performed gene-set analysis on 53 tissue expression profiles obtained from the GTEx portal, and 24 cell-type specific expression profiles.

For all gene sets, we computed competitive *P*-values, which result from testing whether the combined effect of genes in a gene set is significantly larger than the combined effect of a same number of randomly selected genes (in contrast to testing against the null hypothesis of no effect; self-contained test). Here, we only report Bonferroni corrected ($a=0.05/7{,}323= 6.83 \times 10^{-6}$) competitive *P*-values, which are more conservative compared to self-contained *P*-values.

Cell type specific expression analysis

Definition and calculation of gene sets for cell type specific expression is described in detail elsewhere[28,50]. Briefly, brain cell type expression data was drawn from scRNA-seq data from mouse brain[28]. For each gene, the value for each cell type was calculated by dividing the mean Unique Molecular Identifier (UMI) counts for the given cell type by the summed mean UMI counts across all cell types[28]. MAGMA[25] was used to calculate associations between gene-wise *P*-values from the meta-analysis and cell type specific gene expression. Genes were grouped into 40 equal bins by specificity of expression, and subsequently bin-membership was regressed on gene-wise association with neuroticism in the meta-analysis. Results were deemed significant if the association *P*-values were smaller than the relevant Bonferroni threshold.

Conditional gene-set analyses

Conditional gene-set analyses were performed using MAGMA[25] to determine which tissue expression levels and MsigDB gene-sets represent independent associations. In these regression-based analyses, the effect of a gene-set (or tissue expression) of interest is conditioned on the effects of another gene-sets (or tissue expressions) to correct the association of the tested gene-set for any effect it shares with the conditioned-on gene-sets.

For the MSigDB gene-sets we conducted two series of conditional analyses. First, we performed a forward selection on the initially significant gene-sets, in each step selecting the most strongly associated gene-set after conditioning on all already selected gene-sets. (**Supplementary Table 19**). Second, to test whether the association of gene-sets to neuroticism is primarily driven by association signal of one specific subcluster, we also reran the GO gene-set analyses conditioning on the gene *Z*-scores of *depressed affect* or *worry*, respectively (**Supplementary Table 35**). If the gene-set association decreases after conditioning on one cluster but not, or less so, when conditioning on the

other, then this suggests that neuroticism's association to that gene-set is primarily driven by the genetic effects of the first, and not the second, item cluster.

Genetic correlations

Genetic correlations (r_g) were computed using LD Score regression[18,30]. The significance of the genetic correlations of neuroticism, depression, *depressed affect* and *worry* with 35 behavioral, social and (mental) health phenotypes for which summary statistics were available was determined by correcting for multiple testing through a stringent Bonferroni corrected threshold of $P < 0.05 / (4\times35) = 3.6 \times 10^{-4}$).

Mendelian randomization analyses

We performed Mendelian randomization (MR) to test whether genetic correlations could be explained by directional effects between traits. Generalised summary-data based Mendelian randomization (GSMR[31]) was used for MR analysis: a summary statistics-based MR method that uses independent genome-wide significant variants as instrumental variables. Causal associations were tested between the four traits, and the 21 traits that showed significant genetic correlations (r_g) in LD Score regression analysis with at least one of the four traits. To test for uni- and bidirectional effects, we performed both forward and reverse GSMR analysis (i.e., using the four GWAS traits either as predictor or as outcome). Associations were Bonferroni corrected for multiple testing $P < 0.05/(21*4*2) = 2.98 \times 10^{-4}$).

Partitioned heritability

To investigate the relative contribution to the overall SNP-based heritability annotated to 22 specific genomic categories, we partioned SNP heritability by binary annotations using stratified LD Score regression[22]. Information about binary SNP annotations were obtained from the LD Score website. Enrichment results reflect the X-fold increase in h^2 proportional to the number of SNPs (e.g., enrichment = 13.79 for SNPs in conserved regions implies that a 13,79-fold increase in h^2 is carried by SNPs in these region, corrected for the proportion of SNPs in these regions compared to all tested SNPs).

Gene drug targets

We aimed to identify potential druggable targets by performing lookup of implicated genes (by one of the gene-mapping strategies) in the drug-gene interaction database (DGIdb[32,33], version 3.0). The DGIdb database contains mined data from several resources, and provides a comprehensive overview of the druggability of gene targets. First, we searched 20 drug-gene databases for interactions with existing medicines based on 48 known interaction types with genes that were implicated in each of the four phenotypes. Filtering was performed based on known interaction types, and interactions with FDA-approved pharmaceutical compounds. Second, to identify genes that may form targets for novel therapies in addition to existing medicines, we searched for potential gene druggability of gene targets and performed an additional search in 10 DGIdb databases containing information about gene targetability.

Polygenic risk scoring

To test the predictive accuracy (ΔR^2) of the our meta-analytic results for neuroticism, we calculated a polygenic risk score (PGS) based on the SNP effect sizes of the current analysis. As independent samples we used 3 holdout samples; we removed 3,000 individuals from the discovery sample (UKB only, as we only had access to individual-level data from this sample) and reran the genome-wide analyses. We repeated this three times, to create 3 randomly drawn, independent hold-out samples. Next, we calculated a PGS on the individuals in each of the 3 holdout samples. PGS were calculated using LDpred[24] and PRSice[23] (clumping followed by *P*-value thresholding).

For LDpred, PGS were calculated based on different LDpred priors (P_{LDpred} = 0.01, 0.05, 0.1, 0.5, 1 and infinitesimal). The explained variance (R^2) was derived from the linear model, using the neuroticism summary score as the outcome, while correcting for age, gender, array, townsend deprivation index and genetic principal components.

References

1. Kendler, K. S. & Myers, J. The genetic and environmental relationship between major depression and the five-factor model of personality. *Psychol. Med.* **40**, 801–806 (2010).
2. Middeldorp, C. M. *et al*. The association of personality with anxious and depressive psychopathology. *Biol. Personal. Individ. Differ.* 251–272 (2006).
3. Hettema, J. M., Neale, M. C., Myers, J. M., Prescott, C. A. & Kendler, K. S. A population-based twin study of the relationship between neuroticism and internalizing disorders. *Am. J. Psychiatry* **163**, 857–864 (2006).
4. Hayes, J. F. *et al.* Association of late adolescent personality with risk for subsequent serious mental illness among men in a Swedish nationwide cohort study. *JAMA Psychiatry* **54**, 948–963 (2017).
5. Smeland, O. B. *et al*. Identification of genetic loci shared between schizophrenia and the Big Five personality traits. *Sci. Rep.* **7**, 1–9 (2017).
6. Van Os, J. & Jones, P. B. Neuroticism as a risk factor for schizophrenia. *Psychol. Med*. **31**, 1129–1134 (2001).
7. Genetics of Personality Consortium. Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA psychiatry* **72**,

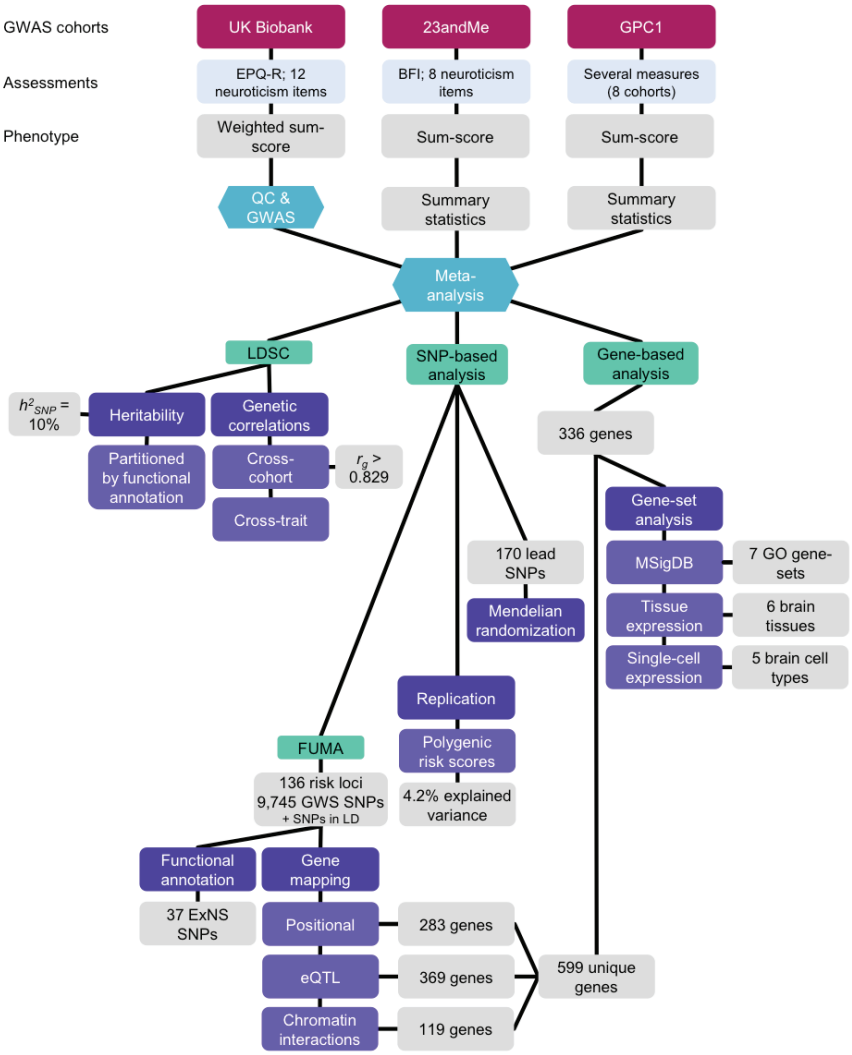
642–650 (2015).
8. Terracciano, A. *et al*. Genome-wide association scan for five major dimensions of personality. *Mol. Psychiatry* **15**, 647–656 (2010).
9. Moor, M. H. M. De *et al*. Meta-analysis of genome-wide association studies for personality. *Mol. Psychiatry* **17**, 337–349 (2012).
10. Lo, M. *et al*. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2016).
11. Smith, D. J. *et al*. Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Mol. Psychiatry* **21**, 1–9 (2016).
12. Okbay, A. *et al*. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–636 (2016).
13. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & Van der Sluis, S. Item-level Analyses Reveal Genetic Heterogeneity in Neuroticism. *Nat. Commun.* **9**, 905 (2018).
14. Bycroft, C. *et al*. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).
15. Eriksson, N. *et al*. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, 1–20 (2010).
16. Willer, C. J., Li, Y., Abecasis, G. R. & Overall, P. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
17. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
18. Bulik-Sullivan, B. K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
19. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
20. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
21. Kircher, M. *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
22. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
23. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, 1466–1468 (2015).
24. Vilhjálmsson, B. J. *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
25. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, 1–19 (2015).
26. Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
27. GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
28. Skene, N. G. *et al*. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **58**, 825–833 (2018).
29. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
30. Bulik-Sullivan, B. K. *et al*. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1–9 (2015).
31. Zhu, Z. *et al*. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
32. Griffith, M. *et al*. DGIdb: mining the druggable genome. *Nat. Methods* **10**, 1209–1210 (2013).
33. Cotto, K. C. *et al*. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.* **46**, D1068-1073 (2017).
34. Luciano, M. *et al*. 116 independent genetic variants influence the neuroticism personality trait in over 329,000 UK Biobank individuals. *Nat. Genet.* **50**, 6–11 (2017).
35. Hyde, C. L. *et al*. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
36. Eysenck, B. G., Eysenck, H. J. & Barrett, P. A revised version of the psychoticism scale. *Pers. Individ. Dif.* **6**, 21–29 (1985).
37. John, O. P. & Srivastava, S. The Big Five trait taxonomy: history, measurement, and theoretical perspectives. *Handb. Personal. Theory Res.* **2**, 102–138 (1999).
38. Soto, C. J. & John, O. P. Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *J. Res. Pers.* **43**, 84–90 (2009).
39. Costa, P. & McCrae. Professional Manual: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor-Inventory (NEO-FFI). (Psychological Assessment Resources: Odessa, FL, 1992).
40. Auton, A. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
41. Webb, B. T. *et al*. Molecular genetic influences on normative and problematic alcohol use in a population-based sample of college students. *Front. Genet.* **8**, 1–11 (2017).
42. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, 1–5 (2014).
43. 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).
44. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1 (2015).
45. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
46. Boyle, A. P. *et al*. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
47. Schmitt, A. D. *et al*. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
48. Roadmap Epigenomics Consortium *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
49. Croft, D. *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).
50. Coleman, J. R. I. *et al*. Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87,740 individuals. *Mol. Psychiatry* **24**, 182–197 (2019).

Supplementary information

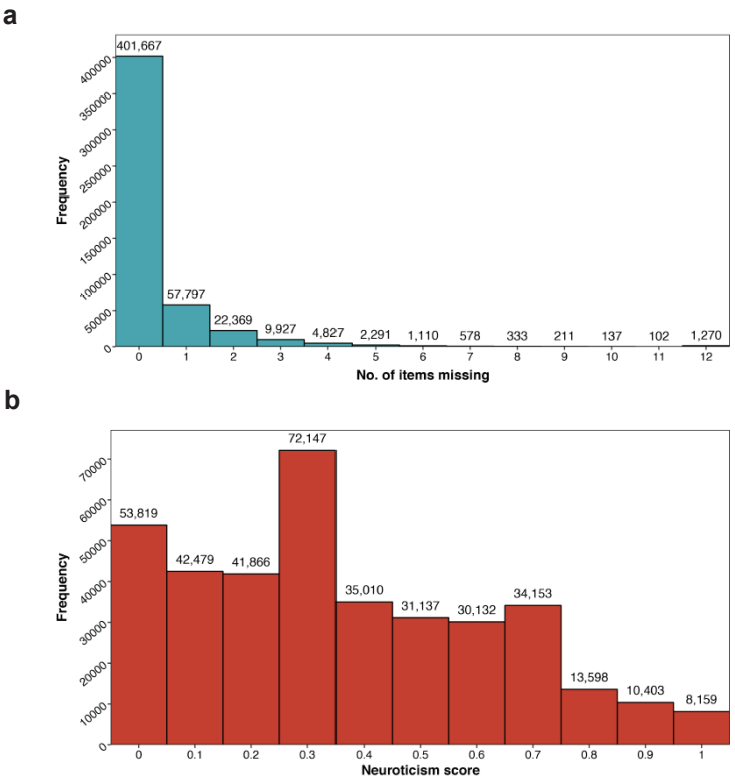
Supplementary information, figures (1-17) and tables (1-42) are in the online version of the manuscript:
https://www.nature.com/articles/s41588-018-0151-7

Supplementary information

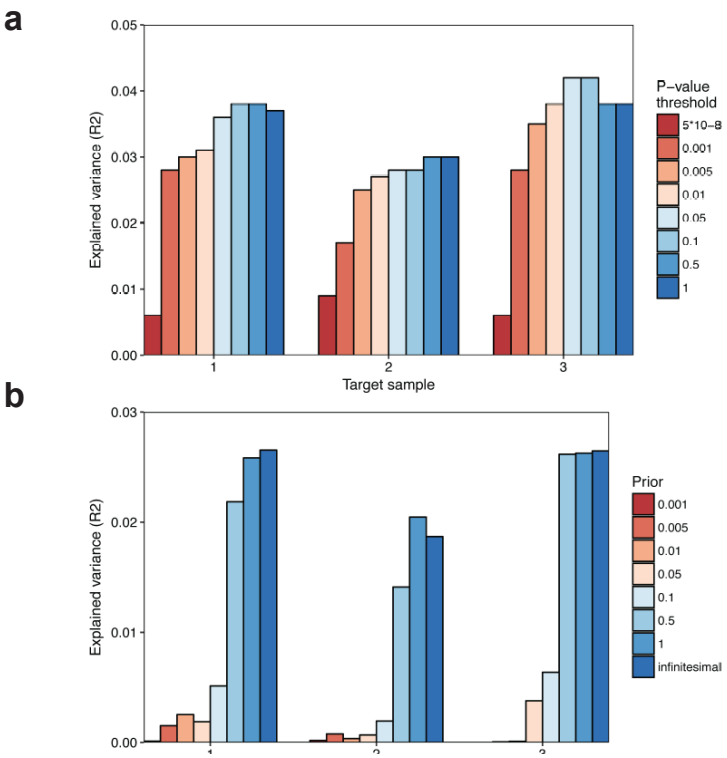
Supplementary Fig. 1 | Flow chart of analyses conducted in the study of neuroticism. Schematic representation of the analyses conducted for neuroticism. Additional analyses performed on the depression, *depressed affect*, and *worry* phenotypes are not included in this chart.



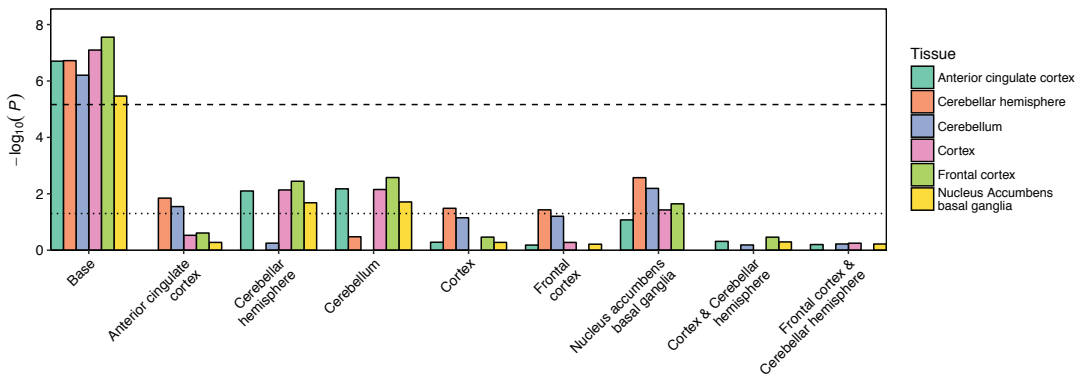
Supplementary Fig. 2 | Distribution of missingness and weighted sum-score for neuroticism in the UK biobank sample. (a) Distribution of the number of items missing. Individuals with invalid responses to more than two items missing were excluded from further analysis. **(b)** For the remaining participants, the neuroticism sum-score was established by summing the individual item responses and dividing by the total number of completed items for that participant. The distributions of these scores is shown in panel **b**.



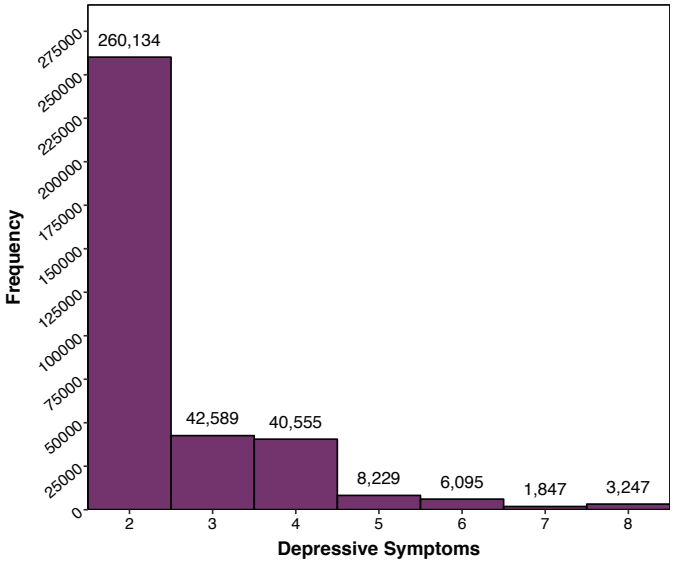
Supplementary Fig. 4 | Explained variance in neuroticism by polygenic risk scores (PGS) in 3 holdout samples (n = 3,000). To estimate the variation in neuroticism that could be explained by our GWA meta-analysis in independent samples, we re-ran out GWA meta-analysis for neuroticism three times, each time excluding a UKB hold-out sample of n = 3,000 randomly drawn individuals (n = 449,484-3,000 = 446,484). We then calculated polygenic scores (PGS) using two methods: **(a)** PRSice[23] (P-value thresholding and clumping) and **(b)** LDpred[24]. See **Supplementary Table 14**.



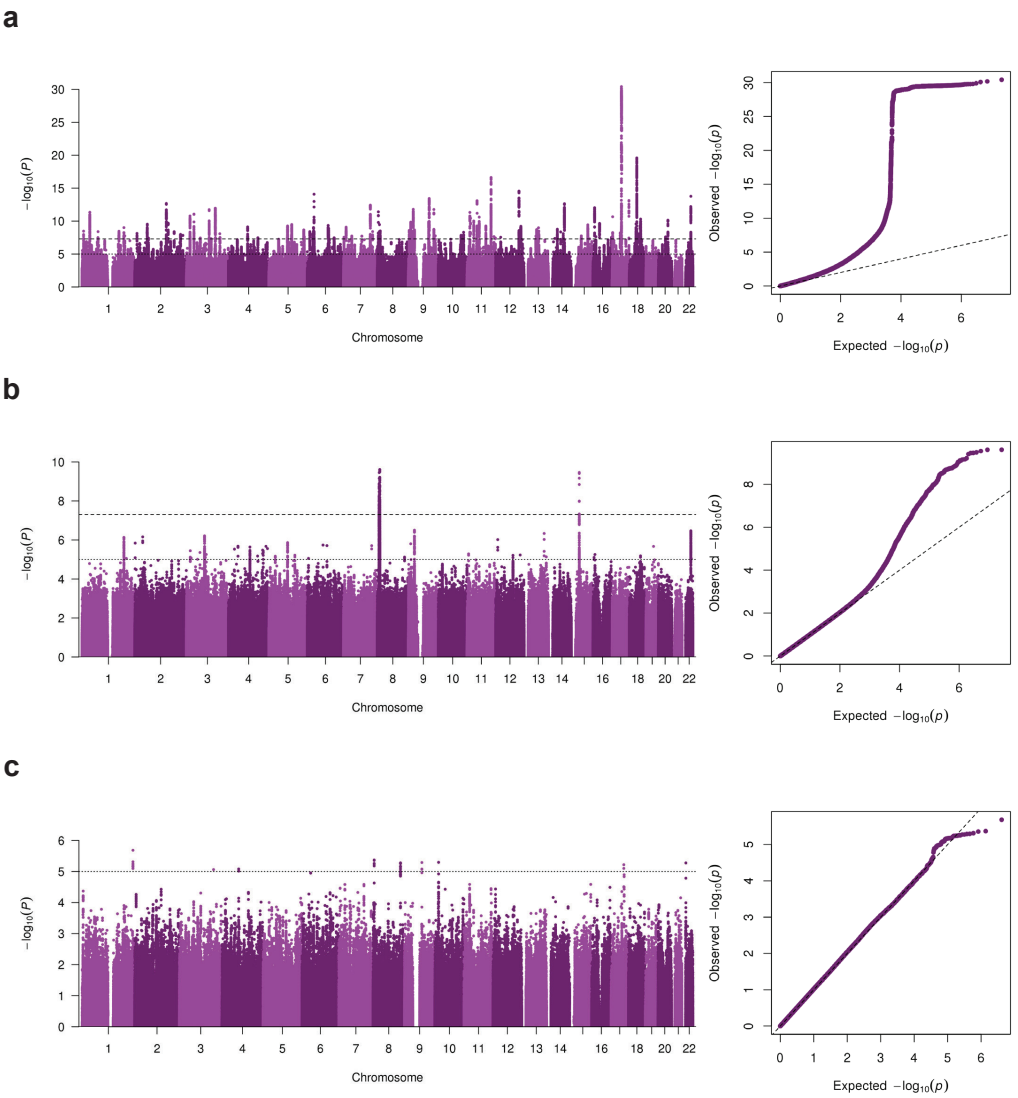
Supplementary Fig. 5 | Conditional gene-set analyses on 6 significant brain tissues. Conditional gene-set analysis was conducted in MAGMA, using the gene-based analysis results as input. The dotted line indicates nominal significance at a = 0.05; the dashed line indicates the Bonferroni corrected significance threshold (a=0.05/7,299=6.85 × 10^{-6}) used in the initial gene expression analysis (base).



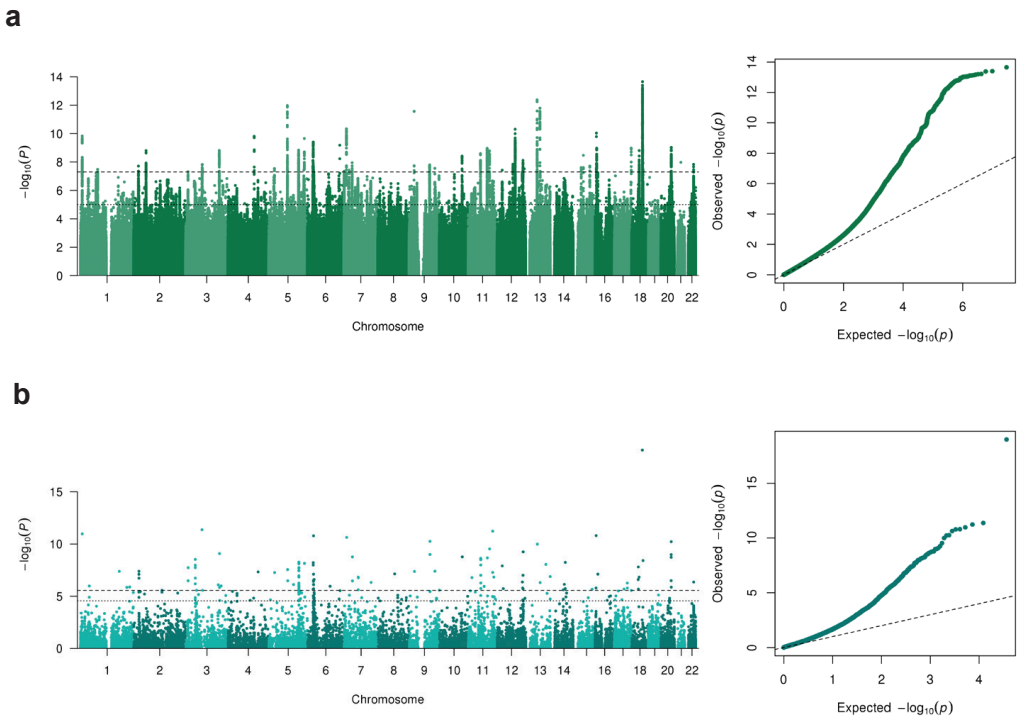
Supplementary Fig. 6 | Distribution of the depressive symptoms score in the UK Biobank sample. The distribution of the continuous depressive symptoms score that we calculated in the UKB sample. Depression was operationalized by adding up the scores on two continuous items (both evaluated on a 4-point Likert scale), resulting in a continuous depression score. Individuals that did not complete both items were excluded from further analysis. For the remaining participants, the depressive symptoms score was established by summing the individual item responses.



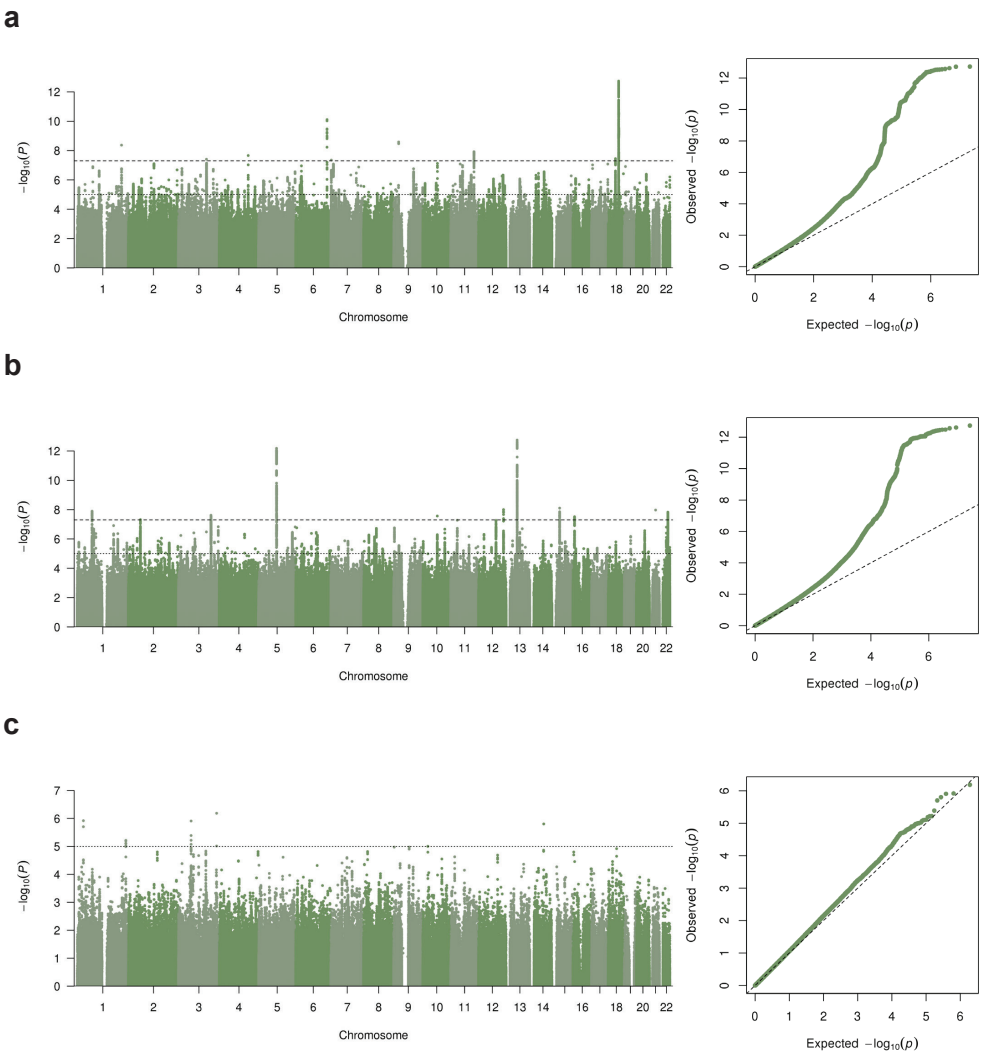
Supplementary Fig. 7 | Manhattan and Q-Q plots of SNP-based association with neuroticism in the individual cohorts. SNP Association results from the GWAS on neuroticism in **(a)** UK Biobank (n = 372,903 individuals) **(b)** 23andMe (n = 59,206 individuals) and **(c)** GPC1 (n = 17,375 individuals) cohorts. P-values for the UKB cohort were computed using a linear regression model of additive allelic effects and covariates in PLINK (see **Online Methods** for covariates and more information on the 23andMe and GPC1 cohorts). Dashed lines indicate genome-wide significance ($P<5\times10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$).



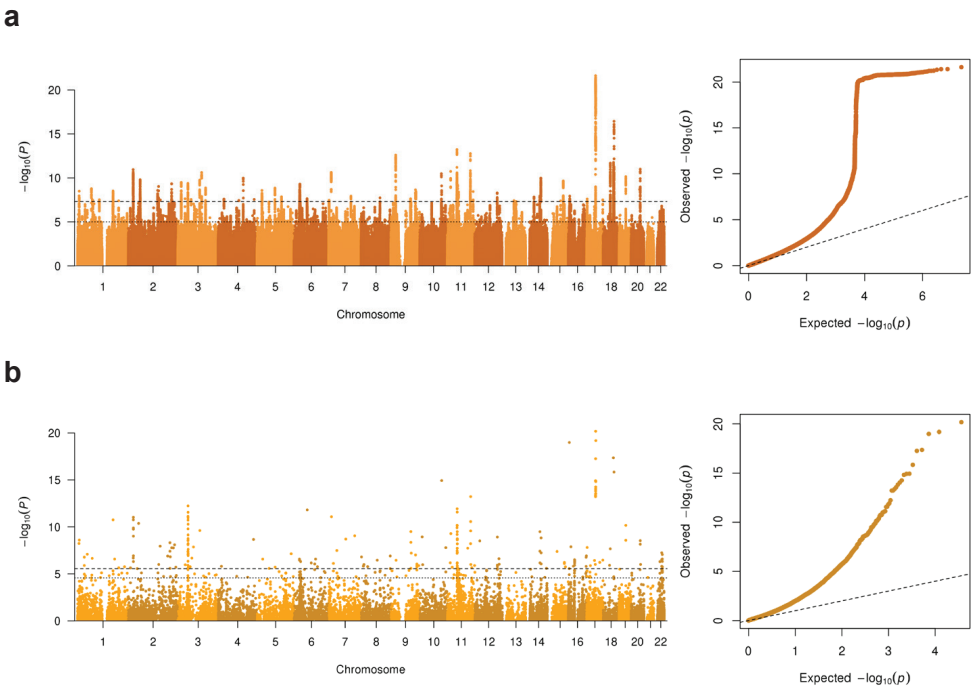
Supplementary Fig. 8 | Manhattan and Q-Q plots of SNP- and gene-based association with depression (n = 688,809 individuals). **(a)** SNP-based association results from the GWA meta-analysis (UK biobank, 23andMe and PGC) on depression. SNP *P*-values were computed in METAL using a two-sided, weighted z-score method. Dashed lines indicate genome-wide significance ($P < 5 \times 10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$). An overview of these genome-wide significant (GWS) SNPs and their *P*-values can be found in **Supplementary Table 24**. **(b)** Gene-based association results from the meta-analysis (UK biobank, 23andMe and PGC) on depression. Gene-based *P*-values were computed using MAGMA's gene-based test (where the summary statistics from GWAS were used as input). Dashed lines indicate genome-wide significance ($P < 2.75 \times 10^{-6}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 2.75 \times 10^{-5}$). An overview of these GWS genes and their *P*-values can be found in **Supplementary Tables 16** and **28**.



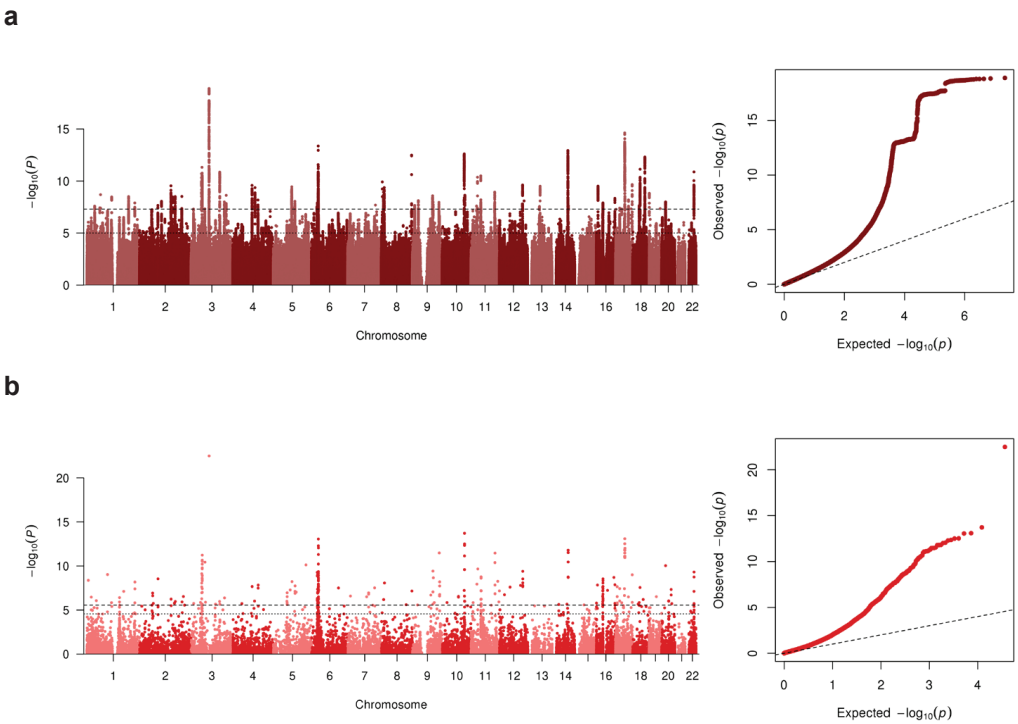
Supplementary Fig. 9 | Manhattan and Q-Q plots of SNP-based association with depression in the individual cohorts. SNP Association results from the GWAS on depression in **(a)** UK Biobank (n = 362,696 individuals) **(b)** 23andMe (n = 307,354 individuals) and **(c)** PGC (n = 18,759 individuals) cohorts. P-values for the UKB cohort were computed using a linear regression model of additive allelic effects and covariates in PLINK (see **Online Methods** for covariates and more information on the 23andMe and PGC cohorts). Dashed lines indicate genome-wide significance ($P < 5 \times 10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$).



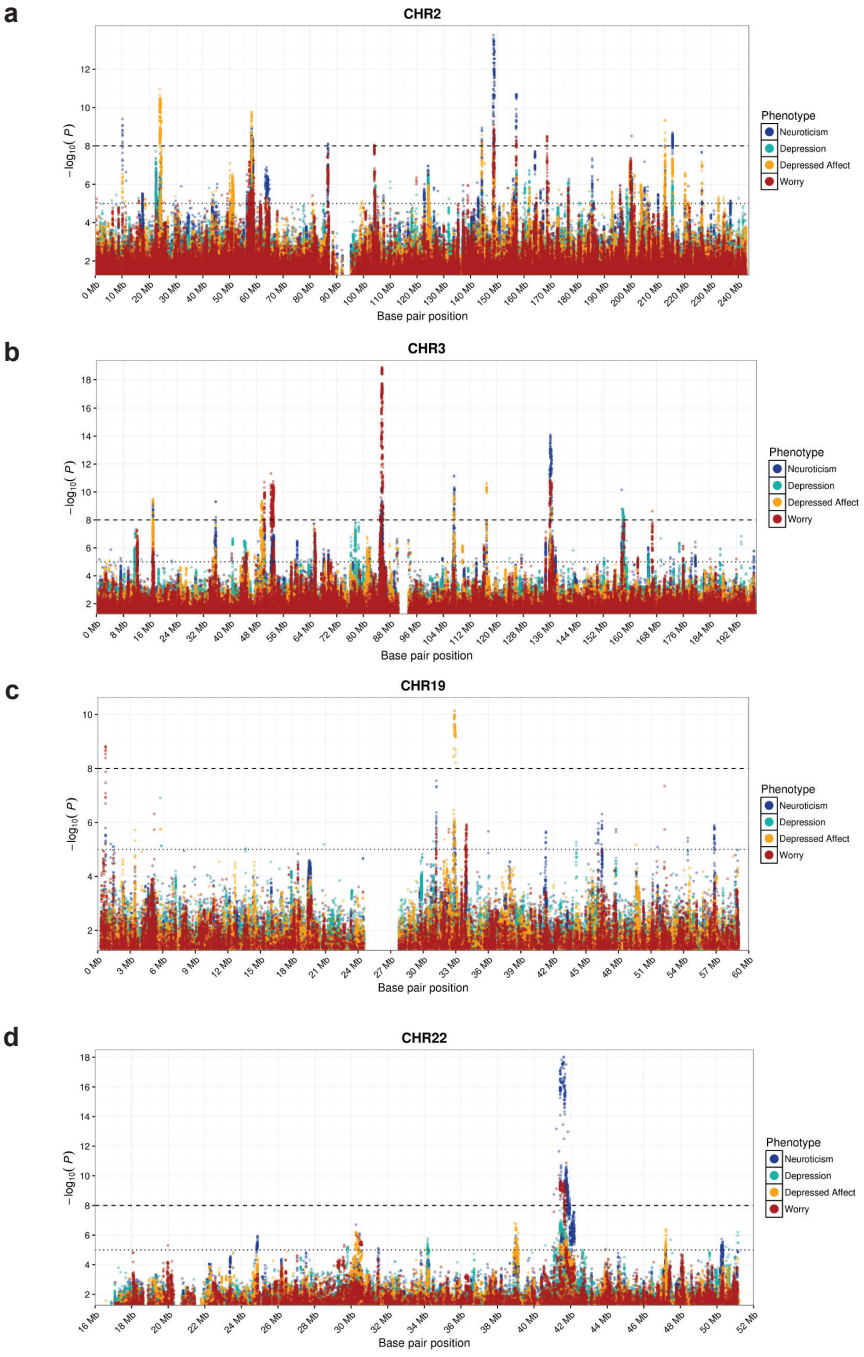
Supplementary Fig. 10 | Manhattan and Q-Q plots of SNP- and gene-based association with depressed affect (n = 357,957 individuals). (a) SNP-based association results from the GWAS on *depressed affect* in the UK biobank sample SNP *P*-values were computed using a linear regression model of additive allelic effects and covariates in PLINK. Dashed lines indicate genome-wide significance ($P < 5 \times 10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$). **(b)** Gene-based association results from the GWAS on *depressed affect* in the UK Biobank sample. Gene-based *P*-values were computed in MAGMA using a two-sided, weighted z-score method. Dashed lines indicate genome-wide significance ($P < 2.75 \times 10^{-6}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 2.75 \times 10^{-5}$).



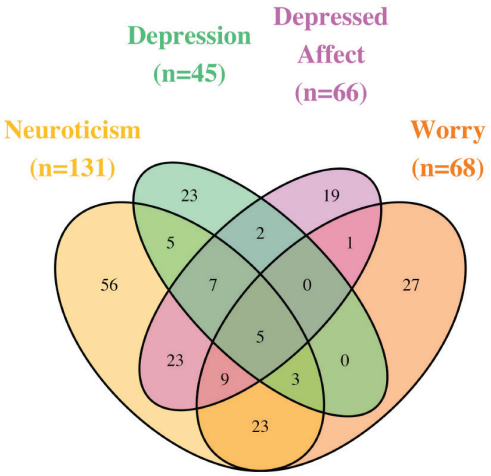
Supplementary Fig. 11 | Manhattan and Q-Q plots of SNP- and gene-based association with worry (n = 348,219 individuals), (a) SNP-based association results from the GWAS on *worry* in the UK biobank sample. SNP *P*-values were computed using a linear regression model of additive allelic effects and covariates in PLINK. Dashed lines indicate genome-wide significance ($P < 5 \times 10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$). **(b)** Gene-based association results from the GWAS on *worry* in the UK biobank sample. Gene-based *P*-values were computed in MAGMA using a two-sided, weighted z-score method. Dashed lines indicate genome-wide significance ($P < 2.75 \times 10^{-6}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 2.75 \times 10^{-5}$).



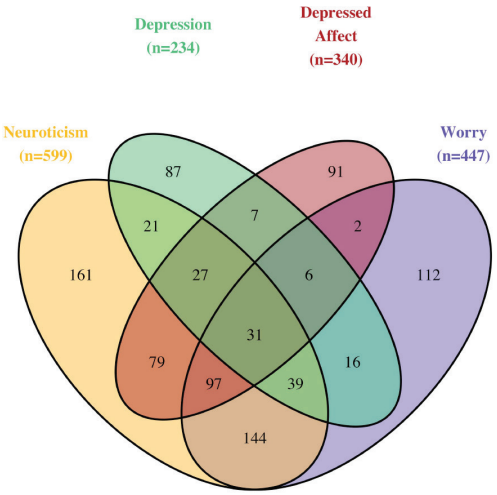
Supplementary Fig. 12 | Manhattan plots showing SNP associations for neuroticism, depression, depressed affect and worry on 4 chromosomes. Color-coded SNP-based association results of all 4 phenotypes plotted in the same plot. SNP *P*-values were computed in METAL using a two-sided, weighted z-score method (neuroticism and depression meta-analyses) or linear regression in PLINK (clusters). Dashed lines indicate genome-wide significance ($P < 5 \times 10^{-8}$) and dotted lines indicate the 'suggestive' significance threshold ($P < 1 \times 10^{-5}$). significance threshold ($P < 2.75 \times 10^{-5}$). SNP-based association results on **(a)** chromosome 2, **(b)** chromosome 3, **(c)** chromosome 19 and **(d)** chromosome 22.



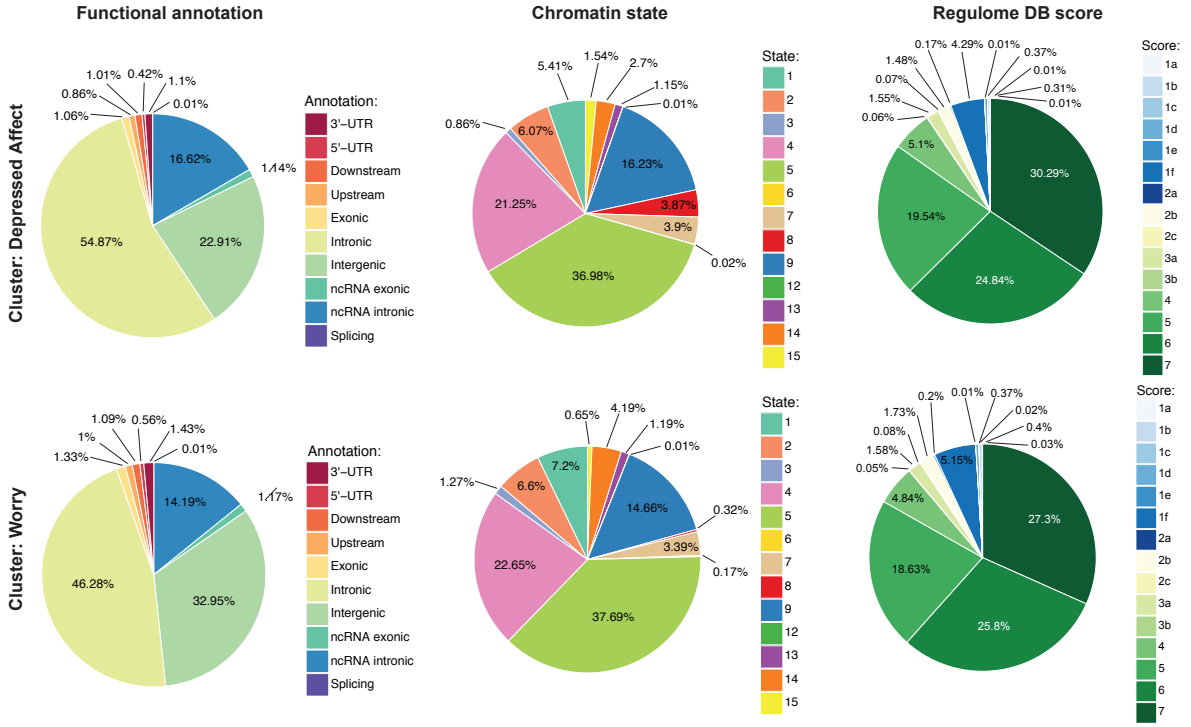
Supplementary Fig. 13 | Venn diagram showing overlap in associated genomic loci for neuroticism, depression, depressed affect and worry. Loci for neuroticism include the two low confidence loci, discussed in the **Supplementary Note**.



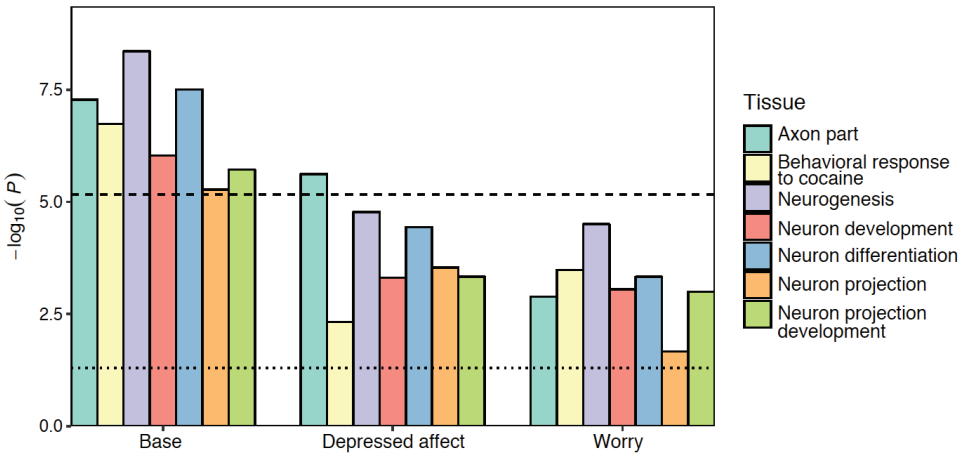
Supplementary Fig. 14 | Venn diagram showing overlap in associated genes for neuroticism, depression, depressed affect and worry.



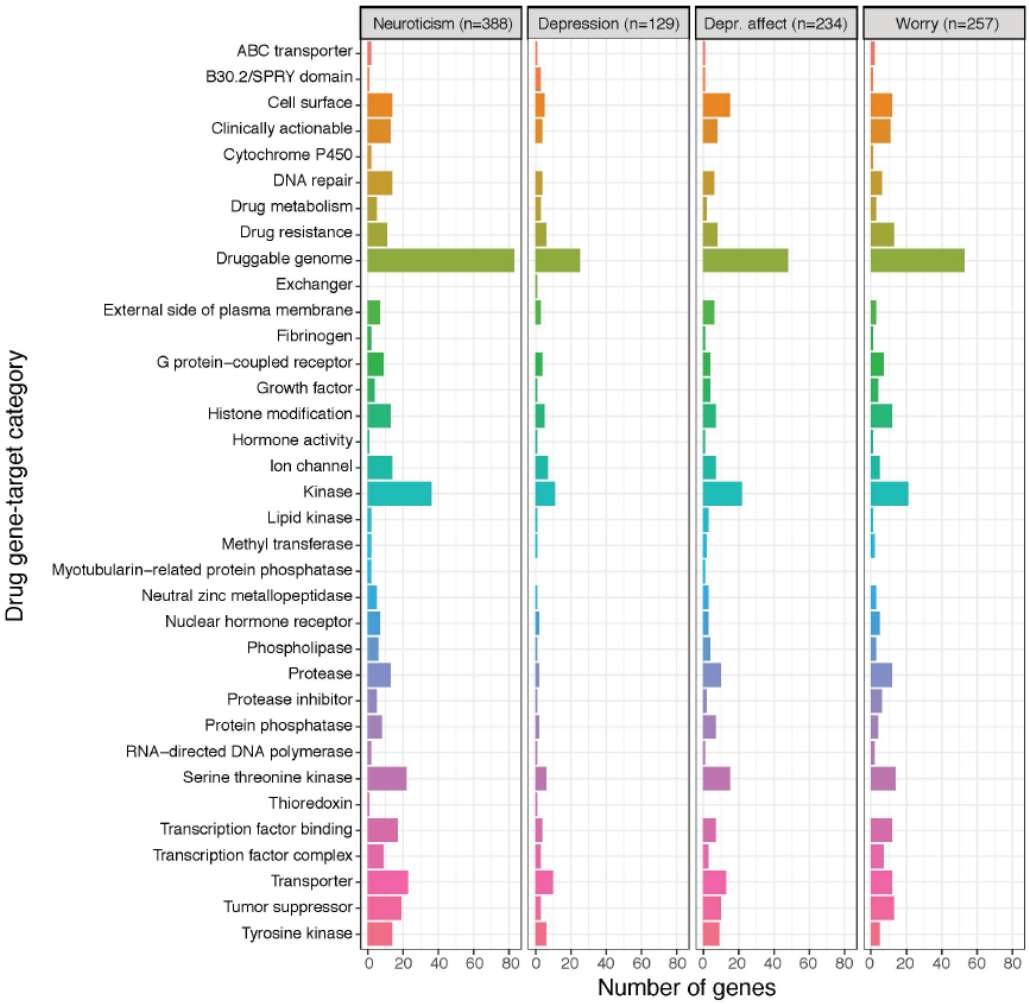
Supplementary Fig 15 | Functional categories, chromatin state and regulome DB score for SNPs that were GWS for the depressed affect and worry clusters. The functional annotation indicates the functional consequences on the gene to which a given SNP was annotated (**Supplementary Table 31**). The chromatin state refers to the minimum (i.e., most active) chromatin state across 127 tissues for all genome-wide significant SNPs. The lower the regulome DB (database) score, the more likely it is that the SNP has a regulatory function. Coding of the chromatin states and regulome DB scores is presented in **Supplementary Tables 32-33**. Annotation was performed in FUMA[20].



Supplementary Fig. 16 | Gene-set analysis conditional on depressed affect and worry. Association *P*-values of gene-sets for neuroticism, conditional on the scores on the neuroticism item clusters depressed affect and worry. The initial conditional gene-set analysis (**Supplementary Table 35**) showed that 3 of the 6 gene-sets (axon part, behavioral response to cocaine and neurogenesis) had largely independent associations with neuroticism. Repeating the conditional analyses, now conditioning on depressed affect and worry scores, respectively, shows that the involvement of 'axon part' in neuroticism may largely originate in the worry component of neuroticism (compared to conditioning on depressed affect, the *P*-value drops more substantially). Conditional gene-set analyses were conducted in MAGMA using theresults of the gene-based analyses as input.



Supplementary Fig. 17 | Potentially druggable gene categories for all four traits. Bar plot showing the number of potentially druggable gene-targets based on data from the drug-gene interaction database[32,33] (DGIdb). The search in the DGIdb was performed on all genes that were implicated by one of the gene-mapping strategies for each of the four traits. These categories highlight gene-targets that may form a potential drug-target, but which is not necessarily already targeted by currently approved drugs. A full overview of the results is provided in **Supplementary Table 41**.



Chapter 5

Genome-wide Association Meta-analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence

Jeanne E. Savage*, Philip R. Jansen*, Sven Stringer, Kyoko Watanabe, Julien Bryois, Christiaan A. de Leeuw, Mats Nagel, Swapnil Awasthi, Peter B. Barr, Jonathan R. I. Coleman, Katrina L. Grasby, Anke R. Hammerschlag, Jakob Kaminski, Robert Karlsson, Eva Krapohl, Max Lam, Marianne Nygaard, Chandra A. Reynolds, Joey W. Trampush, Hannah Young, Delilah Zabaneh, Sara Hägg, Narelle K. Hansell, Ida K. Karlsson, Sten Linnarsson, Grant W. Montgomery, Ana B. Muñoz-Manchado, Erin B. Quinlan, Gunter Schumann, Nathan Skene, Bradley T. Webb, Tonya White, Dan E. Arking, Deborah K. Attix, Dimitrios Avramopoulos, Robert M. Bilder, Panos Bitsios, Katherine E. Burdick, Tyrone D. Cannon, Ornit Chiba-Falek, Andrea Christoforou, Elizabeth T Cirulli, Eliza Congdon, Aiden Corvin, Gail Davies, Ian J. Deary, Pamela DeRosse, Dwight Dickinson, Srdjan Djurovic, Gary Donohoe, Emily Drabant Conley, Johan G. Eriksson, Thomas Espeseth, Nelson A. Freimer, Stella Giakoumaki, Ina Giegling, Michael Gill, David C. Glahn, Ahmad R. Hariri, Alex Hatzimanolis, Matthew C. Keller, Emma Knowles, Bettina Konte, Jari Lahti, Stephanie Le Hellard, Todd Lencz, David C. Liewald, Edythe London, Astri J. Lundervold, Anil K. Malhotra, Ingrid Melle, Derek Morris, Anna C. Need, William Ollier, Aarno Palotie, Antony Payton, Neil Pendleton, Russell A. Poldrack, Katri Räikkönen, Ivar Reinvang, Panos Roussos, Dan Rujescu, Fred W. Sabb, Matthew A. Scult, Olav B. Smeland, Nikolaos Smyrnis, John M. Starr, Vidar M. Steen, Nikos C. Stefanis, Richard E. Straub, Kjetil Sundet, Aristotle N. Voineskos, Daniel R. Weinberger, Elisabeth Widen, Jin Yu, Goncalo Abecasis, Ole A. Andreassen, Gerome Breen, Lene Christiansen, Birgit Debrabant, Danielle M. Dick, Andreas Heinz, Jens Hjerling-Leffler, M. Arfan Ikram, Kenneth S. Kendler, Nicholas G. Martin, Sarah E. Medland, Nancy L. Pedersen, Robert Plomin, Tinca J.C. Polderman, Stephan Ripke, Sophie van der Sluis, Patrick F. Sullivan, Henning Tiemeier, Scott I. Vrieze, Margaret J. Wright, Danielle Posthuma

* = *authors contributed equally*

Published in *Nature Genetics* 2018 Jul;50(7):912-919

Intelligence is highly heritable[1] and a major determinant of human health and wellbeing[2]. Recent genome-wide meta-analyses have identified 24 genomic loci linked to variation in intelligence[3-7], but much about its genetic underpinnings remains to be discovered. Here, we present a large-scale genetic association study of intelligence (n = 269,867), identifying 205 associated genomic loci (190 new) and 1,016 genes (939 new) via positional mapping, expression quantitative trait locus (eQTL) mapping, chromatin interaction mapping, and gene-based association analysis. We find enrichment of genetic effects in conserved and coding regions and associations with 146 nonsynonymous exonic variants. Associated genes are strongly expressed in the brain, specifically in striatal medium spiny neurons and hippocampal pyramidal neurons. Gene set analyses implicate pathways related to nervous system development and synaptic structure. We confirm previous strong genetic correlations with multiple health-related outcomes, and Mendelian randomization analysis results suggest protective effects of intelligence for Alzheimer's disease and ADHD and bidirectional causation with pleiotropic effects for schizophrenia. These results are a major step forward in understanding the neurobiology of cognitive function as well as genetically related neurological and psychiatric disorders.

We performed a genome-wide association study (GWAS) meta-analysis of 14 independent epidemiological cohorts of European ancestry and 9,295,118 genetic variants passing quality control (**Table 1, Supplementary Fig. 1**, and **Supplementary Table 1**). A flowchart of the study methodology is presented in **Supplementary Fig. 2**, and additional details of the methods and results are presented in the **Supplementary Note**. Intelligence was assessed using various neurocognitive tests, primarily gauging fluid domains of cognitive functioning (**Supplementary Note**). Despite variation in form and content, cognitive test scores display a positive manifold of correlations, a robust empirical phenomenon that is observed in multiple populations[8]. Statistically, the variance common across cognitive tasks can be modeled as a latent factor denoted as *g* (the general factor of intelligence)[9,10]. In addition, twin and family studies show strong genetic correlations across diverse cognitive domains[11], suggesting pleiotropy, and across levels of ability[11], substantiating the view of general intelligence as an etiological continuum (with rare syndromic forms of severe intellectual disability being the exception[12]). Additionally, *g* factors extracted from different sets of cognitive tests correlate very strongly (>0.98[13,14]), supporting the universality of *g*[15,16]. In performing meta-analysis of cognitive scores obtained using a variety of tests, we aimed to boost the statistical power to detect genetic variants underlying *g*, which are likely to have pleiotropic effects across multiple domains of cognitive functioning.

Despite sample and methodological variations, genetic correlations (r_g) between cohorts were considerable (mean = 0.67), and there was no evidence of heterogeneity between cohorts in the SNP associations (**Supplementary Table 2** and **Supplementary Note**). Age-stratified meta-analyses indicated high genetic correlations (r_g > 0.62) and comparable heritability across age groups, as captured by the SNPs included in the analysis (r_g = 0.19–0.22) (**Supplementary Table 3** and **Supplementary Note**). The full-sample was 0.19 (standard error (s.e.) = 0.01), in line with previous findings[4,5], and a linkage disequilibrium (LD) score intercept[17] of 1.08 (s.e. = 0.02) indicated that most of the inflation (λ = 1.92) could be explained by polygenic signal[6] (**Supplementary Fig. 3** and **Supplementary Table 4**).

In the meta-analysis, 12,110 variants indexed by 242 lead SNPs in approximate linkage equilibrium ($r < 0.1$) reached genome-wide significance ($P < 5 \times 10^{-8}$) (**Fig. 1a, Supplementary Figs. 4 and 5**, and **Supplementary Tables 5–7**). These were located in 205 distinct genomic loci (**Supplementary Note**). We tested for replication using the proxy phenotype of educational attainment, which is correlated phenotypically (r_g ~0.40)[18] and genetically (r_g ~0.70)[19] with intelligence. We confirmed this high genetic correlation (r_g = 0.73) and observed sign concordance with educational attainment for 93% of genome-wide significant SNPs ($P < 1 \times 10^{-300}$), with replication for 48 loci (**Supplementary Table 8** and **Supplementary Note**). Using polygenic score (PGS) prediction[20,21], the current results explain up to 5.2% of the variance in intelligence in four independent samples (**Supplementary Table 9** and **Supplementary Note**). We observed enrichment for heritability of SNPs in conserved regions ($P = 2.01 \times 10^{-12}$), coding regions ($P = 1.67 \times 10^{-6}$), and acetylated Lys9 of histone H3 (H3K9ac) histone regions/peaks ($P < 6.26 \times 10^{-5}$), and among common (minor allele frequency (MAF) > 0.3) variants (**Fig. 1b, Supplementary Figs. 6 and 7, Supplementary Table 10**, and **Supplementary Note**). Conserved and regulatory regions have previously been implicated in cognitive functioning[22], but coding regions have not.

Functional annotation of all candidate SNPs in the asso-

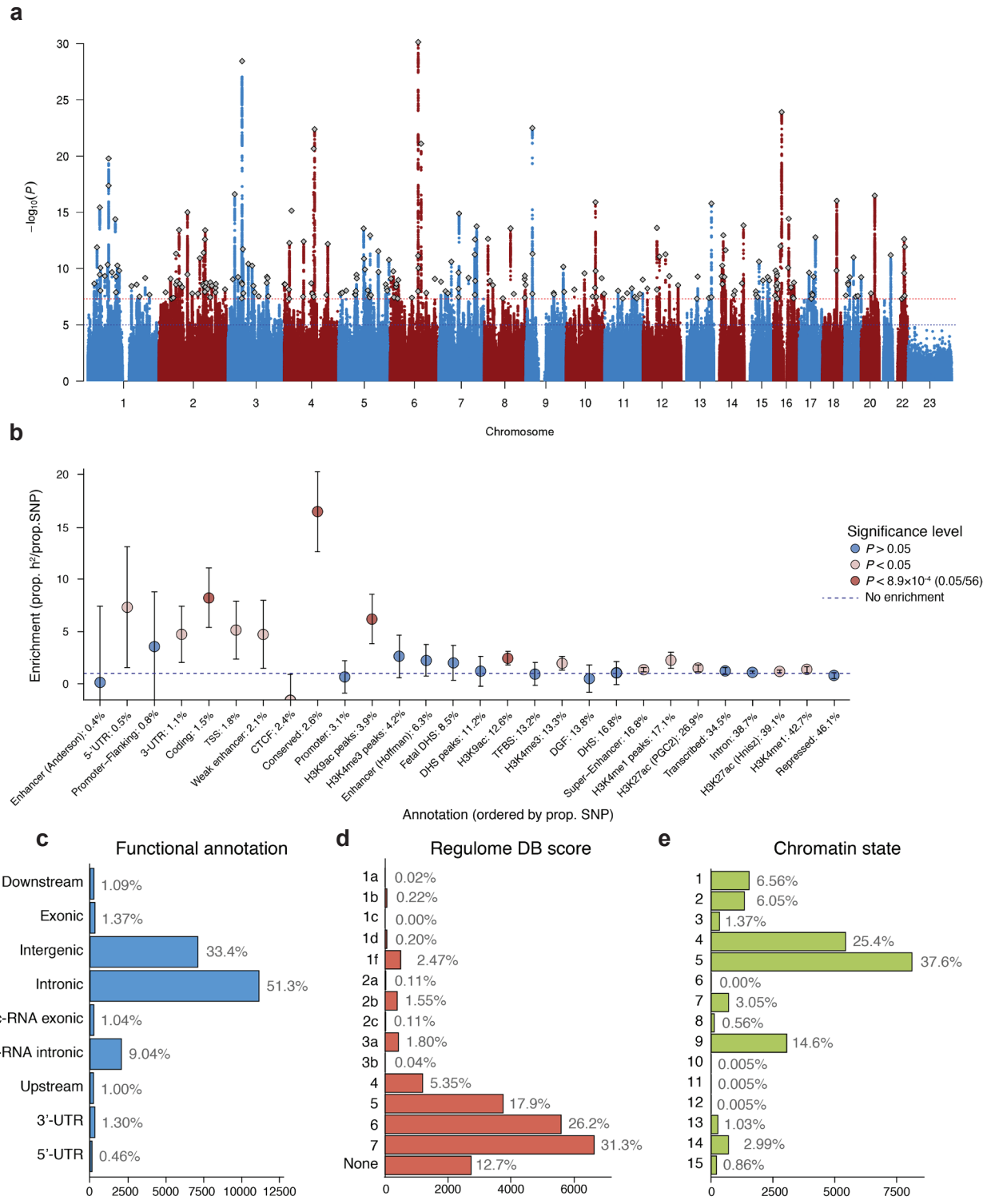

a
−log10(P)
Chromosome
b
Enrichment (prop. h²/prop.SNP)
Significance level
P > 0.05
P < 0.05
P < 8.9×10⁻⁴ (0.05/56)
No enrichment
Enhancer (Anderson): 0.4%
5'-UTR: 0.5%
Promoter-Flanking: 0.8%
3-UTR: 1.1%
Coding: 1.5%
TSS: 1.8%
Weak enhancer: 2.1%
CTCF: 2.4%
Conserved: 2.6%
Promoter: 3.1%
H3K9ac peaks: 3.9%
H3K4me3 peaks: 4.2%
Enhancer (Hoffman): 6.3%
Fetal DHS: 8.5%
DHS peaks: 11.2%
H3K9ac: 12.6%
TFBS: 13.2%
H3K4me3: 13.3%
DGF: 13.8%
DHS: 16.8%
Super-Enhancer: 16.8%
H3K4me1 peaks: 17.1%
H3K27ac (PGC2): 26.9%
Transcribed: 34.5%
Intron: 38.7%
H3K27ac (Hnisz): 39.1%
H3K4me1: 42.7%
Repressed: 46.1%
Annotation (ordered by prop. SNP)
c
Functional annotation
Downstream 1.09%
Exonic 1.37%
Intergenic 33.4%
Intronic 51.3%
nc-RNA exonic 1.04%
nc-RNA intronic 9.04%
Upstream 1.00%
3'-UTR 1.30%
5'-UTR 0.46%
d
Regulome DB score
1a 0.02%
1b 0.22%
1c 0.00%
1d 0.20%
1f 2.47%
2a 0.11%
2b 1.55%
2c 0.11%
3a 1.80%
3b 0.04%
4 5.35%
5 17.9%
6 26.2%
7 31.3%
None 12.7%
e
Chromatin state
1 6.56%
2 6.05%
3 1.37%
4 25.4%
5 37.6%
6 0.00%
7 3.05%
8 0.56%
9 14.6%
10 0.005%
11 0.005%
12 0.005%
13 1.03%
14 2.99%
15 0.86%

Fig. 1 (previous page) | SNP-based associations with intelligence in the GWAS meta-analysis of n = 269,867 independent individuals. a, Manhattan plot showing the $-\log_{10}$ transformed two-tailed *P*-value of each SNP from the GWAS meta-analysis (of linear and logistic regression statistics) on the y axis and base-pair positions along the chromosomes on the x axis. The dotted red line indicates Bonferroni-corrected genome-wide significance ($P < 5 \times 10^{-8}$); the dotted blue line indicates the threshold for suggestive association ($P < 1 \times 10^{-5}$). Independent lead SNPs are indicated by a diamond. **b**, Heritability enrichment of 28 functional annotation categories for SNPs in the meta-analysis, calculated with stratified LD Score regression. Error bars show 95% confidence intervals around the enrichment estimates. The dashed horizontal line indicates no enrichment of the annotation category. Red dots represent significant Bonferroni-corrected two-tailed *P*-values, and beige dots represent suggestive ($P < 0.05$) values. TSS, transcription start site; CTCF, CCCTCbinding factor; DHS, DNase I– hypersensitive site; TFBS, transcription factor binding site; DGF, DNase I digital genomic footprint. **c**, Distribution of the functional consequences of SNPs in genomic risk loci in the meta-analysis. **d**, Distribution of RegulomeDB scores for SNPs in genomic risk loci, with a low score indicating a higher likelihood of the SNP having a regulatory function (**Methods**). **e**, The minimum chromatin state across 127 tissue and cell types for SNPs in genomic risk loci, with lower states indicating higher accessibility and states 1–7 referring to open chromatin states (**Methods**).

ciated loci (SNPs with $r^2 \geq 0.6$ with one of the independent significant SNPs, a suggestive *P*-value ($P < 1 \times 10^{-5}$), and MAF > 0.0001; n = 21,368) showed that these were mostly intronic or intergenic (**Fig. 1** and **Supplementary Table 6**), yet 146 (81 genome-wide significant) SNPs were exonic nonsynonymous (ExNS) (**Supplementary Table 11** and **Supplementary Note**). Convergent evidence of strong association (z = 9.49) and the highest observed probability of a deleterious protein effect (CADD[23] score = 34) were found for rs13107325. This missense mutation (MAF = 0.065, $P = 2.23 \times 10^{-21}$) in *SLC39A8* was the lead SNP in locus 71, and the ancestral C allele was associated with higher scores on intelligence measures. The effect sizes for ExNS SNPs were individually small, with each effect allele accounting for a difference of 0.01 to 0.08 s.d. A detailed catalog of variants in the associated genomic loci is presented in **Supplementary Tables 6** and **11** and in the **Supplementary Note**.

To link the associated variants to genes, we applied three gene-mapping strategies implemented in FUMA[24]. Positional gene mapping aligned SNPs to 522 genes by genomic location, eQTL gene mapping matched cis-eQTL SNPs to 684 genes whose expression levels they influence, and chromatin interaction mapping annotated SNPs to 227 genes on the basis of 3D DNA–DNA interactions (**Fig. 2**, **Supplementary Figs. 8** and **9, Supplementary Tables 12–14**, and **Supplementary Note**). This resulted in 859 unique mapped genes, 435 of which were implicated by at least two mapping strategies and 139 of which were implicated by all three (**Fig. 3**). Although not all of these genes are certain to have a role in intelligence, they point to potential functional links for the GWAS-associated variants and give higher credibility to genes with convergent evidence of association from multiple sources. The FUMA-mapped genes were enriched for brain tissue expression and several regulatory biological gene sets (**Supplementary Note**). Fifteen genes are particularly notable as they are implicated via chromatin interactions between two independent genomic risk loci (**Fig. 2** and **Supplementary Note**). Cross-locus interactions implicated *ELAVL2*, *PTCH1*, *ATF4*, *FBXL17*, and *MAN2A1* in the left ventricle of the heart, *SATB2* in liver tissue, and *MEF2C* in five tissues. Multiple interactions in multiple tissue types were seen for a cluster of eight genes on chromosome 6 encoding zinc-finger proteins and histones.

We performed genome-wide gene-based association study (GWGAS) analysis using MAGMA[25] to estimate aggregate associations on the basis of all SNPs in a gene (whereas FUMA annotates individually significant SNPs to genes). GWGAS analysis identified 507 associated genes (**Fig. 3a**, **Supplementary Table 15**, and **Supplementary Note**), of which 350 were also mapped by FUMA (**Fig. 3b**). In total, 105 genes were implicated by all four strategies (**Supplementary Table 16**).

In gene set analysis, six Gene Ontology[26] gene sets were significantly associated with intelligence: neurogenesis ($P = 4.78 \times 10^{-7}$), neuron differentiation ($P = 4.82 \times 10^{-6}$), central nervous system neuron differentiation ($P = 3.31 \times 10^{-6}$), regulation of nervous system development ($P = 9.30 \times 10^{-7}$), positive regulation of nervous system development ($P = 1.00 \times 10^{-6}$), and regulation of synapse structure or activity ($P = 5.42 \times 10^{-6}$) (**Supplementary Tables 17** and **18**, and **Supplementary Note**). Conditional analysis indicated that there were three independent associations—regulation of nervous system development, central nervous system neuron differentiation, and regulation of synapse structure or activity—that together accounted for the associations of the other sets.

Linking gene-based *P*-values to tissue-specific gene expression, we observed strong associations with gene expression across multiple brain areas (**Fig. 3c, Supplemen-**

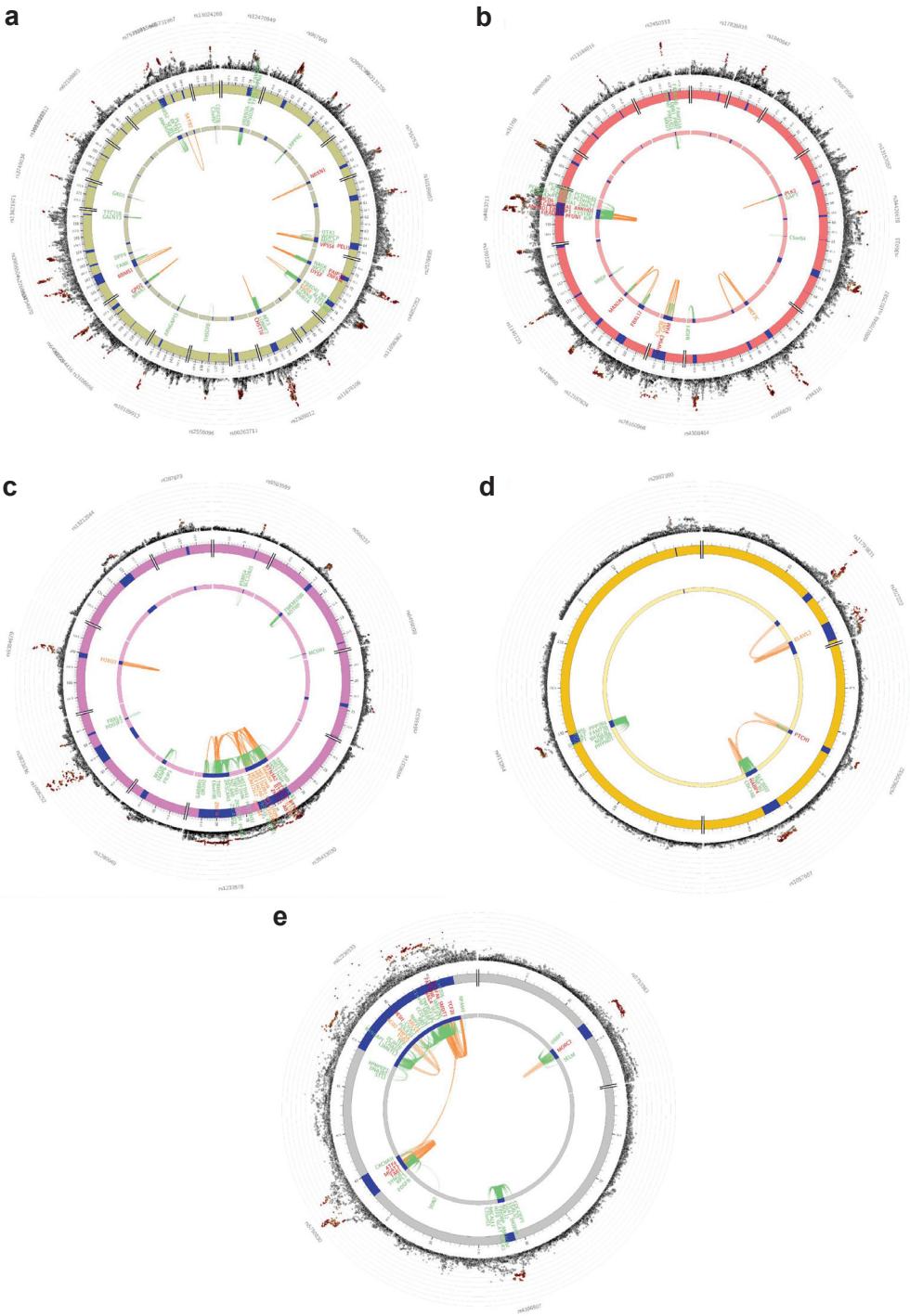


Fig. 2 | Cross-locus interactions for genomic regions associated with intelligence in 269,867 independent individuals. a–e, Circos plots showing genes on chromosomes 2 (**a**), 5 (**b**), 6 (**c**), 9 (**d**), and 22 (**e**) that were linked to genomic risk loci in the GWAS meta-analysis (blue regions) by eQTL mapping (green lines connecting an eQTL SNP to its associated gene) and/or chromatin interactions (orange lines connecting two interacting regions) and showed evidence of interaction across two independent genomic risk loci. Genes implicated by eQTLs are in green, by chromatin interactions are in orange, and by both eQTLs and chromatin interactions are in red. The outer layer shows a Manhattan plot containing the –log transformed two-tailed *P*-value of each SNP from the GWAS meta-analysis (of linear and logistic regression statistics), with genome-wide significant SNPs colored according to LD patterns with the lead SNP. Circos plots for all chromosomes are provided in **Supplementary Fig. 8**.

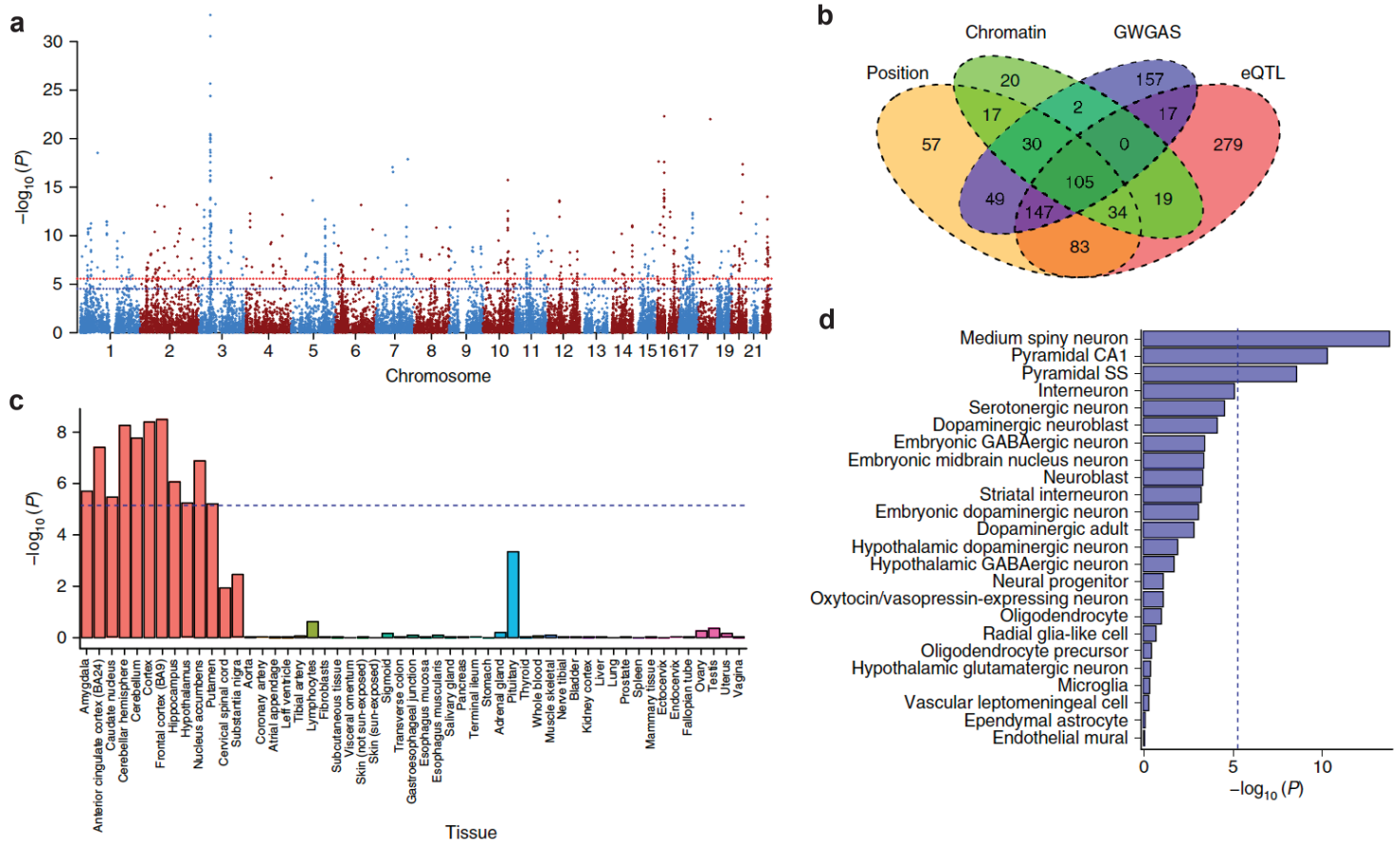


Fig. 3 | Implicated genes, pathways, and tissue and cell expression profiles for intelligence in 269,867 independent individuals. **a**, Manhattan plot of the GWGAS analysis. The y axis shows the – log transformed two-tailed *P*-value of each gene from a linear model, and chromosomal position is shown on the x axis. The red dotted line indicates the Bonferroni-corrected threshold for genome-wide significance of the gene-based test ($P < 2.76 \times 10^{-6}$; 0.05/18,128 genes), and the blue dotted line indicates the suggestive threshold ($P < 2.76 \times 10^{-6}$; 0.5/18,128 genes). **b**, Venn diagram showing overlap of genes implicated by positional mapping, eQTL mapping, chromatin interaction mapping, and GWGAS. **c**, Gene expression profiles of associated genes in 53 tissue types. The y axis shows the –log transformed two-tailed *P*-value of association of GWGAS test statistics with tissue-specific gene expression levels in a linear model. Expression data were extracted from the Genotype-Tissue Expression (GTEx) database. Expression values (RPKM) were log transformed with pseudo-count 1 after Winsorization at 50 and averaged per tissue. The dashed blue line indicates the Bonferroni-corrected significance threshold (P = 0.05/7,323 gene sets = 6.83×10^{-6}). **d**, Single-cell gene expression analysis of genes related to intelligence in 24 cell types. The x axis shows the –log transformed two-tailed *P*-value of association of GWGAS test statistics with cell-specific gene expression levels in a linear model. The dashed blue line indicates the Bonferroni-corrected significance threshold ($P = 0.05/7,323$ gene sets = 6.83×10^{-6}).

tary Table 19, and **Supplementary Note**), particularly the frontal cortex ($P = 3.10 \times 10^{-9}$). In brain single-cell expression gene set analyses, we found significant associations of striatal medium spiny neurons ($P = 2.02 \times 10^{-14}$), and hippocampal CA1 pyramidal neurons in the CA1 hippocampal ($P = 5.67 \times 10^{-11}$) and cortical ($P = 2.72 \times 10^{-9}$) somatosensory regions (**Fig. 3d**, **Supplementary Table 20**, and **Supplementary Note**). Conditional analysis showed that the independent association signal in brain cells was driven by medium spiny neurons, neuroblasts, and pyramidal CA1 neurons.

Intelligence has been associated with a wide variety of human behaviors[15] and brain anatomy[27]. Confirming previous reports[5, 6], we observed negative genetic correlations with attention deficit/hyperactivity disorder (ADHD; $r_g = -0.36$, $P = 4.58 \times 10^{-23}$), depressive symptoms ($r_g = -0.27$, $P = 6.20 \times 10^{-10}$), Alzheimer's disease ($r_g = -0.27$, $P = 2.03 \times 10^{-5}$), and schizophrenia ($r_g = -0.21$, $P = 3.82 \times 10^{-17}$) and positive correlations with longevity ($r_g = 0.43$, $P = 7.96 \times 10^{-8}$) and autism ($r_g = 0.25$, $P = 3.14 \times 10^{-7}$), among others (**Supplementary Fig. 10** and **Supplementary Table 21**). Comparison with previous GWAS[28] supported these correlations, showing numerous shared genetic variants across phenotypes (**Supplementary Tables 22** and **23**, and **Supplementary Note**). Low enrichment (87 of 1,518 genes, $P = 0.05$) was found for genes previously linked to intellectual disability or developmental delay, indicating largely distinct biological processes. However, our results extend previous genetic research on normal variation in general intelligence, as catalogued in **Supplementary Tables 24** and **25**.

We used Generalized Summary-statistic-data-based Mendelian Randomization[29] to test for potential credible causal associations between intelligence and genetically correlated traits (**Supplementary Figs. 11** and **12, Supplementary Table 26**, and **Supplementary Note**). We observed a strong bidirectional effect of cognitive ability on educational attainment ($b_{xy} = 0.549$, $P < 1 \times 10^{-320}$) and of educational attainment on intelligence ($b_{xy} = 0.480$, $P = 6.85 \times 10^{-82}$). Such findings are consistent with previous studies implicating bidirectional causal effects[30, 31]. There was also a bidirectional association showing a strong protective effect of intelligence on schizophrenia (odds ratio (OR) = 0.50, $b_{xy} = -0.685$, $P = 2.02 \times 10^{-57}$) and a relatively smaller reverse effect ($b_{xy} = -0.214$, $P = 4.19 \times 10^{-52}$), with additional evidence for pleiotropy (**Supplementary Note**). A number of previous reports support both a causal link and genetic overlap between these phenotypes[32, 33]. Our results also suggested that higher intelligence had a protective effect on ADHD (OR = 0.48, $b_{xy} = -0.734$, P

Table 1 | Overview of the cohorts included in a GWAS meta-analysis of general intelligence.

Cohort	Cohort name	n	Age (years)	Phenotype
1	UKB	195,653	39 - 72	Verbal and mathematical reasoning
2	COGENT	35,289	8 - 96	One or more neuropsychological tests from three or more domains of cognitive performance
3	RS	6,182	45 - 98	Letter–digit substitution, Stroop, verbal fluency, delayed recall
4	GENR	1,929	5 - 9	SON-R (spatial visualization and abstract reasoning subsets)
5	STR	3,215	18	Logical, verbal, spatial, and technical ability subtests
6	S4S	2,818	17 - 18	SAT test scores
7	HiQ/HRS	9,410	NA[a]	High-IQ cases/unselected population controls
8	TEDS	3,414	12	WISC-III verbal and nonverbal reasoning; Raven's progressive matrices
9a	DTR-MADT	737	55 - 80	Verbal fluency, digit span, immediate and delayed recall tests
9b	DTR-LSADT	253	73 - 94	Verbal fluency, digit span, immediate and delayed recall tests
10	IMAGES	1,343	14	WISC-IV, CANTAB factor score
11a	BLTS-children	530	12 - 13	VSRT-C factor score
11b	BLTS-adolescents	2,598	15 - 30	MAB-II IQ score
12	NESCOG	252	18 - 79	WAIS IQ score
13	GfG	5,084	15 - 91	ICAR verbal reasoning test
14a	STSA-SATSA+GENDER	703	50 - 94	Verbal, spatial, episodic memory, and processing speed tests
14b	STSA-HARMONY	448	65 - 96	Verbal, spatial, episodic memory, and processing speed tests

[a] The HiQ/HRS sample used a case–control design rather than a cognitive test score ascertained at a specific age; see the **Methods** and **Supplementary Note**.

= 2.57×10^{-46}) and Alzheimer's disease (OR = 0.65, b_{xy} = −0.435, $P = 3.59 \times 10^{-14}$), but was associated with higher risk of autism (OR = 1.38, $b_{xy} = 0.321$, $P = 1.12 \times 10^{-3}$).

In the present study, we have affirmed and expanded existing knowledge of the genetics of general intelligence, identifying 190 new loci and 939 new associated genes and replicating previous associations with 15 loci and 77 genes. The combined strategies of functional annotation and gene mapping using biological data resources provide extensive information on the likely consequences of relevant genetic variants and put forward a rich set of plausible gene targets and biological mechanisms for functional follow-up. Gene set analyses contribute novel insight into underlying neurobiological pathways, confirming the importance of brain-expressed genes and neurodevelopmental processes in fluid domains of intelligence and pointing toward the involvement of specific cell types. Our results indicate overlap in the genetic processes involved in both cognitive functioning and neurological and psychiatric traits and provide suggestive evidence of causal associations that may drive these correlations. These results are important for understanding the biological underpinnings of cognitive functioning and contribute to understanding of related neurological and psychiatric disorders.

Online Methods

Methods Study cohorts

The meta-analysis included new and previously reported GWAS summary statistics from 14 cohorts: UK Biobank (UKB), the Cognitive Genomics Consortium (COGENT), the Rotterdam Study (RS), the Generation R Study (GENR), the Swedish Twin Registry (STR), Spit for Science (S4S), the HighIQ/Health and Retirement Study (HiQ/HRS), the Twins Early Development Study (TEDS), the Danish Twin Registry (DTR), IMAGEN, the Brisbane Longitudinal Twin Study (BLTS), the Netherlands Study of Cognition, Environment, and Genes (NESCOG), Genes for Good (GfG), and the Swedish Twin Studies of Aging (STSA). All samples were obtained from epidemiological cohorts ascertained for research on a variety of physical and psychological outcomes. Participants ranged from children to older adults, with older samples being screened for cognitive decline to exclude the possibility of dementia affecting performance on cognitive tests. Different measures of intelligence were assessed in each cohort but were all operationalized to index a common latent g factor underlying multiple dimensions of cognitive functioning. With the exception of HiQ/HRS, all cohorts extracted a single sum score, mean score, or factor score from a multidimensional set of cognitive performance tests and used this normally distributed score as the phenotype in a covariate-adjusted (for example, age, sex, ancestry principal components) GWAS using linear regression methods. For HiQ/HRS, a logistic regression GWAS was run with 'case' status reflecting whether participants were drawn from an extreme-sampled population of very high intelligence (i.e., at the upper ~0.03% of the tail of the normal distribution) versus an epidemiological sample of unselected population 'controls'. Detailed descriptions of the samples, measures, genotyping, quality control, and analysis procedures for each cohort are provided in the **Supplementary Note, Supplementary Table 1**, and in the Nature Research Reporting Summary.

Meta-analysis

Stringent quality control measures were applied to the summary statistics for each GWAS cohort before combining. All files were checked for data integrity and accuracy. SNPs were filtered from further analysis if they met any of the following criteria: imputation quality (INFO/R^2) score < 0.6, Hardy–Weinberg equilibrium $P < 5 \times 10^{-6}$, study-specific minor allele frequency (MAF) corresponding to a minor allele count (MAC) < 100, and mismatch of alleles or allele frequency difference greater than 20% from the Haplotype Reference Consortium (HRC) genome reference panel[16]. Some cohorts used more stringent criteria (**Supplementary Note**). Indels and SNPs that were duplicated, multiallelic, monomorphic, or ambiguous (A/T or C/G with MAF > 0.4) were also excluded. Visual inspection of the distribution of the summary statistics was performed, and Manhattan plots and quantile–quantile plots were created for the cleaned summary statistics from each cohort (**Supplementary Fig. 1**). The SNP association *P*-values from the GWAS cohorts were subjected to meta-analysis with METAL[34] in two phases. First, we performed meta-analysis on all cohorts with quantitative phenotypes (all except HiQ/HRS) using a sample-size-weighted scheme. In the second phase, we added the HiQ/HRS study results to the results from the first phase, weighting each set of summary statistics by their respective non-centrality parameter (NCP). This method improves power when using an extreme case sampling design such as that in HiQ[35] and provides a comparable metric with which to combine information from different analytic designs while accounting for their differences in power/effective sample size. NCPs were estimated using the Genetic Power Calculator[36], as described by Coleman et al.[37]. After combining all data, meta-analysis results were further filtered to exclude any variants with n < 50,000. We additionally included a random-effects

meta-analysis for each phase, as implemented in METAL, to evaluate potential heterogeneity in the SNP association statistics between cohorts. The X-chromosome was treated separately in the meta-analysis because imputed genotypes were not available for the X-chromosome in the largest cohort (UKB), and there was little overlap between the UKB called genotypes and imputed data from other cohorts (n < 500). We therefore included only the called X-chromosome variants in UKB for these analyses after performing X-chromosome-specific qualitycontrol steps[38]. We conducted a series of meta-analyses on subsets of the full sample using the same methods as above. Age-group-specific meta-analyses were run in the cohorts of children (age < 17 years; GENR, TEDS, IMAGEN, BLTS; n = 9,814), young adults (age ~17–18 years; S4S, STR; n = 6,033), adults (age > 18 years, primarily middle-aged or older: UKB, RS, DTR, NESCOG, STSA; n = 204,228), and older adults (mean age > 60 years, RS, DTR, STSA; n = 8,323), excluding studies whose samples overlapped children/young adult and adult groups (COGENT, HiQ/HRS, GfG; n = 49,792). To create independent discovery samples for use in polygenic score validation, we also conducted meta-analyses with a 'leave-oneout' strategy in which summary statistics from four validation datasets were each excluded from the meta-analysis (see "Polygenic scoring").

Cohort heritability and genetic correlation

LD Score regression[17] was used to estimate genomic inflation and heritability of the intelligence phenotypes in each of the 14 cohorts using their post-quality-control summary statistics and to estimate the cross-cohort genetic correlations[39]. Pre-calculated LD scores from the 1000 Genomes European reference population were obtained online. Genetic correlations were calculated on HapMap 3 SNPs only. LD Score regression was also used on the agesub-group meta-analyses to estimate heritability and cross-age-group genetic correlations.

Genomic risk locus definition

Independently associated loci from the meta-analysis were defined using FUMA[24], an online platform for functional mapping of genetic variants. We first identified 'independent significant SNPs', which had a Bonferroni-corrected genome-wide significant two-tailed P-value ($P < 5 \times 10$) and represented signals that were independent from each other at $r^2 < 0.6$. These SNPs were further represented by 'lead SNPs', which are a subset of the independent significant SNPs that are in approximate linkage equilibrium with each other at $r^2 < 0.1$. We then defined associated 'genomic loci' by merging any physically overlapping lead SNPs (LD blocks < 250 kb apart). Borders of the associated genomic loci were defined by identifying all SNPs in LD ($r^2 \geq 0.6$) with one of the independent significant SNPs in the locus, and the region containing all of these 'candidate SNPs' was considered to be a single independent genomic locus. All LD information was calculated from UKB genotype data.

Proxy replication with educational attainment

We conducted GWAS of educational attainment, an outcome with a high genetic correlation with intelligence[5], in a nonoverlapping European subset of the UKB sample (n = 188,435) who did not complete the intelligence measure. Educational attainment was coded as maximum years of education completed, using the same methods as earlier analyses[40], and GWAS was conducted using the same qualitycontrol and analytic procedures as described for the UKB intelligence phenotype (**Supplementary Note**). To test replication of the SNPs with this proxy phenotype, we performed a sign concordance test for all genome-wide significant SNPs from the meta-analysis using the two-tailed exact binomial test. For each independent genomic locus, we considered it to be evidence for replication if the lead SNP or another correlated SNP in the region was sign concordant with the corresponding SNP in the intelligence meta-analysis and had a two-tailed P-value of association with educational attainment smaller than 0.05/242 independent tests = 0.0002.

Polygenic scoring

We calculated polygenic scores (PGSs) based on the SNP effect sizes of the leave-one-out meta-analyses, from which four cohorts were (separately) excluded and reserved for score validation. These included child (GENR), young adult (S4S), and adult (RS) samples. We also included the UKB-wb sample to test for validation in a very large (n = 53,576) cohort with the greatest phenotypic similarity to the largest contributor to the meta-analysis statistics (UKB-ts), to maximize potential predictive power. PGSs were calculated on the genotype data using LDpred[21], a Bayesian PGS method that uses a prior on effect size distribution to remodel the SNP effect size and account for LD, and PRSice[20], a PLINK[41] based program that automates optimization of the set of SNPs included in the PGS based on high-resolution filtering of the GWAS P-value threshold. LDpred PGSs were applied to the called, cleaned, genotyped variants in each of the validation cohorts with UKB as the LD reference panel. PRSice PGSs were calculated on hard-called imputed genotypes using P-value thresholds from 0.0 to 0.5 in steps of 0.001. The explained variance (ΔR^2) was derived from a linear

model in which the GWAS intelligence phenotype was regressed on each PGS while controlling for the same covariates as in each cohort-specific GWAS, compared to a linear model with GWAS covariates only.

Stratified heritability

We partitioned SNP heritability using stratified LD sScore regression[42] in three ways: (i) by functional annotation category, (ii) by MAF in six percentile bins, and (iii) by chromosome. Annotations for 28 binary categories of putative functional genomic characteristics (for example, coding or regulatory regions) were obtained from the LD score website. With this method, enrichment/depletion of heritability in each category is calculated as the proportion of heritability attributable to SNPs in the specified category divided by the proportion of total SNPs annotated to that category. The Bonferroni-corrected significance threshold was 0.05/56 annotations = 0.0009.

Functional annotation of SNPs

Functional annotation of SNPs implicated in the meta-analysis was performed using FUMA[24]. We selected all candidate SNPs in associated genomic loci having r ≥0.6 with one of the independent significant SNPs, a suggestive *P*-value ($P < 1 \times 10^{-5}$), and MAF > 0.0001 for annotations. Predicted functional consequences for these SNPs were obtained by matching SNPs' chromosome, base-pair position, and reference and alternate alleles to databases containing known functional annotations, including ANNOVAR[43] categories, combined annotationdependent depletion (CADD) scores[23], RegulomeDB[44] (RDB) scores, and chromatin states[45, 46]. ANNOVAR categories identify the SNP's genic position (for example, intron, exon, intergenic) and associated function. CADD scores predict how deleterious the effect of a SNP is likely to be for protein structure/function, with higher scores referring to higher deleteriousness. A CADD score above 12.37 is the threshold to be potentially pathogenic[23]. The RegulomeDB score is a categorical score based on information from eQTLs and chromatin marks, ranging from 1a to 7, with lower scores indicating an increased likelihood of having a regulatory function. Scores are as follows: 1a, eQTL + transcription factor (TF) binding + matched TF motif + matched DNase footprint + DNase peak; 1b, eQTL + TF binding + any motif + DNase footprint + DNase peak; 1c, eQTL + TF binding + matched TF motif + DNase peak; 1d, eQTL + TF binding + any motif + DNase peak; 1e, eQTL + TF binding + matched TF motif; 1f, eQTL + TF binding/DNase peak; 2a, TF binding + matched TF motif + matched DNase footprint + DNase peak; 2b, TF binding + any motif + DNase footprint + DNase peak; 2c, TF binding + matched TF motif + DNase peak; 3a, TF binding + any motif + DNase peak; 3b, TF binding + matched TF motif; 4, TF binding + DNase peak; 5, TF binding or DNase peak; 6, other; 7, not available. The chromatin state represents the accessibility of genomic regions (every 200 bp) with 15 categorical states predicted by a hidden Markov model based on 5 chromatin marks for 127 epigenomes in the Roadmap Epigenomics Project[46]. A lower state indicates higher accessibility, with states 1–7 referring to open chromatin states. We annotated the minimum chromatin state across tissues to SNPs. The 15 core chromatin states as suggested by Roadmap are as follows: 1, active transcription start site (TSS); 2, flanking active TSS; 3, transcription at gene 5' and 3' ends; 4, strong transcription; 5, weak transcription; 6, genic enhancer; 7, enhancers; 8, zinc-finger gene and repeats; 9, heterochromatic; 10, bivalent/poised TSS; 11, flanking bivalent/poised TSS/enhancer; 12, = bivalent enhancer; 13, repressed Polycomb; 14, weak repressed Polycomb; 15, quiescent/low. Standardized SNP effect sizes were calculated for the SNPs with the greatest impact by transforming the samplesize-weighted meta-analysis z score, as described by Zhu et al.[47].

Gene mapping

Genome-wide significant loci obtained by the GWAS meta-analysis were mapped
to genes in FUMA[24] using three strategies:

1. Positional mapping maps SNPs to genes based on physical distance (within a 10kb window) from known protein-coding genes in the human reference assembly (GRCh37/hg19)
2. eQTL mapping maps SNPs to genes with which they show a significant eQTL association (i.e., allelic variation at the SNP is associated with the expression level of that gene). eQTL mapping uses information from 45 tissue types in 3 data repositories (GTEx[48], Blood eQTL browser[49], BIOS QTL browser[50]) and is based on ciseQTLs that can map SNPs to genes up to 1 Mb away. We used a false discovery rate (FDR) of 0.05 to define significant eQTL associations
3. Chromatin interaction mapping was performed to map SNPs to genes when there was a 3D DNA–DNA interaction between the SNP region and a gene region. Chromatin interaction mapping can involve long-range interactions, as it does not have a distance boundary. FUMA currently contains HiC data for 14 tissue types from the study of Schmitt et al.[51]. Because chromatin interactions are often defined in a certain resolution, such as 40 kb, an interacting region can span multiple genes. If a SNP is located

in a region that interacts with a region containing multiple genes, it will be mapped to each of those genes. To further prioritize candidate genes, we selected only interaction-mapped genes in which one region involved in the interaction overlapped with a predicted enhancer region in any of the 111 tissue/cell types from the Roadmap Epigenomics project[46] and the other region was located in a gene promoter region (from 250 bp upstream to 500 bp downstream of the TSS and also predicted by Roadmap to be a promoter region). This reduced the number of genes mapped but increased the likelihood that those identified would have a plausible biological function. We used an FDR of 1×10^{-5} to define significant interactions, based on previous recommendations[51] modified to account for the differences in cell lines used here. Functional annotation of mapped genes Genes implicated by mapping of significant GWAS SNPs were further investigated using the GENE2FUNC procedure in FUMA[24], which provides hypergeometric tests of enrichment of the list of mapped genes in 53 GTEx[48] tissue-specific gene expression sets, 7,246 MSigDB gene sets[52], and 2,195 GWAS catalog gene sets[28]. The Bonferronicorrected significance threshold was 0.05/9,494 gene sets = 5.27×10^{-6}. Gene-based analysis

SNP-based *P*-values from the meta-analysis were used as input for GWGAS. 18,128 protein-coding genes (each containing at least 1 GWAS SNP) from the NCBI 37.3 gene definitions were used as the basis for GWGAS in MAGMA[25]. The Bonferroni-corrected genome-wide significance threshold was 0.05/18,128 genes = 2.76×10^{-6} .

Gene set analysis

Results from the GWGAS analyses were used to test for association in three types of predefined gene sets:

1. 7,246 curated gene sets representing known biological and metabolic pathways were derived from 9 data resources, catalogued by and obtained from MSigDB version 5.2[29];
2. Gene expression values from 53 tissues obtained from GTEx[48], log transformed with pseudo-count 1 after Winsorization at 50 and averaged per tissue;
3. Cell-type-specific gene expression in 24 types of brain cells, which were calculated following the method described in Skene et al.[53] and Coleman et al.[37]. Briefly, brain-celltype expression data were drawn from single-cell RNAseq data from mouse brains. For each gene, the value for each cell type was calculated by dividing the mean unique molecular identifier (UMI) counts for the given cell type by the summed mean UMI counts across all cell types. Single-cell gene sets were derived by grouping genes into 40 equal bins by specificity of expression.

These gene sets were tested for association with the GWGAS gene-based test statistics using MAGMA. We computed competitive *P*-values, which represent the test of association for a specific gene set in comparison to other gene sets. This method is more robust to type I error than self-contained tests that only test for association of a gene set against the null hypothesis of no association[25]. The Bonferroni-corrected significance threshold was 0.05/7,323 gene sets = 6.83×10^{-6} . Conditional analyses were performed as a follow-up using MAGMA to test whether each significant association observed was independent of all others. The association between each gene set was tested conditional on the most strongly associated set, and then—if any substantial ($P < 0.05$/number of gene sets) associations remained—by conditioning on the first and second most strongly associated set, and so on until no associations remained. Gene sets that retained their association after correcting for other sets were considered to be independent signals. We note that this is not a test of association per se, but rather a strategy to identify, among gene sets with known significant associations whose defining genes may overlap, which set(s) are responsible for driving the observed association.

Cross-trait genetic correlation

Genetic correlations (r_g) between intelligence and 38 phenotypes were computed using LD Score regression[39], as described above, based on GWAS summary statistics obtained from publicly available databases (**Supplementary Table 18**). The Bonferroni-corrected significance threshold was 0.05/38 traits = 1.32×10^{-3}.

GWAS catalog lookup

We used FUMA to identify SNPs with previously reported ($P < 5 \times 10^{-5}$) phenotypic associations in published GWAS listed in the NHGRIEBI catalog[28] that overlapped with the genomic risk loci identified in the meta-analysis. As an additional relevant phenotype of interest, we examined whether the genes associated with intelligence in this study (by FUMA mapping or GWGAS) were overrepresented in a set of 1,518 genes linked to intellectual disability and/or developmental delay, as compiled by Region-Annotater. Many of these have been identified by non-GWAS sources and are not represented in the NHGRI catalog. We tested for enrichment using a hypergeometric test with a background set of 19,283 genomic

protein-coding genes, as in FUMA. Manual lookups were also performed to identify overlapping loci/genes with known previous GWAS of intelligence.

Mendelian randomization

To infer credible causal associations between intelligence and traits that are genetically correlated with intelligence, we performed Generalized Summary-data-based Mendelian Randomization[29] (GSMR). This method uses summary-level data to test for causal associations between a putative risk factor (exposure) and an outcome by using independent genome-wide significant SNPs as instrumental variables. HEIDI outlier detection was used to filter genetic instruments that showed clear pleiotropic effects on both the exposure phenotype and the outcome phenotype. We used a threshold *P*-value of 0.01 for the outlier detection analysis in HEIDI, which removes 1% of SNPs by chance if there is no pleiotropic effect. To test for a potential causal effect of intelligence on various outcomes, we selected traits in nonoverlapping samples that showed significant genetic correlations (r_g) with intelligence. We tested for bidirectional causation by repeating the analyses while switching the role of each correlated phenotype as an exposure and intelligence as the outcome. For each trait, we selected independent ($r_g \leq 0.1$), genome-wide significant lead SNPs as instrumental variables in the analyses. For traits with fewer than ten genome-wide significant lead SNPs (i.e., the minimum number of SNPs on which GSMR can perform a reliable analysis), the genome-wide significance threshold was lowered to 1×10^{-5} , allowing a sufficient number of SNPs to conduct the reverse GSMR analysis for former smoker status, autism, and intracranial volume, and ADHD. The method estimates a putative causal effect of the exposure on the outcome (b_{xy}) as a function of the relationship between the SNPs' effects on the exposure (b_{zx}) and the SNPs' effects on the outcome (b_{zy}), given the assumption that the effect of non-pleiotropic SNPs on an exposure (x) should be related to their effect on the outcome (y) in an independent sample only via mediation through the phenotypic causal pathway (b_{xy}). The estimated causal effect coefficients (b_{xy}) are approximately equal to the natural log odds ratio (OR) for a case–control trait[29]. An OR of 2 can be interpreted as a doubled risk in comparison to the population prevalence of a binary trait for every s.d. increase in the exposure trait. For quantitative traits, b_{xy} can be interpreted as a 1 s.d. increase explained in the outcome trait for every s.d. increase in the exposure trait. This method can help differentiate the likely causal direction of association between two traits but cannot make any statement about the intermediate mechanisms involved in any potential causal process.

References

1. Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
2. Wraw, C., Deary, I. J., Gale, C. R. & Der, G. Intelligence in youth and health at age 50. *Intelligence* **53**, 23–32 (2015).
3. Davies, G. *et al.* Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N = 53949). *Mol. Psychiatry* **20**, 183–192 (2015).
4. Davies, G. *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N = 112 151). *Mol. Psychiatry* **21**, 758–767 (2016).
5. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
6. Trampush, J. W. *et al.* GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. *Mol. Psychiatry* **22**, 336–345 (2017).
7. Zabaneh, D. *et al.* A genome-wide association study for extremely high intelligence. *Mol. Psychiatry* **23**, 1226–1232 (2018).
8. Jensen, A.R. The G Factor: The Science of Mental Ability (Praeger, 1998). AQ8
9. Carroll, J. B. Human Cognitive Abilities: A Survey of Factor-Analytic Studies. (Cambridge University Press, Cambridge, UK, 1993).
10. Spearman, C. "General intelligence," objectively determined and measured. *Am. J. Psychol.* **15**, 201–292 (1904).
11. Plomin, R. & Kovas, Y. Generalist genes and learning disabilities. *Psychol. Bull.* **131**, 592–617 (2005).
12. Plomin, R. & von Stumm, S. The new genetics of intelligence. *Nat. Rev. Genet.* **19**, 148–159 (2018).
13. Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M. & Gottesman, I. I. Just one g: consistent results from three test batteries. *Intelligence* **32**, 95–107 (2004).
14. Johnson, W., Nijenhuis, Jt & Bouchard, T. J. Still just 1 g: consistent results from five test batteries. *Intelligence* **36**, 81–95 (2008).
15. Deary, I. J., Penke, L. & Johnson, W. The neuroscience of human intelligence differences. *Nat. Rev. Neurosci.* **11**, 201–211 (2010).
16. Deary, I. J. Intelligence. *Annu. Rev. Psychol.* **63**, 453–482 (2012).
17. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genomewide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Deary, I. J., Strand, S., Smith, P. & Fernandes, C. Intelligence and educational achievement. *Intelligence* **35**, 13–21 (2007).
19. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
20. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
21. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
22. Hill, W. D. *et al.* Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions. *Transl. Psychiatry* **6**, e980 (2016).
23. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
24. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with

FUMA. *Nat. Commun.* **8**, 1826 (2017).
25. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized geneset analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
26. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
27. Posthuma, D. *et al.* The association between brain volume and intelligence is of genetic origin. *Nat. Neurosci.* **5**, 83–84 (2002).
28. MacArthur, J. *et al.* The new NHGRIEBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 45(D1), D896–D901 (2017).
29. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
30. Johnson, W., Deary, I. J. & Iacono, W. G. Genetic and environmental transactions underlying educational attainment. *Intelligence* **37**, 466–478 (2009).
31. Richards, M. & Sacker, A. Is education causal? Yes. *Int. J. Epidemiol.* **40**, 516–518 (2011).
32. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. IQ and schizophrenia in a Swedish national sample: their causal relationship and the interaction of IQ with genetic risk. *Am. J. Psychiatry* **172**, 259–265 (2015).
33. Le Hellard, S. *et al.* Identification of gene loci that overlap between schizophrenia and educational attainment. *Schizophr. Bull.* **43**, 654–664 (2017).
34. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient metaanalysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
35. Peloso, G. M. *et al.* Phenotypic extremes in rare variant study designs. *Eur. J. Hum. Genet.* **24**, 924–930 (2016).
36. Purcell, S., Cherny, S. S. & Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
37. Coleman, J.R.I. *et al.* Biological annotation of genetic loci associated with intelligence in a metaanalysis of 87,740 individuals. *Mol. Psychiatry* **24**, 182–197 (2019).
38. König, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to include chromosome X in your genomewide association study. *Genet. Epidemiol.* **38**, 97–103 (2014).
39. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
40. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
41. Chang, C. C. *et al.* Secondgeneration PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
42. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genomewide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
43. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from highthroughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
44. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
45. Ernst, J. & Kellis, M. ChromHMM: automating chromatinstate discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
46. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
47. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
48. GTEx Consortium. The GenotypeTissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
49. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
50. Zhernakova, D. V. *et al.* Identification of contextdependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
51. Schmitt, A. D. *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
52. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
53. Skene, N.G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *bioRxiv*

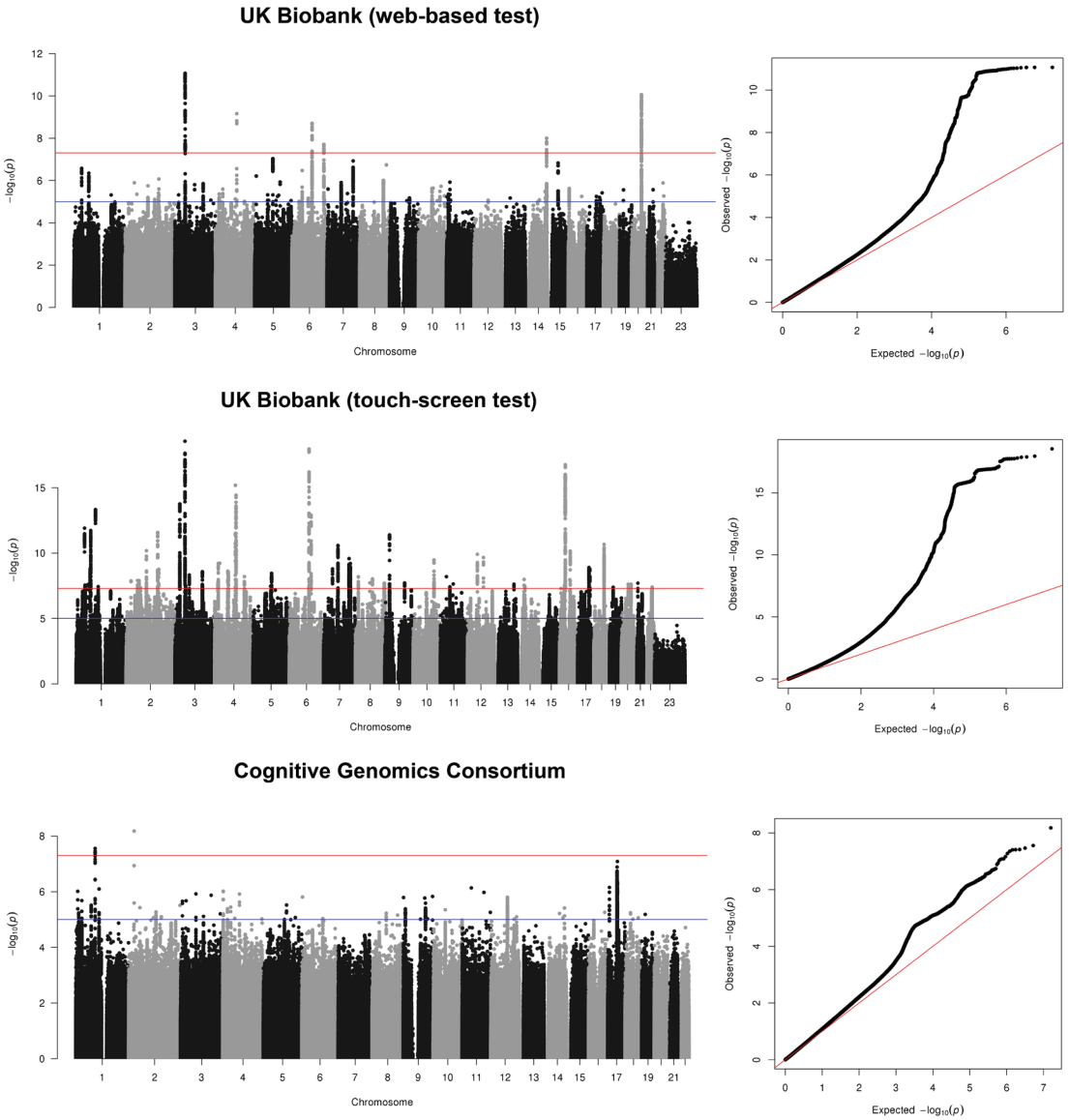
Supplementary information

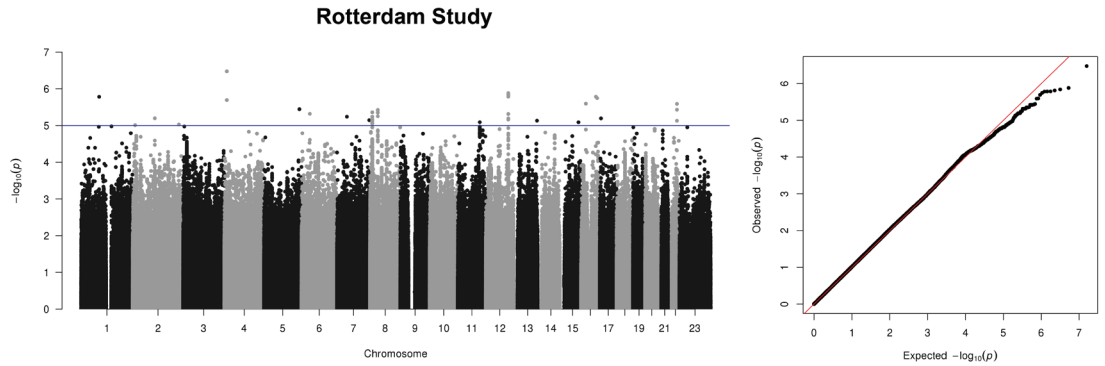
Supplementary information, figures (1-12) and tables (1-26) can be found in the online version of the manuscript:

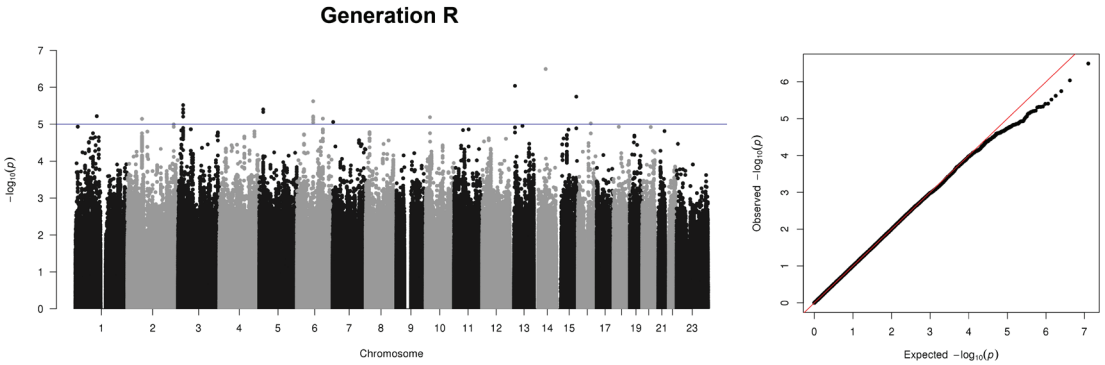
https://www.nature.com/articles/s41588-018-0152-6

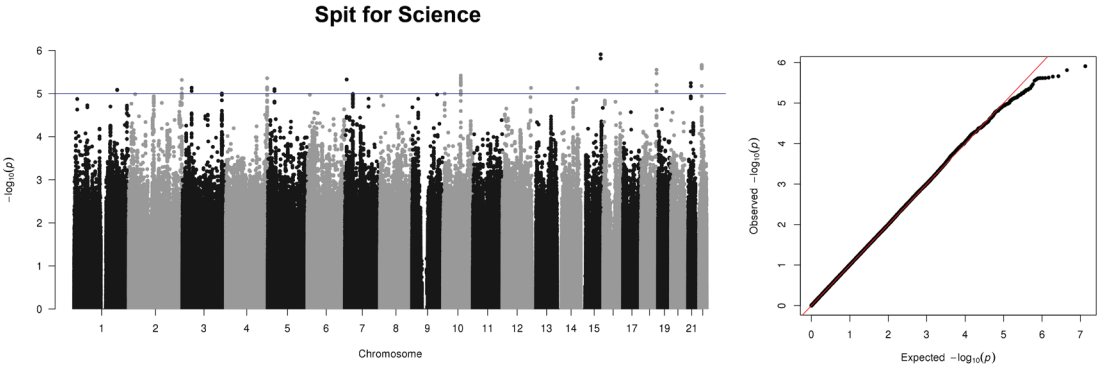
Supplementary information

Supplementary Fig. 1 | Manhattan and QQ plots of the individual cohort GWAS results included in a meta-analysis of intelligence in 269,867 independent individuals. For each of 14 cohorts included in the GWAS meta-analysis, $-\log_{10}$ transformed two-tailed *P*-values of SNP associations with intelligence measures in a linear or logistic regression model are presented against their chromosomal position in a Manhattan plot (left) and against expected null *P*-values in a QQ plot (right). Sample sizes and details of the statistical analyses for each cohort are presented in **Supplementary Information 1.1**. The dotted red line indicates Bonferroni-corrected genome-wide significance ($P < 5 \times 10^{-8}$), the blue line the threshold for suggestive associations ($P<1\times10^{-5}$).

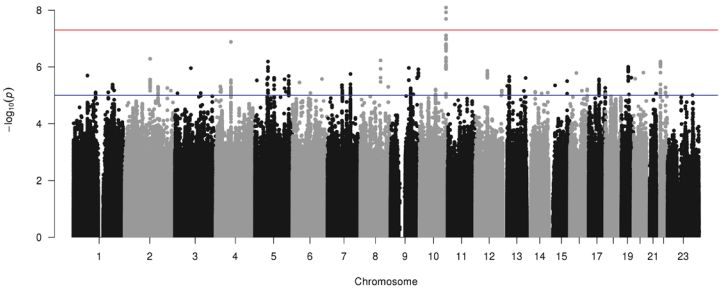


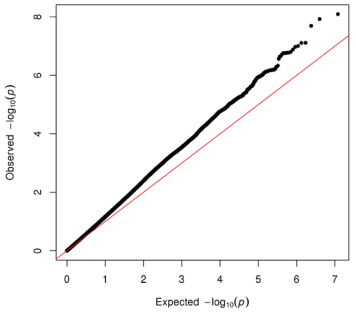

Rotterdam Study
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)


Generation R
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)

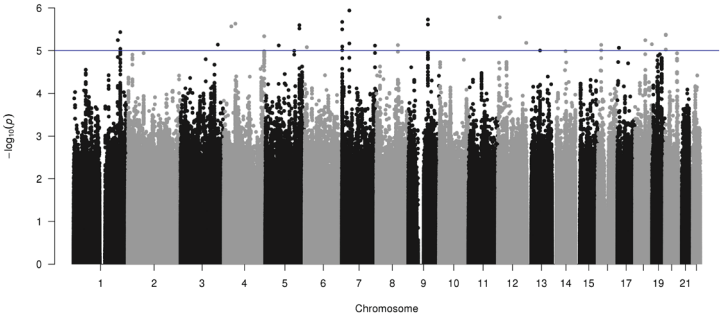

Spit for Science
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)

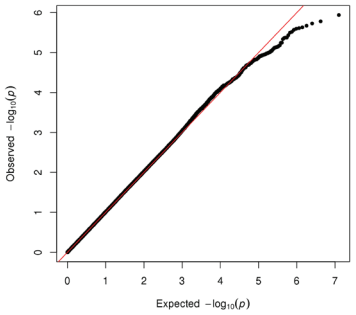
High IQ / Health and Retirement Study



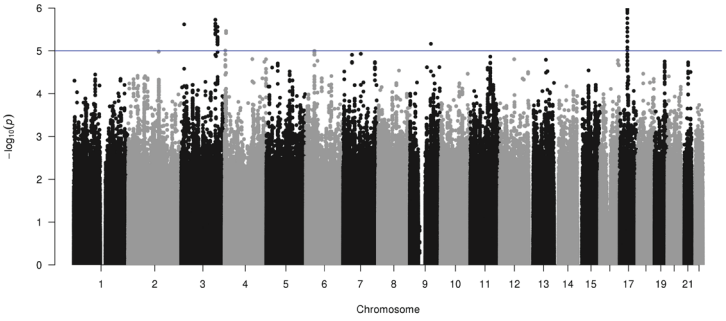


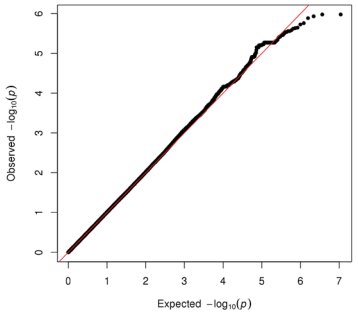
Twins Early Development Study 1

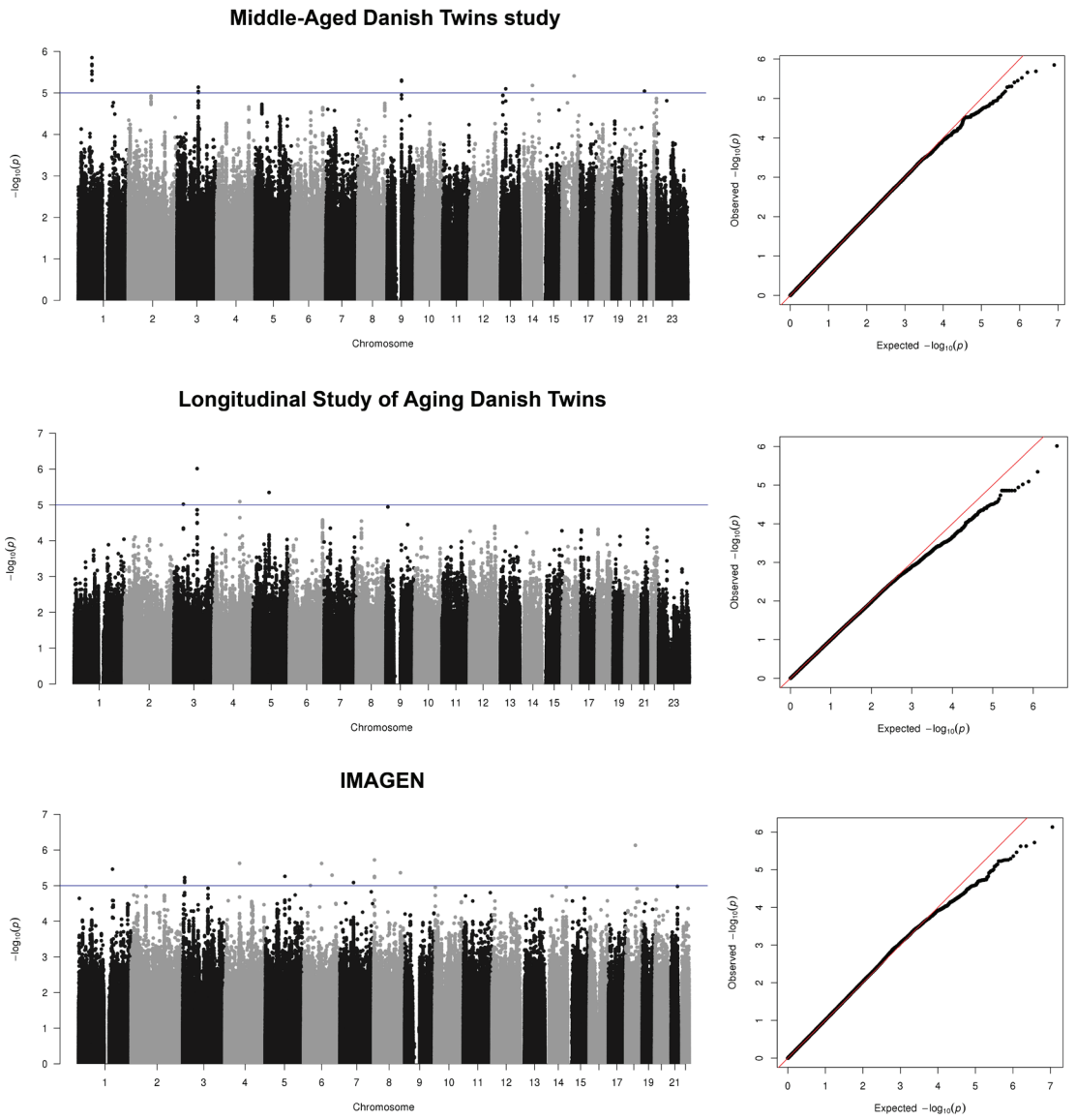


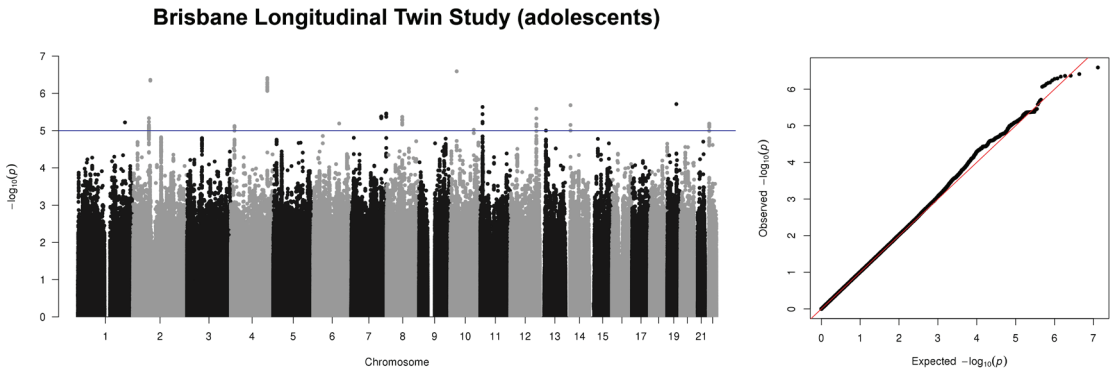


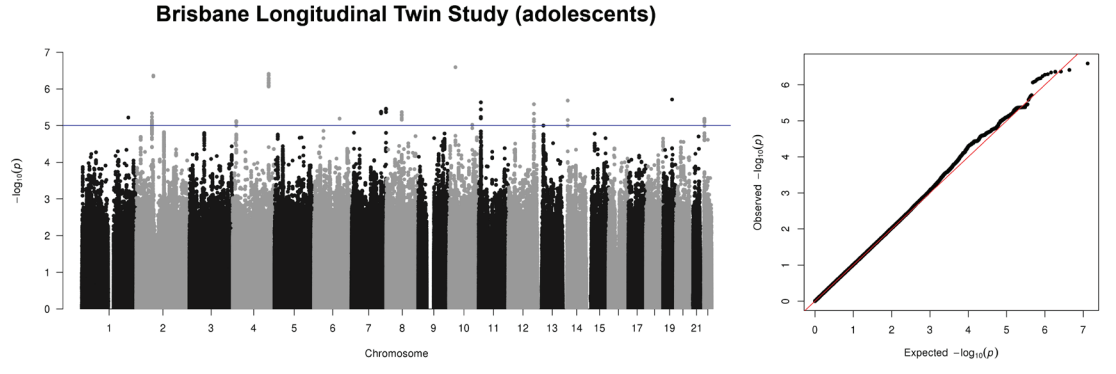
Twins Early Development Study 2

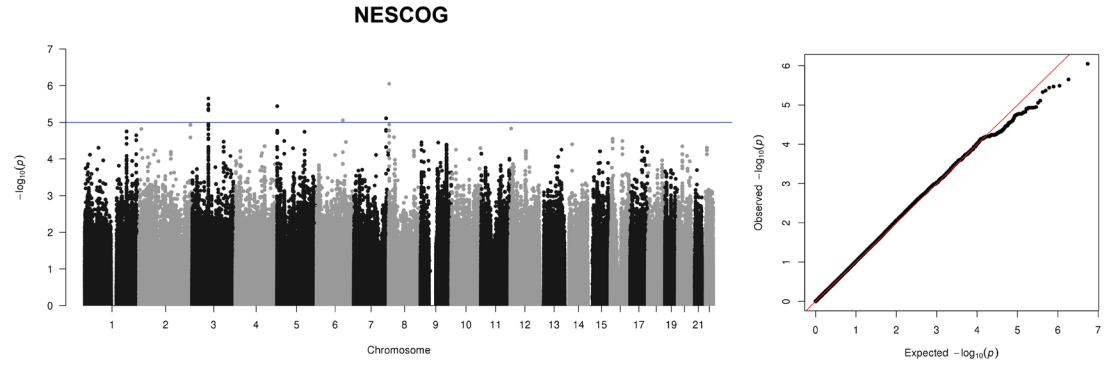


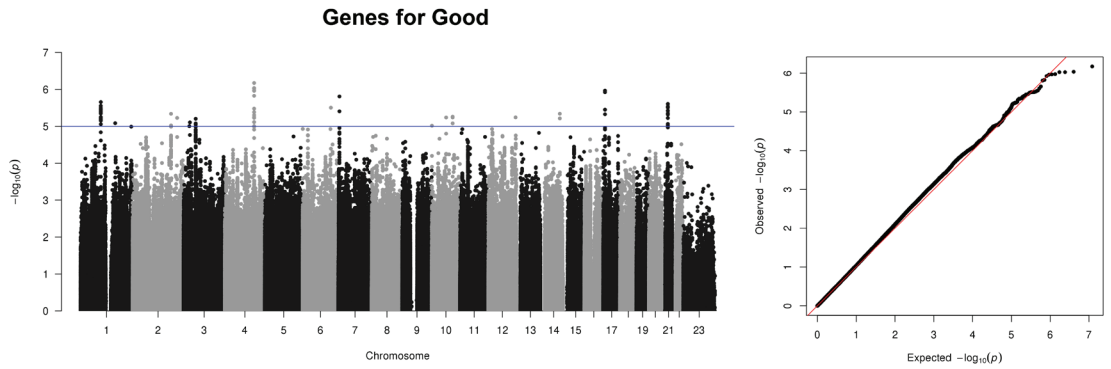


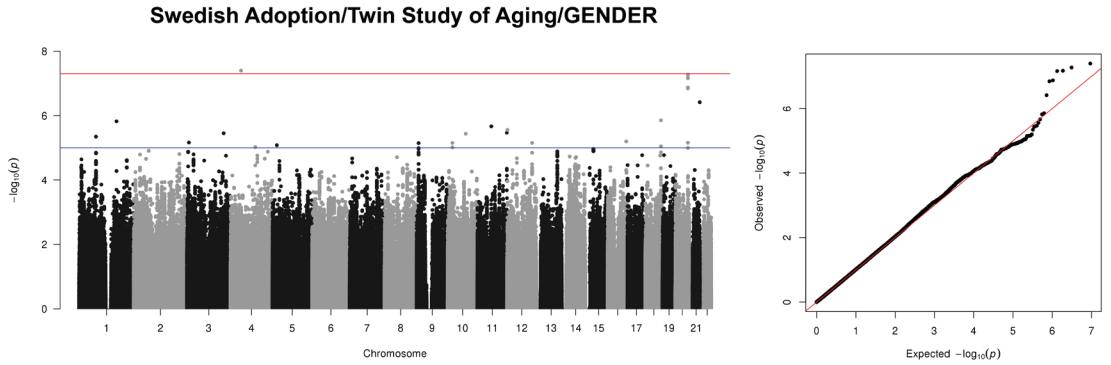

Middle-Aged Danish Twins study
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)
Longitudinal Study of Aging Danish Twins
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)
IMAGEN
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)

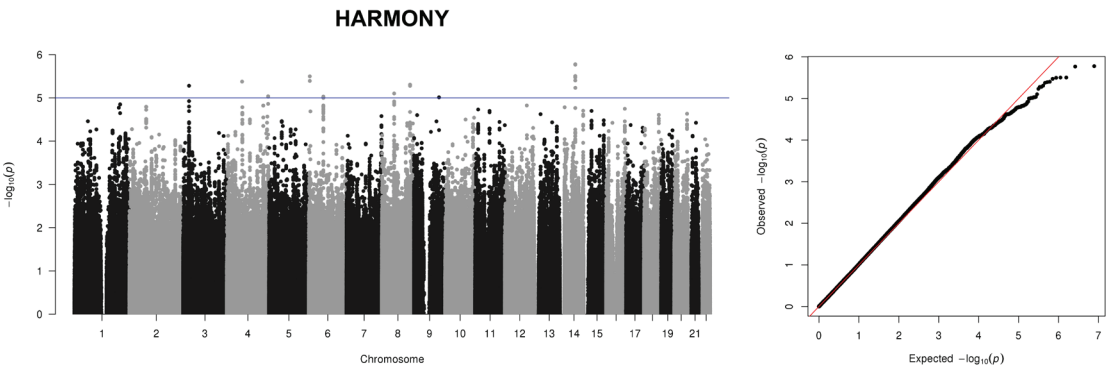

Brisbane Longitudinal Twin Study (adolescents)
−log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21
Observed −log10(p)
Expected −log10(p)


Brisbane Longitudinal Twin Study (adolescents)
−log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21
Observed −log10(p)
Expected −log10(p)

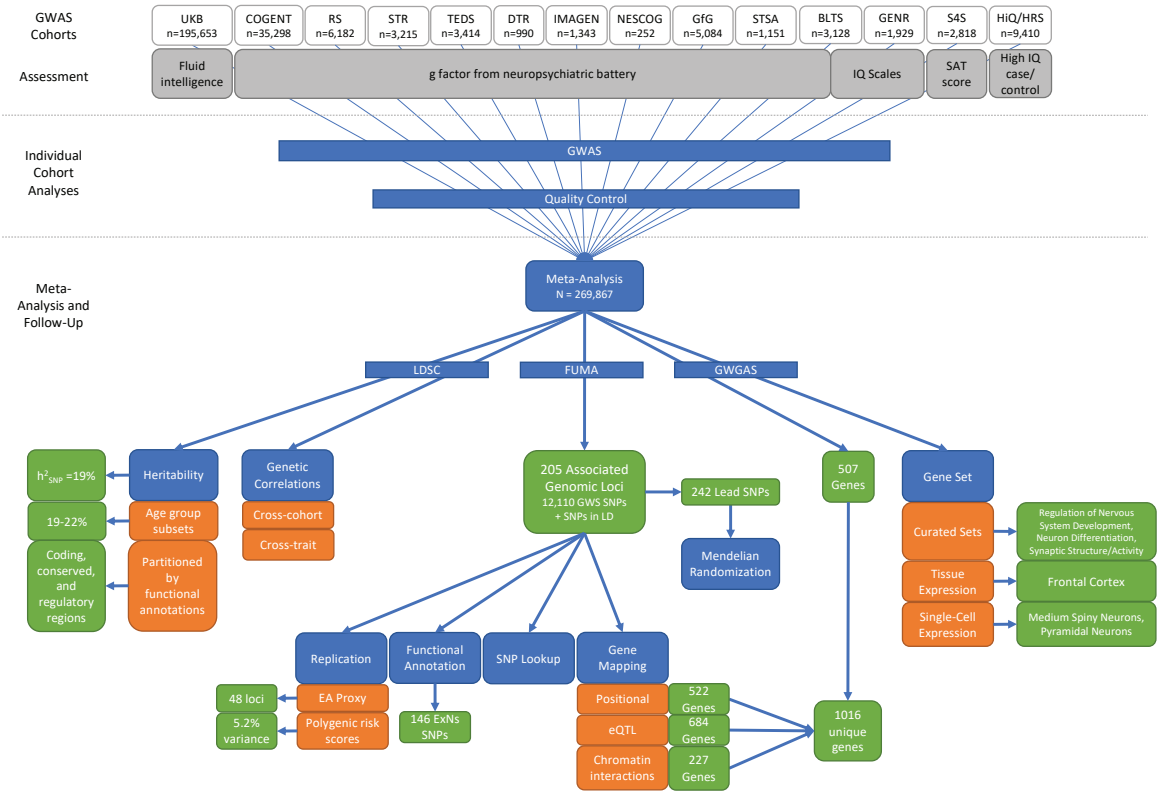

NESCOG
−log10(p)
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 19 21
Observed −log10(p)
Expected −log10(p)


Genes for Good
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)

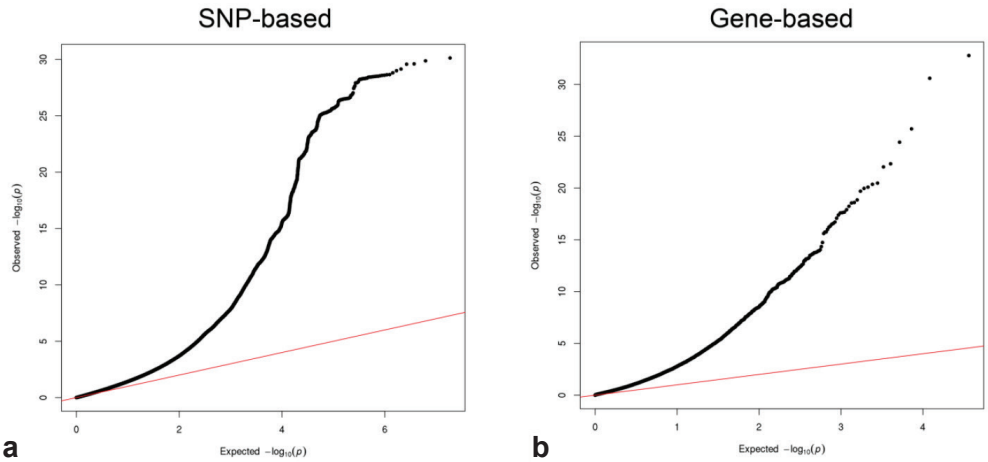

Swedish Adoption/Twin Study of Aging/GENDER
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)


HARMONY
−log10(p)
Chromosome
Observed −log10(p)
Expected −log10(p)

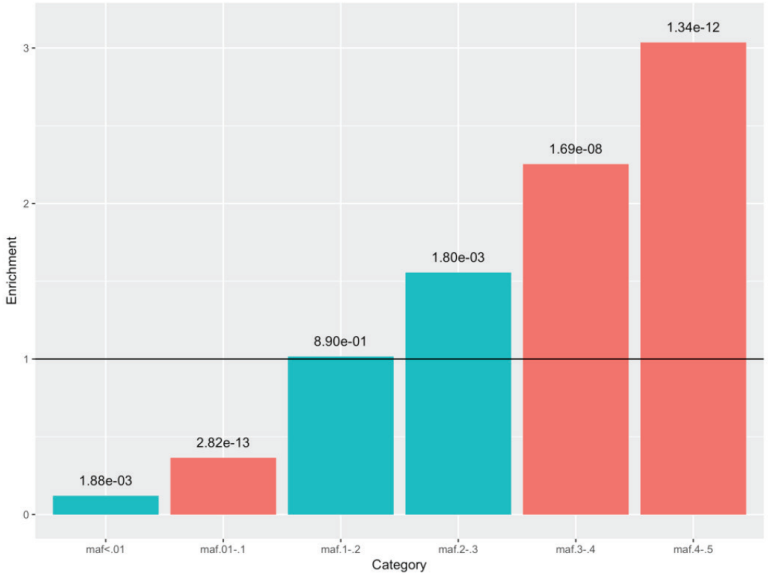
Supplementary Fig. 2 | Flowchart of study methodology and results of a GWAS meta-analysis of intelligence in 269,867 independent individuals. White boxes indicate study cohorts, grey boxes indicate assessment measures, blue boxes indicate groups of analytic procedures, orange boxes indicate sub-analyses, green boxes indicate results. GWAS = genome-wide association; LDSC = linkage disequilibrium (LD) score regression; FUMA = functional mapping and annotation; GWGAS = genome-wide gene-based association; SNP = single nucleotide polymorphism; EA = educational attainment; ExNS = exonic non-synonymous; eQTL = expression quantitative trait locus.



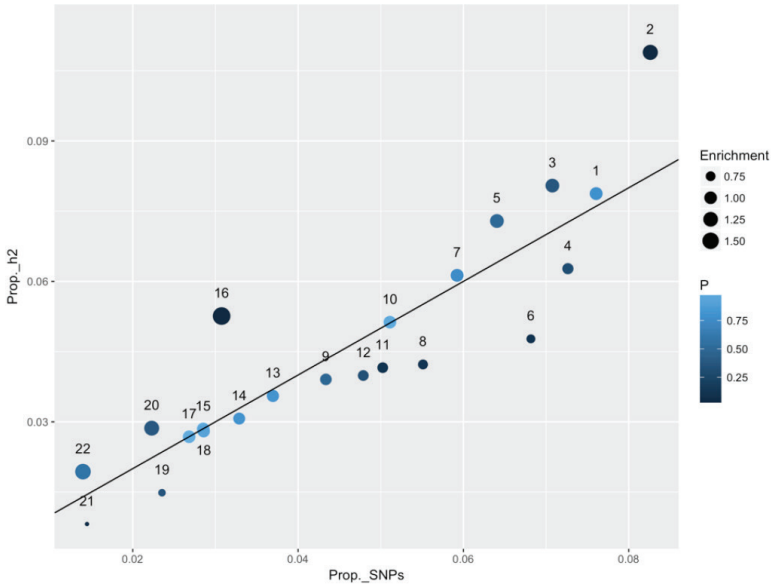
Supplementary Fig. 3 | QQ-plots of SNP and gene association results in a meta-analysis of intelligence in 269,867 independent individuals. Observed $-\log_{10}$ transformed two-tailed *P*-values of associations with intelligence measures are plotted against expected null *P*-values for **a)** all SNPs in the GWAS meta-analysis, and **b)** all genes in the gene-based meta-analysis.



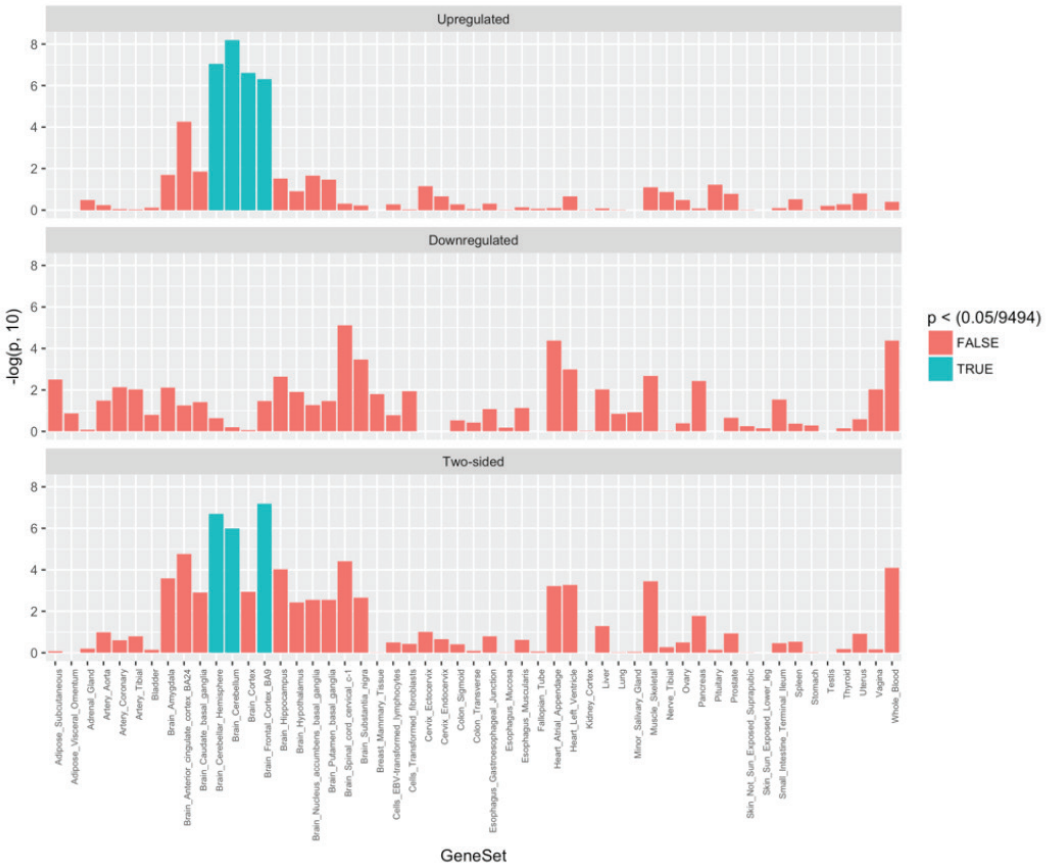
Supplementary Fig. 6 | Heritability of intelligence in a GWAS meta-analysis of 269,867 independent individuals, stratified by minor allele frequency (MAF) bins. Tests of enrichment/depletion of SNP heritability within a bin relative to the proportion of SNPs in the bin were conducted with LD score regression. Two-tailed *P*-values are presented above the bins; red bins are significant after Bonferroni correction for 56 total strata. Horizontal line (=1.0) indicates no enrichment.



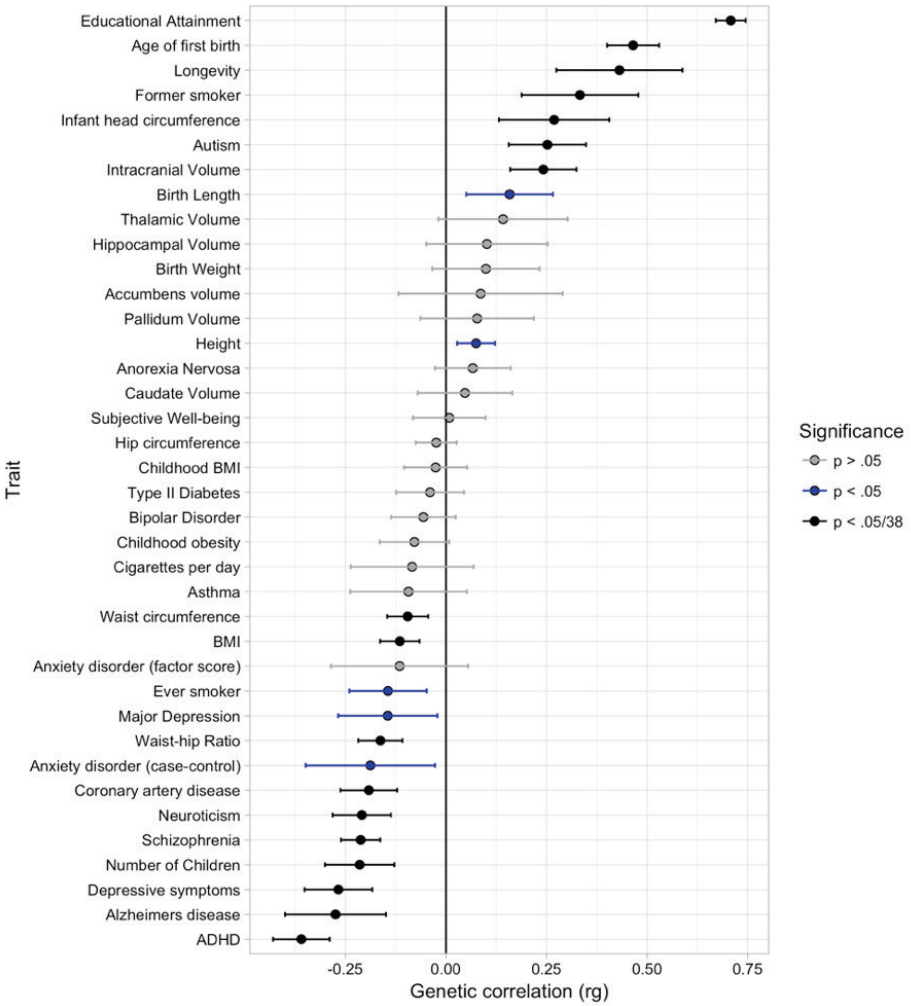
Supplementary Fig. 7 | Heritability of intelligence in a GWAS meta-analysis of 269,867 independent individuals, stratified by chromosome. Tests of enrichment/depletion of SNP heritability within a chromosome relative to the proportion of SNPs on the chromosome were conducted with LD score regression. The chromosome size (proportion of total SNPs) is on the x-axis and the proportion of total heritability (h^2) attributable to each chromosome is on the y-axis. The X-chromosome was not available in the reference panel for LD scores. Chromosomes are colored by two-tailed *P*-values; none were significant after Bonferroni correction for 56 total strata. Diagonal line indicates no enrichment.



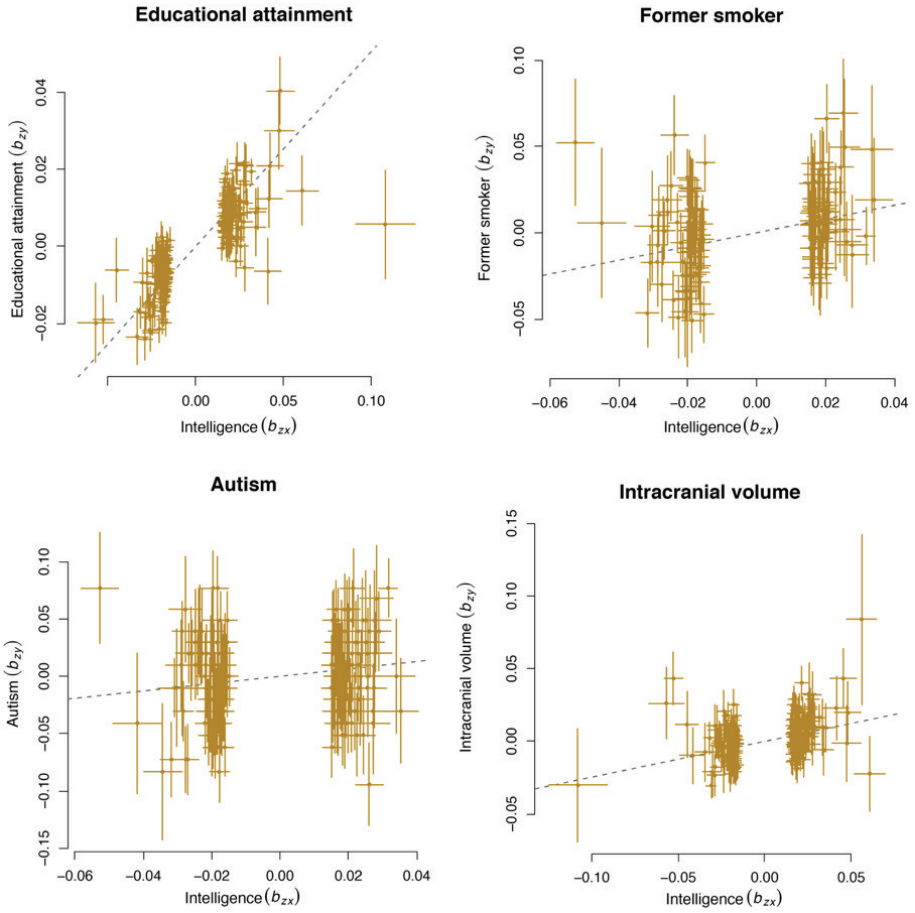
Supplementary Fig. 9 | Tissue enrichment for differential gene expression in genes associated with intelligence in a GWAS meta-analysis of 269,867 independent individuals. Hypergeometric tests of enrichment were conducted for 53 GTEx tissue types for 872 genes implicated by positional, eQTL, or chromatin interaction mapping of GWS SNPs in LD (r^2 >= 0.6) with one of the independent GWS SNPs. Enrichment difference are shown for higher (Upregulated), lower (Downregulated), or two-sided differences in gene expression. The tests do not include genes implicated only by GWGAS. Two-tailed *P*-values are presented; categories in blue are significant after Bonferroni correction for 9,494 total gene-sets.

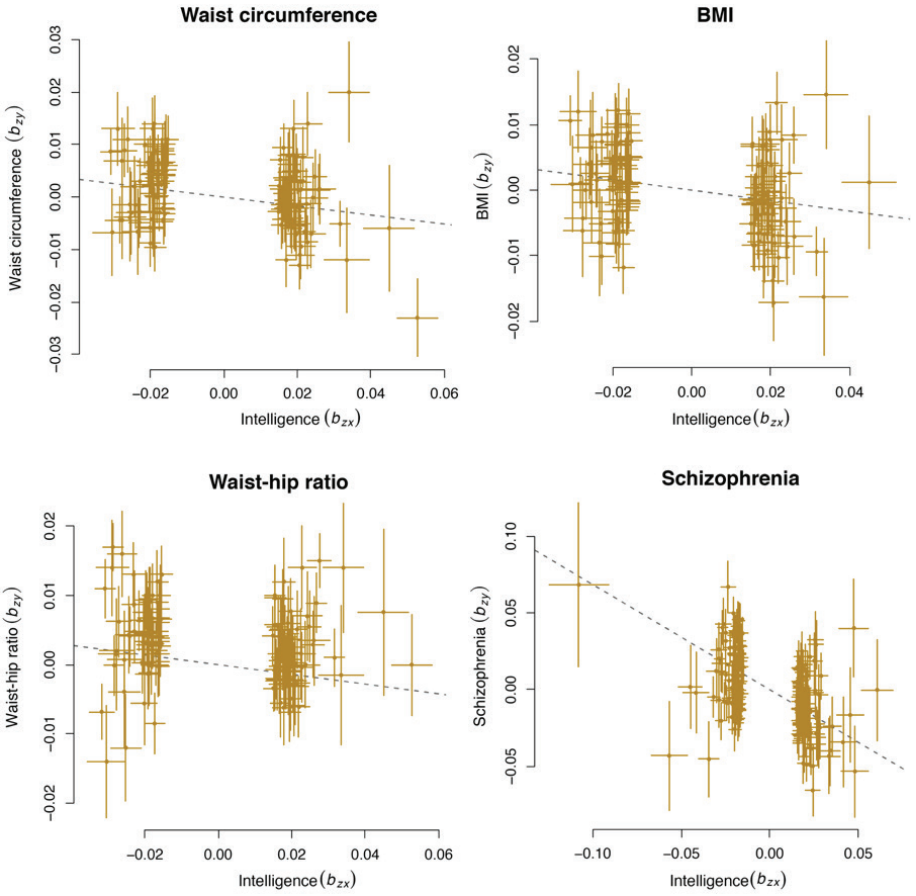


Supplementary Fig. 10 | Genetic correlations between intelligence and other traits previously investigated with GWAS. Correlations were calculated with LD score regression using SNP summary statistics from the GWAS meta-analysis of intelligence in 269,867 individuals and publically available summary statistics for other traits (**Supplementary Table 21**). Point estimates for correlations and 95% confidence intervals are shown; black dots indicate significant two-tailed *P*-values after Bonferroni correction for 38 pairs of traits.

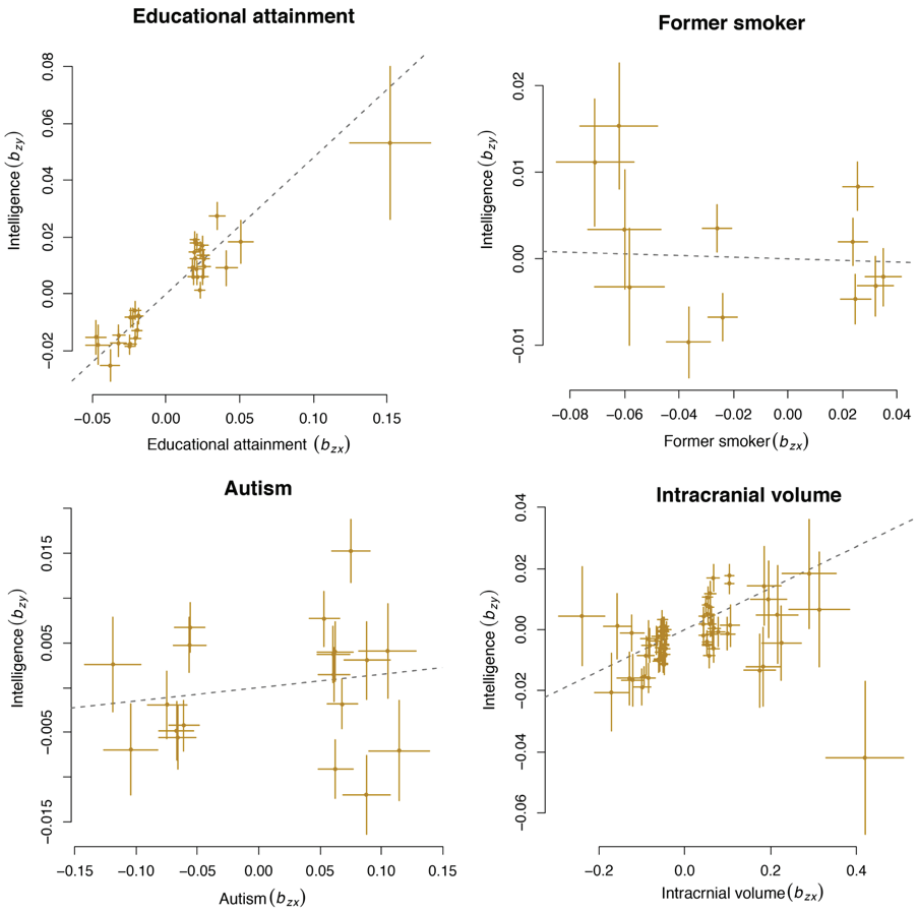


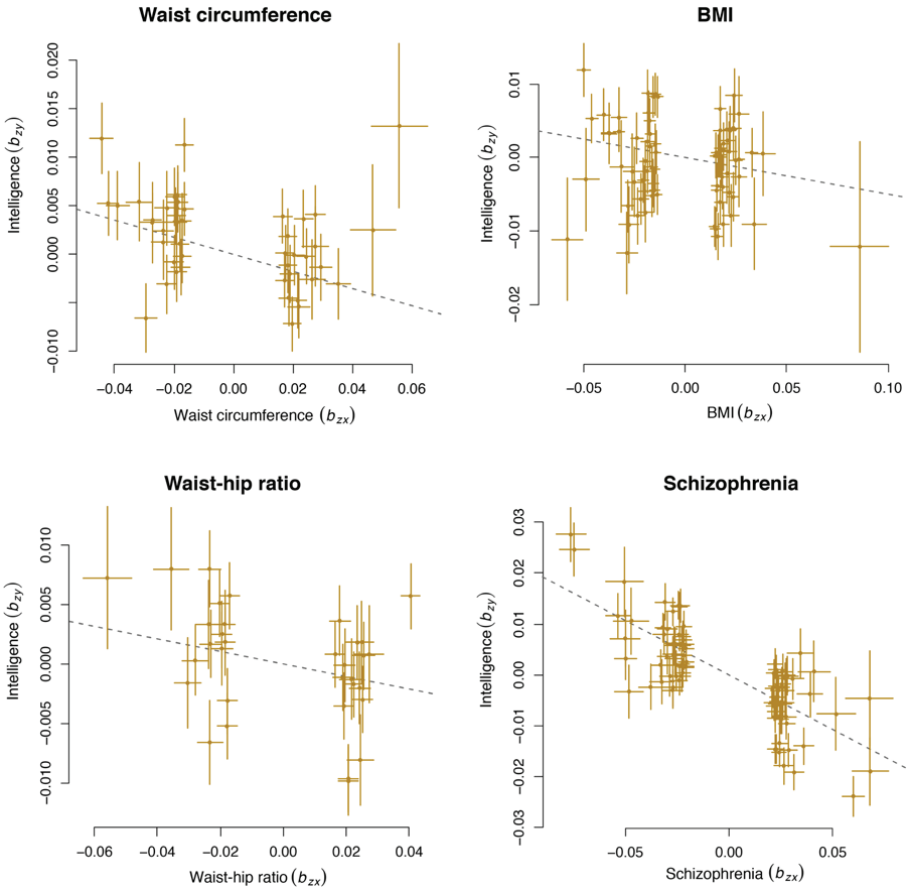
Supplementary Fig. 11 | Mendelian Randomization tests for the effect of intelligence on other correlated traits. Plots of effect sizes of independent lead SNPs from a GWAS meta-analysis of intelligence in 269,867 independent individuals (b_{zx}) on the x-axis and SNP GWAS effect sizes for correlated traits on the y-axis (b_{zx}). The dotted line represents a line with slope of (b_{xy}) and an intercept of 0. Error bars show 95% confidence intervals for the effect sizes for each trait.




Waist circumference
Waist circumference (b_{zy})
Intelligence (b_{zx})
BMI
BMI (b_{zy})
Intelligence (b_{zx})
Waist-hip ratio
Waist-hip ratio (b_{zy})
Intelligence (b_{zx})
Schizophrenia
Schizophrenia (b_{zy})
Intelligence (b_{zx})

Supplementary Fig. 12 | Mendelian Randomization reverse tests for the effect of other correlated traits on intelligence. Plots of effect sizes of independent lead SNPs from a GWAS meta-analysis of intelligence in 269,867 independent individuals (b_{zy}) on the y-axis and SNP GWAS effect sizes for correlated traits on the x-axis (b_{zx}). The dotted line represents a line with slope of (b_{xy}) and an intercept of 0. Error bars show 95% confidence intervals for the effect sizes for each trait.




Waist circumference
Intelligence (b_{zy})
0.020
0.015
0.010
0.005
0.000
−0.010
−0.04
−0.02
0.00
0.02
0.04
0.06
Waist circumference (b_{zx})
BMI
Intelligence (b_{zy})
0.01
0.00
−0.01
−0.02
−0.05
0.00
0.05
0.10
BMI (b_{zx})
Waist-hip ratio
Intelligence (b_{zy})
0.010
0.005
0.000
−0.005
−0.010
−0.06
−0.04
−0.02
0.00
0.02
0.04
Waist-hip ratio (b_{zx})
Schizophrenia
Intelligence (b_{zy})
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.05
0.00
0.05
Schizophrenia (b_{zx})

Part II:
Brain Imaging and Genetics

Chapter 6

Incidental Findings on Brain Imaging in the General Pediatric Population.

Philip R. Jansen, Marjolein Dremmen, Aaike van den Berg, Ilona A. Dekkers, Laura M.E. Blanken, Ryan L. Muetzel, Koen Bolhuis, Rosa H. Mulder, Desana Kocevska, Toyah A. Jansen, Marie-Claire de Wit, Rinze F. Neuteboom, Tinca J.C. Polderman, Danielle Posthuma, Vincent W.V. Jaddoe, Frank C. Verhulst, Henning Tiemeier, Aad van der Lugt, Tonya White

Published in abbreviated form in *New England Journal of Medicine*, 2017 Oct 19;377(16):1593-1595

Background: While the prevalence of incidental findings on brain magnetic resonance imaging (MRI) is well described in adult populations, not much is known about these findings in children. Here, we describe the prevalence of incidental findings on brain MRI in the general pediatric population.
Methods: Between April 2013 and November 2015, nine-to-twelve year-old children of the Generation R Study underwent MRI scanning of the brain. Scans were systematically reviewed for the presence of incidental findings, and prevalence estimates were derived. Cases were referred to the hospital when clinical follow-up or treatment was deemed necessary.
Results: A total of 3966 participants (mean age: 10.1 years, range: 8.6 - 11.9) underwent high resolution MRI and were evaluated for incidental findings. Incidental findings were observed in 1,015 children (25.6%). The most common findings were cysts of the pineal gland (diameter <1cm: 16.4%; >1 cm: 0.33%), arachnoid cysts (2.16%) and developmental venous anomalies (1.59%). A total of 17 children (0.43%) were referred to a pediatric neurologist for clinical imaging and followup, where indicated. Seven of these children (0.18%) had suspected primary brain tumors, of whom two children (0.05%) underwent surgical intervention.
Conclusions: Incidental findings requiring clinical referral were present in 1 out of 233 children, of whom approximately 1 in 3 cases involved suspected primary brain tumors. These results indicate that clinically relevant findings are present in the non-clinical, general population, and adequate managment protocols should be developed by researchers and clinicans that collect neuroimaging data in children.

Introduction

Incidental findings discovered on brain magnetic resonance imaging (MRI) in otherwise healthy individuals pose important medical and ethical considerations regarding the management of these findings[1-4]. It has been estimated that incidental findings requiring medical follow-up are found in approximately three percent of the adult population[5]. However, the current literature on incidental findings in children is limited, as most large-scale studies to date focus exclusively on adults. Although the prior probability of finding asymptomatic abnormalities is expected to be substantially lower in children compared to adults, the discovery of unfavorable anomalies at early age may subsequently lead to life-long and potentially preventable health consequences, and generate substantial levels of distress for the children and their parents[6,7]. Thus, obtaining more precise information on the frequency of brain incidental findings in children is important for well-informed medical decision making and tailoring adequate management protocols. Moreover, accurate information about the probability of finding clinically important abnormalities is necessary for informing families before participation in imaging studies, and to provide consistent information when incidental findings are disclosed.

Available data from brain MRI studies in children report varying prevalence rates between 0.3 to 4.4 percent of intracranial findings on MRI scans that require clinical follow-up[8-12]. Large variation in these estimates is due to a combination of differences in the applied threshold of considering findings as 'clinically relevant', the relatively small sample size of most studies that often cover a wide age range, the careful selection of healthy participants as controls for scientific research, and the image resolution of the collected data.

Here we report the prevalence of incidental findings on brain MRI in children from an unselected sample of the general population. For this purpose, we used a large population-based imaging study of nearly 4,000 nine-to-twelve year-old children who underwent brain MRI scanning. Such findings are expected to become increasingly important, as the utilization of MRI in children for both clinical and scientific purposes is expected to continue to expand in the coming years and the neuroimaging field moves into the era of 'big data'[13,14].

Methods

Population

The participants involved nine to twelve year old children from the Generation R Study, a large prospective population-based cohort investigating the development of children[15]. MRI scanning was performed between April 2013 and November 2015, and participants were invited when they were approximately 10 years of age. The exclusion criteria for the MRI study only included contraindications for MRI and claustrophobia. Children who fully completed the T1-weighted sequence were included in the current analyses. Informed consent for participation in the imaging study was obtained from the legally authorized representative of all study participants.

Brain Imaging Protocol

MRI scanning was preceded by a mock scan that simulated a real MRI scanner, allowing children to become accustomed to the scanning environment[16]. Study participants were scanned on a single wide-bore 3 Tesla

MR750W Discovery scanner (GE, Milwaukee, Wisconsin). For the current study, incidental findings evaluation was performed using three of the collected sequences.
First, a high-resolution 3D T1-weighted sequence was obtained using a coronal inversion recovery fast spoiled gradient recalled (IR-FSPGR, BRAVO) sequence (scan parameters: TR=8.77ms, TE=3.4ms, TI=600ms, flip angle=10°, field of view=220x220mm, acquisition matrix=220x220, slice thickness=1mm, number of slices=230, acceleration factor=2). A small number of scans were acquired with slightly different parameters due to small adjustments in the protocol: (n = 20 children, TR=7.82ms, TE=2.13ms, TI=600ms, flip angle=10°, field of view=220x220mm, acquisition matrix=220x220, slice thickness=1mm, number of slices=226, acceleration factor=3). Next, a structural 3D T2-weighted image was collected using a sagittal fast spin echo sequence (3D CUBE T2, TR=1440ms, TE=130ms, flip angle=90°, field of view=256x256mm, acquisition matrix=256x256, slice thickness = 1mm, number of slices =176, acceleration factor=3). In addition, a resting-state fMRI sequence (rs-FMRI) was collected using echo-planar imaging (EPI), with the following parameters: TR=2000ms, TI=350, TE=30ms, flip angle=85°, acquisition matrix=230x230, slice thickness=4mm.

Ratings of incidental findings on MRI
Brain images were reviewed for incidental findings by a group of researchers (PJ, KB, RM, DK, TJ) who were trained in evaluating scans by an experienced neuroradiologist (AvdL). Before rating MRI scans, raters had become experienced by rating a training set consisting of a large dataset of scans containing confirmed incidental findings. The rating protocol was designed to systematically evaluate brain scans and raters were instructed to report any finding that was seen on the scan. Imaging findings of potential clinical relevance, defined as imaging findings that could potentially lead to negative health consequences for the participant, were subjected to additional evaluation by a radiologist. First, the structural T1-weighted scan was reviewed in the coronal, sagittal and axial planes. Next, the T2-weighted image was assessed for findings not detected on the T1 scans. Finally, the first volume of the echo-gradient functional MRI sequence was evaluated. Evaluation of the MRI scans was performed using Synedra View Personal DICOM viewer, version 16.0.0 (Munich, Germany). No clinical information about the study participants was available to the raters at the time of assessment.

Clinical management
The protocol regarding the management of incidental findings was decided upon before initiation of the MRI study. In those cases where a potentially clinically relevant finding was found, the scans were reviewed by a neuroradiologist (AvdL, AvdB, MD) and discussed with a pediatric neurologist when necessary (MW). Subsequently, a multidisciplinary case-consensus decision was made whether clinical follow-up was indicated based on expert opinion.

Statistical Analyses
Baseline characteristics of the study population are expressed as mean (s.d.), range, or as a percentage. Prevalence rates were calculated with 95% confidence intervals (CI) to indicate the precision of the estimate, based on the total number of children that had completed the T1-weighed sequence.
All statistical analyses were performed using the R statistical software package, version 3.3.1[17].

Results

Sample characteristics
In total, 3,992 children participated in the brain imaging study of the Generation R cohort. Of these participants, 3966 (99.3%) had fully completed the T1-weighted sequence and were included in the current analyses. Sample characteristics of the study population are shown in **Table 1**. Participants were on average 10.12 years old (range: 8.55 – 11.99), and the sample consisted of an approximately equal distribution of gender (49,5% boys).

Overall & Clinical referral rates
A complete overview of the observed prevalence of the incidental findings is shown in **Table 2**. Of the total number of reviewed MRI scans, the vast majority (n = 2,951, 74.4%, 95% CI [73.0, 75.8]) did not show any intracranial incidental findings. In 1,015 children (25.6%, 95% CI [24.3, 27.0]) at least one incidental finding was found. In 75 participants (1.9%, 95% CI [1.5, 2.4]), two incidental findings were found, and two participants (0.05%, 95% CI [0.01, 0.20]) had three coexisting findings. Of the total sample, 17 children (0.43%, 95% CI [0.26, 0.70]), i.e. approximately 1 out of 233 children, were referred for clinical follow-up as a direct consequence of participation in the MRI study. All referred cases were discussed in multidisciplinary meetings, and clinical work-up often included neurological examination and an additional contrast-enhanced MRI scan (CE-MRI). One case showed imaging findings suspicious for central skull base fibrous dysplasia and was referred for clinical follow-up, includ-

Table 1 | Sample characteristics of the study population.

Characteristic	Total sample (n = 3,966)
Child	
Age of MRI (mean, range)	10.1 (8.6 – 11.9)
Gender (N boys, %)	1,963 (49.5)
Ethnicity	
Dutch (%)	61.7
Other Western (N, %)	8.5
Non-Western (N, %)	29.8
Gestational age (weeks) (mean, sd)	39.8 (1.9)
Child IQ (mean, sd)	102.5 (14.9)
Mother	
Maternal age at delivery (mean, range)	31.1 (15.3 – 46.3)
Maternal education	
Low (%)	5.8
Middle (%)	33.8
High (%)	60.4

ing additional CT-scanning, evaluation of the neuro-endocrine axis and genetic testing. None of the referred participants showed neurological symptoms at the time of clinical follow-up

Common incidental findings

Among the most common intracranial findings were pineal gland cysts (*n* = 665, 16.8%, 95% CI [15.6, 18.0]). The majority of the pineal gland cysts were smaller than 1 centimeter in diameter (*n* = 652, 16.4%, 95% CI [15.3, 17.6]), while only a small number (*n* = 13, 0.33%, 95% CI [0.18, 0.58]) was larger than 1 centimeter (largest diameter: 1.5 centimeter). Only one cyst had a solid component and was referred for subsequent CE-MR imaging, which did not show contrast enhancement. Arachnoid cysts were most often located in the middle cranial fossa and in the retro-cerebellar region and were referred for further clinical assessment in two cases (0.05%, 95% CI [0.01, 0.20]), as these exhibited a very large cyst with marked shifting of surrounding structures (**Figure 1A**).

Less common findings

Less frequent findings included a Chiari I malformation (*n* = 25, 0.63%, 95% CI [0.42, 0.94]), defined as tonsillar ectopia more than 5 millimeters below the foramen magnum18. One case with a large tonsillar herniation was referred for neurological evaluation (Figure 1B). An additional MRI scan in this subject did not reveal syrinx formation in the spinal cord, which commonly co-occurs with this abnormality.

Focal white matter damage, showing image characteristics that were most likely to be related to perinatal injury, was considered to be present in 7 participants (*n* = 7, 0.18%, 95% CI [0.08, 0.38]). One case was identified with lesions suggestive of demyelinating disease, classified as a Radiologically Isolated Syndrome[19]. Other infrequent findings were migration disorders (**Figure 1A**), including cortical dysplasia (*n* = 1, 0.03%, 95% CI [0.01, 0.16]) and subependymal heterotopia (*n*=19, 0.48%, 95% CI [0.30, 0.76]), partial agenesis of the corpus callosum (*n* = 2, 0.05%, 95% CI [0.01, 0.20]) (**Figure 1C**), and partial agenesis of the septum pellucidum (*n* = 3, 0.08%, 95% CI [0.02, 0.24]),

Incidental neoplasms

Incidental brain tumors were observed in 7 participants (0.18%, 95% CI [0.08, 0.38]) (**Figure 2A-D**). In two cases, after a detailed clinical assessment and CE-MRI, direct neurosurgical intervention was warranted due to the location, signal characteristics and mass effect on the surrounding structures. In the first case, a large heterogeneous lobulated mass was located near the temporal horn of the lateral ventricle that was histopathologically proven to be an ependymoma (WHO Grade II) (**Figure 2A**). In the second case, a tumor was found in the suprasellar region, which was later classified by histopathological examination as a craniopharyngeoma (WHO Grade I) (**Figure 2B**). In the other five cases, no surgical interventions were performed and the patients are followed up with

Table 2 | Overview of incidental findings in the study population (n=3,966).

Category	Finding	N cases	Prevalence (%) [95%CI]	Clinical referral	Clinical Management
Normal variations	Cavum septum pellucidum	79	1.99 [1.59, 2.49]	none	-
	Mega cisterna magna	104	2.62 [2.16, 3.18]	none	-
	Empty sella configuration	7	0.18 [0.08, 0.38]	none	-
Congenital malformations	Chiari I malformation	25	0.63 [0.42, 0.94]	n=1	MRI follow-up
	Partial agenesis corpus callosum	2	0.05 [0.01, 0.20]	n=2	Neurological examination
	Septum pellucidum agenesis	3	0.08 [0.02, 0.24]	none	-
	Ventriculomegaly	2	0.05 [0.01, 0.20]	n=1	MRI follow-up
Cysts	Arachnoid cyst	86	2.17 [1.75, 2.68]		-
	< 3 cm	75	1.89 [1.50, 2.38]	none	-
	> 3 cm	11	0.28 [0.15, 0.51]	n=2	MRI follow-up
	Pineal gland cyst	665	16.8 [15.6, 18.0]		
	< 1 cm	652	16.4 [15.3, 17.6]	none	-
	> 1 cm	13	0.33 [0.18, 0.58]	N=1	CE-MRI, lumbar puncture
	Porencephalic cyst	3	0.08 [0.02, 0.24]	none	-
	Intraventricular cysts	7	0.18 [0.08, 0.38]	N=1	MRI follow-up
Vascular anomalies	Developmental venous anomaly	63	1.59 [0.12, 2.04]	none	-
	Cavernous angioma	7	0.18 [0.08 0.38]	none	-
	Capillary teleangiectasia	2	0.05 [0.01, 0.20]	none	-
Migration disorders	Supependymal gray matter heterotopia	19	0.48 [0.30, 0.76]	none	-
	Transmantle dysplasia	1	0.03 [0.01, 0.16]	none	-
	Focal cortical dysplasia	1	0.03 [0.01, 0.16]	none	-
White-matter abnormalities	Focal white matter hyperintensity	7	0.18 [0.08, 0.38]	none	-
	Radiological Isolated Syndrome	1	0.03 [0.01, 0.16]	n=1	CE-MRI
Neoplasms	Low-grade glioma[a]	4	0.10 [0.03, 0.28]	n=4	CE-MRI
	DNET[a]	1	0.03 [0.01, 0.16]	n=1	CE-MRI
	Ependymoma[b]	1	0.03 [0.01, 0.16]	n=1	CE-MRI, neurosurgery
	Craniopharyngeoma[b]	1	0.03 [0.01, 0.16]	n=1	CE-MRI, neurosurgery
Other	Fibrous Dysplasia	1	0.03 [0.01, 0.16]	n=1	CT

[a] = radiological diagnosis, [b] = confirmed by histopathology, CE-MRI = contrast-enhanced MRI

additional CE-MRI scans. Clinical follow-up period in these children ranged between 1.96 to 3.12 years. Of the five children who did not undergo surgical intervention, one child with a presumed dysembryoplastic neuroepithelial tumor (DNET) developed epileptic symptoms that were adequately controlled by pharmacological treatment (**Figure 2D**). None of the tumors in these five children showed signs of growth, and no subsequent surgical treatment was performed until the last follow-up.

Discussion

We reported the prevalence of incidental findings in the largest single-site brain imaging study of the general pediatric population to date. Our results showed a relatively high prevalence of a variety of common findings in approximately 1 out of 4 (25.6%) children, although the prevalence of findings requiring clinical follow-up was much lower at 1 in 233 (0.43%) children.

The low prevalence of intracranial findings requiring clinical work-up of 0.43 percent is in line with previous studies in children (overview of previous studies in **Table 3**), although the present rates are relatively lower than those previously published. Reported referral rates range between 0.48 and 0.90 percent in samples from the general population[11,12], and between 0.3 and 3.2 percent in clinical studies[9,10,20]. Out of 3,966 included children, 17 participants had findings that after careful multidisciplinary evaluation were considered potentially clinically relevant. From the referred cases, subsequent clinical management in 15 children included additional radiological imaging,

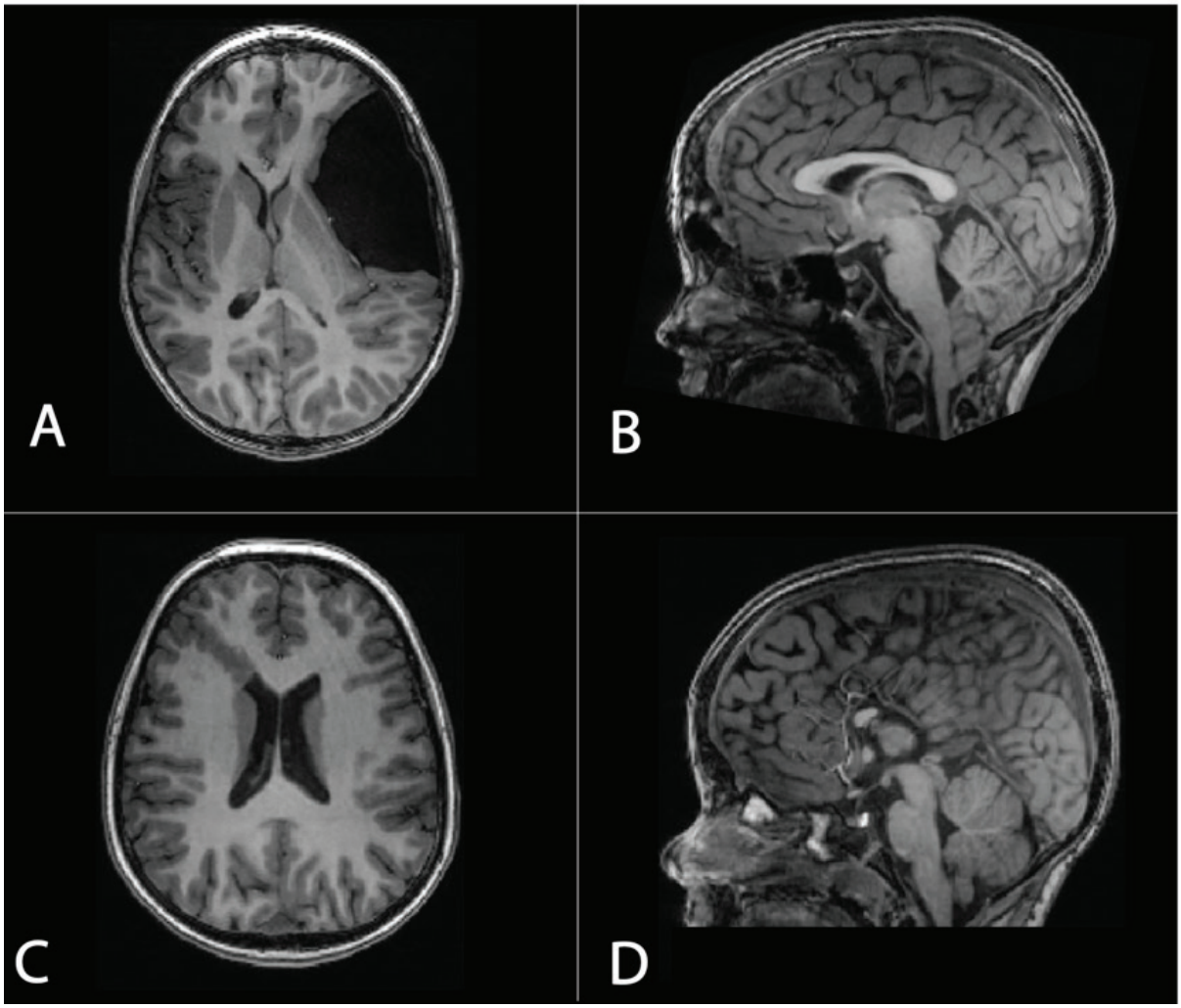


Fig. 1 | Incidental findings observed in the study population. A) Large arachnoid cyst of the temporal lobe with mass-effect on the basal ganglia of the left hemisphere; **B)** Cascading tonsillar tissue herniating through the foramen magnum, Chiari I malformation; **C)** Isolated abnormal gyral patterning and cortical thickening extending towards the lateral ventricle, focal cortical dysplasia; **D)** Partial corpus callosum agenesis with only remnant rostral and genu.

mainly consisting of additional (contrast enhanced) MRI scanning. In the other two cases, subsequent neurological examination was considered sufficient and no further clinical follow-up was indicated. Two main factors may contribute to a low prevalence of clinically relevant findings in our sample. First, our study population consisted entirely of individuals from an unselected population-based sample with a correspondingly low likelihood of carrying asymptomatic abnormalities of clinical importance. Second, considering that our study sample consisted of children from a non-clinical study sample, we maintained a high threshold for referring participants for clinical follow-up on a case consensus basis, given the higher probability of false positive findings leading to unnecessary distress for children and their parents[21].

Importantly, we observed an unexpected high number of brain tumor cases in our study cohort. Our population-based sample included two children (0.05%) with an asymptomatic histologically confirmed primary brain tumor, whereas five other children showed imaging findings consistent with the radiological diagnosis of a primary brain tumor (0.13%). These numbers are higher than the prevalence estimated from cancer registries of symptomatic primary brain tumors that have reported a prevalence of approximately 3 in 10,000 (0.003%) in the US for individuals below 19 years of age[22]. However, in the

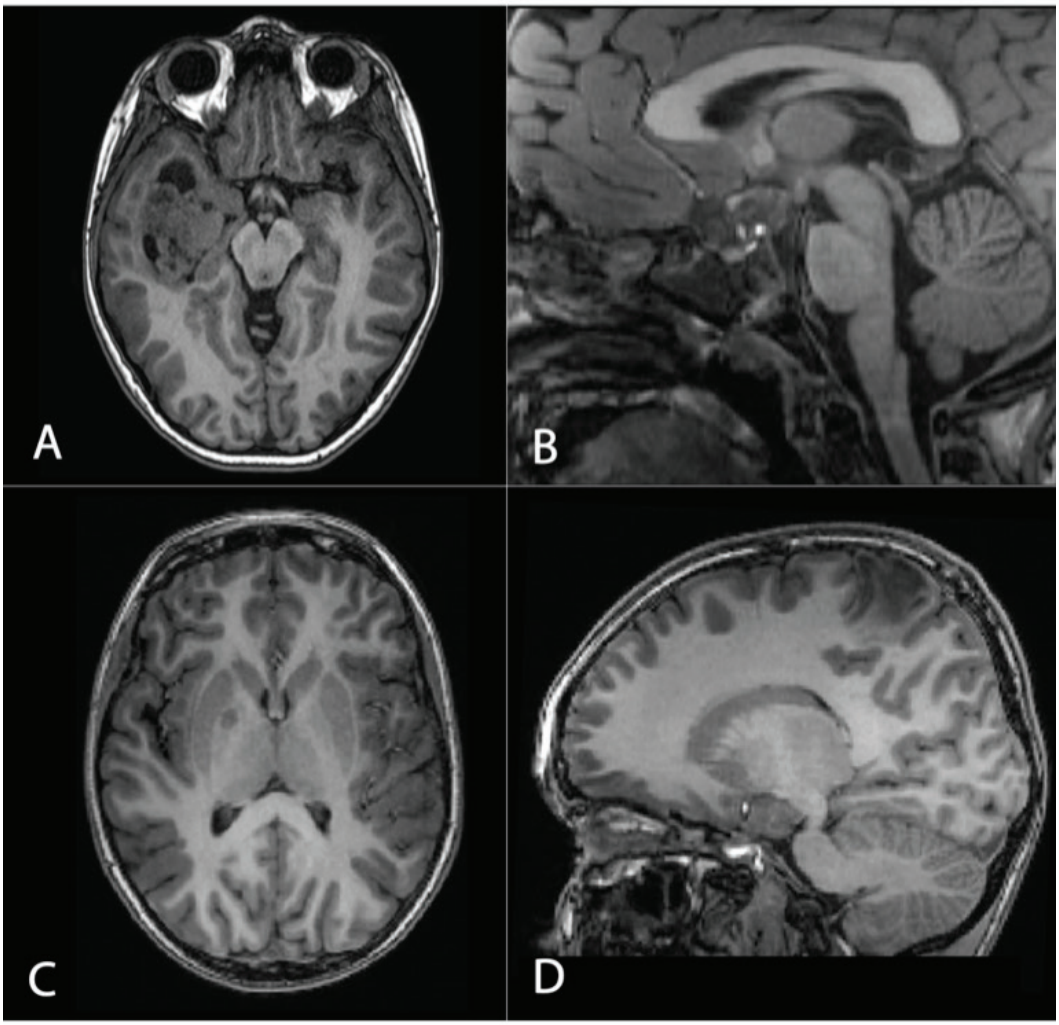


Fig. 2 | Incidental brain tumors. A) Mixed cystic and solid heterogeneous mass in the temporal horn of the lateral ventricle, histologically proven ependymoma (WHO Grade II); **B)** Suprasellar lesion, with calcifications and cystic components with T1 shortening corresponding to proteinaceous contents, craniopharyngeoma; **C)** Focal well defined small hypo-intense lesion in the globus pallidus, suspected low-grade glioma; **D)** Well-defined, cortically based lesion with typical lobulated internal architecture, dysembryoplastic neuroepithelial tumour (DNET).

current pediatric literature no reliable statistics are available of asymptomatic brain tumors in children[13]. The relatively high rate of asymptomatic primary brain tumors suggests that these are possibly more common in children than estimated from clinical cases, but can stay asymptomatic for a longer period of time. Longitudinal studies are necessary to provide the course of asymptomatic tumors in these individuals, as these results would aid in the decision whether or not these lesions should be treated. Absence of clinical symptoms or tumor growth during follow-up in the cases in our sample supports a watchful-waiting policy over immediate surgical treatment of asymptomatic low-grade gliomas[23,24]. In contrast, several findings were not observed that may have been expected given their prevalence and our large sample size, such as arteriovenous malformations[25], pituitary adenomas[26], and vascular aneurysms[27].

There is an increasing need for standardized protocols for incidental findings management in children, including standardized reporting, disclosure to parents and subsequent follow-up when deemed necessary[28,29]. Considering the fact that we found serious incidental abnormalities, including primary brain tumors, that could have detrimental consequences if not referred timely, we recommend

the development of systematic image evaluation protocols in close collaboration with radiologists and neurologists. The responsibility of creating the protocols and necessary infrastructure for the detection and follow-up of incidental findings should be carried out carefully by researchers before neuroimaging studies in children are initiated[30].

The current study has several strengths. Our study population comprised a community-based sample that reflects the healthy pediatric population at large. Moreover, the current study has an adequate sample size for estimating the prevalence of frequent, as well as less frequently observed incidental findings. There are also several limitations. Primary evaluation of the images were performed by trained researchers rather than neuroradiologists, which could potentially lead to underreporting of subtle abnormalities. In addition, the MRI protocol was optimized for scientific research related to brain development. Sequences developed primarily for the detection of brain abnormalities, such as T_2 fluid-attention inverse recovery (T_2 FLAIR), might have increased the detection rate of abnormalities, such as focal or diffuse supratentorial white matter intensities.

In conclusion, our study provides insight in the prevalence of incidental findings on brain MRI in the general pediatric population. These results emphasize that careful evaluation of incidental findings on brain scans of asymptomatic children is warranted. Future studies in children should aim to extend this follow-up period and include longitudinal data to determine whether asymptomatic findings may lead to symptoms over time and whether treatment of these findings should be preferred, as was recently done for the elderly population[31].

Table 3 | Overview of previous studies that reported intracranial incidental findings in children.

Study population	Study (ref)	N	Age range	Exclusion criteria	Findings*	Referred cases
Clinical population	Gupta *et al.* (9)[a]	666	0 – 21	Prematurity	Total: 25.7% Referral: 0.3%	Arachnoid cyst, venous malformation
	Jordan *et al.* (10)[b]	953	8 – 14	Stroke	Total: 6.6% Referral: 3.2%	Chiari I with syringomyelia, primary brain tumors
	Gur *et al.* (20)[c]	1,400	8 – 23	General health issues	Total: 10.6% Referral: 0.85%	
Healthy controls	Kim *et al.* (8)[d]	225	0 – 18	Neurological, developmental or psychiatric disorders	Total: 7.56% Referral: 4.44%	Focal white matter lesion, tonsillar ectopia, hypoplasia pons, cerebellar tonsil lesion
General population	Seki *et al.* (11)	110	5 – 8	N.S.	Total: 9.1% Referral: 0.9%	Cervical syringomyelia
	Sullivan *et al.* (12)	833	12 – 21	Alcohol / drug use criteria	Total: 11.8% Referral: 0.48%	Cranio-cervical stenosis, parietal cortical mass, Chiari I malformation, demyelinating disorders
General population	Current study	3,966	8 – 12	None	Total: 25.6% Referral: 0.43%	

* = Prevalence estimates from these studies are intracranial findings and do not cover extra-cranial abnormalities (e.g. sinusitis), N.S = not stated, CC = corpus callosum, RIS = radiological isolated syndrome, a = tertiary pediatric neurology clinic, b = sickle cell patients, c = sampled from individuals who underwent clinical care, d = healthy controls for fMRI studies

References

1. Illes, J. *et al.* Ethical consideration of incidental findings on adult brain MRI in research. *Neurology* **62**, 888–890 (2004).
2. Wolf, S. M. *et al.* Managing incidental findings in human subjects research: analysis and recommendations. *J. law, Med. ethics* **36**, 219–248 (2008).
3. Booth, T. C., Jackson, A., Wardlaw, J. M., Taylor, S. A. & Waldman, A. D. Incidental findings found in "healthy" volunteers during imaging performed for research: current legal and ethical implications. *Br. J. Radiol.* **83**, 456–465 (2010).
4. Kaiser, D., Leach, J., Vannest, J., Schapiro, M. & Holland, S. Unanticipated findings in pediatric neuroimaging research: Prevalence of abnormalities and process for reporting and clinical follow-up. *Brain Imaging Behav.* **9**, 32–42 (2015).
5. Morris, Z. *et al.* Incidental findings on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* **339**, b3016 (2009).
6. Schmidt, C. O. *et al.* Psychosocial consequences and severity of disclosed incidental findings from whole-body MRI in a general population study. *Eur. Radiol.* **23**, 1343–1351 (2013).
7. White, T. The ethics are in the numbers: A bayesian approach

to the management of incidental fi ndings in pediatric magnetic resonance imaging research. *Am. J. Case Rep.* **10**, 22–26 (2009).
8. Kim, B. S., Illes, J., Kaplan, R. T., Reiss, A. & Atlas, S. W. Incidental findings on pediatric MR images of the brain. *Am. J. Neuroradiol.* **23**, 1674–1677 (2002).
9. Gupta, S. N. & Belay, B. Intracranial incidental findings on brain MR images in a pediatric neurology practice: a retrospective study. *J. Neurol. Sci.* **264**, 34–37 (2008).
10. Jordan, L. C. *et al.* Incidental findings on brain magnetic resonance imaging of children with sickle cell disease. *Pediatrics* **126**, 53–61 (2010).
11. Seki, A. *et al.* Incidental findings of brain magnetic resonance imaging study in a pediatric cohort in Japan and recommendation for a model management protocol. *J. Epidemiol.* **20**, S498–S504 (2010).
12. Sullivan, E. V *et al.* Structural brain anomalies in healthy adolescents in the NCANDA cohort: relation to neuropsychological test performance, sex, and ethnicity. *Brain Imaging Behav.* **11**, 1302–1315 (2017).
13. Maher, C. O. & Piatt, J. H. Incidental findings on brain and spine imaging in children. *Pediatrics* **135**, e1084–e1096 (2015).
14. Van Horn, J. D. & Toga, A. W. Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* **8**, 323–331 (2014).
15. Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31**, 1243–1264 (2016).
16. White, T. *et al.* Pediatric population-based neuroimaging and the Generation R Study: the intersection of developmental neuroscience and epidemiology. *Eur. J. Epidemiol.* **28**, 99–111 (2013).
17. R Core Team, A language and environment for statistical computing. (2013).
18. Elster, A. D. & Chen, M. Y. Chiari I malformations: clinical and radiologic reappraisal. *Radiology* **183**, 347–353 (1992).
19. Okuda, D. T. *et al.* Incidental MRI anomalies suggestive of multiple sclerosis: the radiologically isolated syndrome. *Neurology* **72**, 800–805 (2009).
20. Gur, R. E. *et al.* Incidental findings in youths volunteering for brain MRI research. *Am. J. Neuroradiol.* **34**, 2021–2025 (2013).
21. Illes, J. *et al.* Incidental findings in brain imaging research. *Science* **311**, 783-784 (2006).
22. Porter, K. R., McCarthy, B. J., Freels, S., Kim, Y. & Davis, F. G. Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology. *Neuro. Oncol.* **12**, 520–527 (2010).
23. Bredlau, A.-L., Constine, L. S., Silberstein, H. J., Milano, M. T. & Korones, D. N. Incidental brain lesions in children: to treat or not to treat? *J. Neurooncol.* **106**, 589–594 (2012).
24. Perret, C., Boltshauser, E., Scheer, I., Kellenberger, C. J. & Grotzer, M. A. Incidental findings of mass lesions on neuroimages in children. *Neurosurg. Focus* **31**, E20 (2011).
25. Di Rocco, C., Tamburrini, G. & Rollo, M. Cerebral arteriovenous malformations in children. *Acta Neurochir. (Wien).* **142**, 145–158 (2000).
26. Ezzat, S. *et al.* The prevalence of pituitary adenomas: a systematic review. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* **101**, 613–619 (2004).
27. Brown Jr, R. D. & Broderick, J. P. Unruptured intracranial aneurysms: epidemiology, natural history, management options, and familial screening. *Lancet Neurol.* **13**, 393–404 (2014).
28. van der Lugt, A. Incidental findings on brain magnetic resonance imaging. *BMJ* b3107(2009).
29. Borra, R. J. H. & Sorensen, A. G. Incidental findings in brain MRI research: what do we owe our subjects? *J. Am. Coll. Radiol.* **8**, 848–852 (2011).
30. Bunnik, E. M. & Vernooij, M. W. Incidental findings in population imaging revisited. *Eur. J. Epidemiol.* **31**, 1–4 (2016).
31. Illes, J. *et al.* Discovery and disclosure of incidental findings in neuroimaging research. *J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn. Reson. Med.* **20**, 743–747 (2004).

The published version of the article can be found below

https://www.nejm.org/doi/full/10.1056/NEJMc1710724

Chapter 8

Common Polygenic Variations for Psychiatric Disorders and Cognition in Relation to Brain Morphology in the General Pediatric Population

Silvia Alemany, Philip R. Jansen, Ryan L. Muetzel, Natalia Marques, Hanan El Marroun, Vincent W.V. Jaddoe, Tinca J.C. Polderman, Henning Tiemeier, Danielle Posthuma, Tonya White

Published in *Journal of the American Academy of Child and Adolescent Psychiatry* 2019 Jan 9

Objective: This study examined the relation between polygenic scores (PGSs) for 5 major psychiatric disorders and 2 cognitive traits with brain magnetic resonance imaging morphologic measurements in a large population-based sample of children. In addition, this study tested for differences in brain morphology-mediated associations between PGSs for psychiatric disorders and PGSs for related behavioral phenotypes.

Method: Participants included 1,139 children from the Generation R Study assessed at 10 years of age with genotype and neuroimaging data available. PGSs were calculated for schizophrenia, bipolar disorder, major depression disorder, attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder, intelligence, and educational attainment using results from the most recent genome-wide association studies. Image processing was performed using FreeSurfer to extract cortical and subcortical brain volumes.

Results: Greater genetic susceptibility for ADHD was associated with smaller caudate volume (strongest prior = 0.01: $\beta = -0.07$, $P = 0.006$). In boys, mediation analysis estimates showed that 11% of the association between the PGS for ADHD and the PGS attention problems was mediated by differences in caudate volume (n = 535), whereas mediation was not significant in girls or the entire sample. PGSs for educational attainment and intelligence showed positive associations with total brain volume (strongest prior = 0.5: $\beta = 0.14$, $P = 7.12 \times 10^{-8}$; and $\beta = 0.12$, $P = 6.87 \times 10^{-7}$, respectively).

Conclusion: The present findings indicate that the neurobiological manifestation of polygenic susceptibility for ADHD, educational attainment, and intelligence involve early morphologic differences in caudate and total brain volumes in childhood. Furthermore, the genetic risk for ADHD might influence attention problems through the caudate nucleus in boys.

Introduction

Findings from genome-wide association studies (GWASs) indicate that multiple common genetic variants of small effect contribute to the etiology of psychiatric disorders, implying a highly polygenic architecture[1]. However, it remains largely unknown how these common genetic variants ultimately contribute to the development of psychiatric symptoms.

Polygenic scores (PGSs) are increasingly being used to index individual genetic susceptibility for a given disorder or trait and explore shared genetic influences across phenotypes to improve understanding of disease etiology[2]. Studies in childhood have shown that the polygenic risk for schizophrenia (SCZ) is associated with lower cognitive abilities, greater social impairments, more behavioral problems, and psychopathology[3-5]. Interestingly, previous work from our group found that genetic susceptibility for educational attainment (EA; years of schooling) was inversely related to child behavioral problems[4]. In a large prospective study, polygenic risk for major depressive disorder (MDD) was associated with emotional problems in adulthood, but not earlier in life[6]. Similarly, a study of genetic risk for bipolar disorder (BD) in adult samples suggested an association with increased risk for different psychiatric disorders[7]. For childhood-onset psychiatric disorders, PGSs for attention-deficit/hyperactivity disorder (ADHD) have been associated with inattentive and hyperactive-impulsive traits, worse educational outcomes, and lower IQ in children and adolescents from the general population[8,9].

Genetic susceptibility to psychopathology and cognitive function has been linked to behavior[4], which could imply that heritable neurobiological mechanisms are at play in the early presentation of symptoms. Within this context, it is well established that brain morphology during development is strongly influenced by genetic factors[10]. Furthermore, widespread morphologic brain abnormalities have been associated with the pathophysiology of major psychiatric disorders[11-15]. Although genetic and environmental factors can account for these brain abnormalities, we expect that genetic susceptibility for psychiatric disorders are associated with variations in brain morphology. Indeed, several studies have reported relations between PGSs for psychiatric disorders and PGSs for structural brain magnetic resonance imaging (MRI) measurements in adults using medium to large samples in the context of imaging genetics[16-19]. Higher genetic risk for SCZ was related to total brain volume (TBV) in patients with SCZ ($n = 152$) and controls ($n = 142$)[16], although this finding was not replicated using 2 large general population-based samples ($n = 763$ and $n = 707$)[17]. Other studies in healthy populations have related polygenic risk for SCZ and BD to reduced globus pallidus and amygdala volumes ($n = 274$)[18]. However, one of the largest studies to date did not find evidence for associations between polygenic risk for SCZ, BD, or MDD and subcortical brain volumes using data from the UK Biobank study ($n = 978$)[19].

Furthermore, to our knowledge, no study has yet been conducted in a pediatric MRI sample representative of the general population. Thus, whether associations of polygenic susceptibility for major psychiatric disorders

and brain morphology are present earlier in life is largely unclear. Because ASD and ADHD are childhood-onset psychiatric disorders, the study of polygenic risk for these traits in pediatric samples is particularly relevant. To date, this has been hampered by the lack of large-scale imaging studies in children that include genetic data.

Against this backdrop, the goal of this study was to examine the association of polygenic susceptibility for 5 psychiatric disorders and 2 cognitive outcomes with global and subcortical brain volumes in a large population-based sample of school-age children. As a secondary aim, this study investigated the potential mediating role of brain morphologic variation in associations between PGSs for psychiatric disorders and those for related behavioral phenotypes.

We hypothesized that polygenic susceptibility for SCZ, BD, MDD, autism spectrum disorders (ASDs), and ADHD would be associated with brain morphologic characteristics that overlap with brain abnormalities consistently reported in patients affected by these disorders. For EA and intelligence, we hypothesized that PGSs for these traits would be positively associated with global brain morphology measures.

Methods

Study Population

Participants were drawn from the Generation R Study, an ongoing population-based cohort study of many domains of child development[20]. As part of the cohort's MRI study, 3,992 children were scanned from March 2013 through November 2015, corresponding to visits of the 9- to-11-year-old Generation R sample[21]. Of these children, 3,937 had images that were reconstructed using FreeSurfer 6.0. One hundred thirty-one children were excluded due to the use of a different sequence (n = 22), dental braces (n = 87), and the presence of incidental findings (n = 22)[22]. Of the remaining 3,806, 620 scans were excluded due to data rated as unusable after visual inspection of segmentation quality. This left 3,186 children with good-quality MRI data. Of these, genotype data were available for 1,189 children with European ancestry. Relatedness and genotype quality resulted in an additional exclusion of 50 children. Thus, the final sample included 1,139 participants (flowchart in **Figure S1,** available online). The study protocol was approved by the medical ethics committee of the Erasmus University Medical Center (Rotterdam, the Netherlands). Written informed consent was obtained from the legal representatives of all participants.

Magnetic Resonance Imaging

To familiarize participants with the MRI scanning environment, all children underwent a mock scanning session. Structural MRI scans were obtained on a 3-T scanner (Discovery MR750W; GE Worldwide, Milwaukee, WI). Whole-brain high-resolution T1-weighted inversion recovery fast spoiled gradient recalled sequences were obtained using an 8-channel head coil. The scan parameters were repetition time of 8.77 ms, echo time of 3.4 ms, inversion time of 600 ms, flip angle of 10°, field of view of 220 × 220 mm, acquisition matrix of 220 × 220, asset acceleration factor of 2, b of 900 s/mm^2, 230 contiguous slices with a thickness of 1.0 mm, and in-plane resolution of 1.0 × 1.0 mm. Further details on the design and protocol of the Generation R cohort's MRI study can be found elsewhere[21].

Cortical reconstruction and volumetric segmentation were carried out with FreeSurfer Image Analysis Suite 6.0[23]. Specifically, automatic parcellation and segmentation protocols were conducted using the recon-all stream to obtain total, cortical, and subcortical brain volumes. All images were inspected for surface reconstruction accuracy using automated and manual methods[24]. Based on previous research investigating brain abnormalities in psychiatric disorders[11,15], 10 volumetric brain measures were studied as outcomes: TBV, cortical gray matter (GM), total white matter, subcortical GM, ventricular volume, and cerebellum as global segmented brain measurements; and amygdala-hippocampus complex, caudate, putamen, and thalamus as subcortical brain volumes. Correlations between brain measurements are shown in **Figure S2,** available online.

Genotyping

DNA samples were collected from cord blood at birth or from venipuncture during a visit to the research center on Illumina 610K and 660K single-nucleotide polymorphism arrays depending on collection time (Illumina, San Diego, CA). Further details on genotype calling procedures in the Generation R Study can be found elsewhere[25]. Information on quality control procedures of the genotype data and principal component analysis can be found in **Supplement 1**, available online.

Polygenic Scoring

Only participants with European ancestry were selected for polygenic scoring. Genotype data that passed quality control were used to compute PGSs based on GWAS results for 5 psychiatric traits—SCZ, BD, MDD, ADHD, and ASD—from the Psychiatric Genomics Consortium. In addition, we calculated PGSs for EA

and intelligence. **Table S1,** available online, provides an overview of the GWASs used for PGS calculation. For intelligence, we repeated the GWAS meta-analysis after exclusion of the Generation R sample to ensure independence of discovery and target sample. The PGSs were computed using LDpred[26]. This polygenic scoring method infers the posterior mean effect size of each marker using a prior on effect size distribution and linkage disequilibrium information from a reference genotype panel. The LDpred algorithm has improved prediction accuracy compared with traditional methods. Six PGSs were computed for each trait corresponding to 6 priors that determined the proportion of single-nucleotide polymorphisms with a causal effect (0.01, 0.05, 0.1, 0.5, 1, and infinitesimal). All PGSs were standardized to a mean of 0 and a standard deviation (SD) of 1. Correlations between PGSs are shown in Figure S3, available online.

Statistical Analysis

Multiple linear regression analyses were conducted using R 3.3.1 (https://www.r-project.org/). To examine whether genetic susceptibility for major psychiatric disorders and cognition is related to brain morphology, each PGS was tested for association with each brain measure individually. In these models, brain measurements were assigned as dependent variables with PGSs for SCZ, BD, ADHD, ASD, EA, or intelligence generated at 6 LDpred priors as independent variables. Models with TBV as the outcome were adjusted by sex, age, and 4 genetic principal components. Models for the remaining brain measurements also were adjusted by total intracranial volume.

We corrected for multiple testing across all PGSs, generated at 6 different priors, for association with 10 brain measurements using the false discovery rate (FDR) method[27]. Results at a *P*-value less than 0.05 by FDR correction were considered statistically significant.

For statistically significant associations showing a consistent pattern of results, we performed mediation analyses to examine whether differences in the associated brain regions mediated associations between the PGS and the phenotypic manifestation of the polygenic trait. Multiple linear regressions analyses were conducted to examine associations among PGSs, brain measurements, and behavioral phenotypes by adjusting for the same covariates included in the primary analyses and age at behavioral assessment. Direct, indirect, and total effects were estimated using the "mediation" package in R. As long as the assumptions of the mediation analysis are met, the direct effect represents the effect of genetic susceptibility on behavioral phenotypes after controlling for variation in brain morphology, and the indirect effect represents the estimated effect of polygenic susceptibility operating through brain morphology[28]. The proportion of mediation by brain morphology can be calculated as the ratio of indirect effect to total effect. Given the data available in Generation R, mediation analyses were feasible only for associations with PGSs for psychiatric disorders for which behavioral data were assessed when children were 8 to 11 years of age (mean 9.7, SD 0.23, range 8.85–11.54) using the (Child Behavior Checklist [CBCL]/6-18)[29]. Genetic, neuroimaging, and behavioral data were available for 1,053 participants. Further details on behavioral assessment can be found elsewhere[21]. For psychiatric disorders with sex differences in prevalence, we also conducted stratified analysis by sex.

To elucidate whether each cognitive trait independently contributed to the variation in brain measurement, we performed sensitivity analyses for analyses between the PGSs for EA and intelligence and for TBV by mutually adjusting using the PGSs for intelligence and EA, respectively.

Results

Sample Characteristics

A total of 1,139 children were included in the present study (561 girls [49.30%]), and the mean age was 10.16 years (SD 0.60, range 8.72–11.99).

Effects of PGS on Brain Morphology

Figure 1 presents a summary of the associations between the PGS for psychiatric disorders and the PGS for cognition calculated at 6 priors and brain volumes. Full results for these associations are presented in **Table S2,** available online.

No significant associations were observed between PGSs for SCZ and BD and brain measurements.

Greater genetic susceptibility for MDD was consistently related to smaller TBV, with the strongest association for the infinitesimal prior ($\beta = -0.07$, standard error [s.e.] 0.03; $P_{uncorrected} = 0.009$). PGS for MDD also showed negative associations with total white matter (prior = 0.01: $\beta = -0.03$, s.e. = 0.01, $P_{uncorrected} = 0.043$), cerebellum volume (prior = 0.5: $\beta = -0.05$, s.e. = 0.02, $P_{uncorrected} = 0.042$; prior = 1: $\beta = -0.05$, s.e. = 0.02, $P_{uncorrected} = 0.040$), and thalamus volume (prior = 0.01: $\beta = -0.05$, s.e. = 0.02, $P_{uncorrected} = 0.009$). However, after FDR correction, none of these associations remained significant.

PGSs for ADHD were associated with smaller TBV and caudate volume across all priors, although associations did not reach statistical significance for prior 0.01 in the case of TBV and prior 1 in the case of caudate volume.

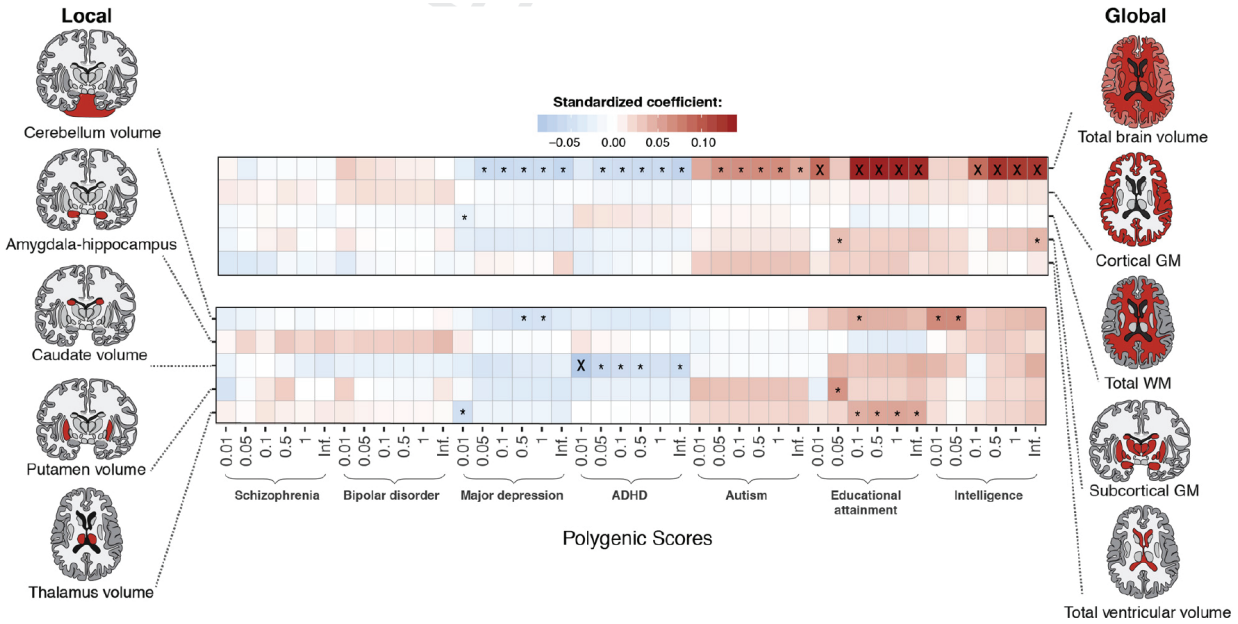


Fig. 1 | Associations Between Polygenic Scores for Psychiatric Disorders and Cognition and Brain Volumes (n = 1,139).

The strongest association with TBV was observed at the infinitesimal prior (β = −0.07, s.e. = 0.03, $P_{uncorrected}$ = 0.006), and the strongest association with caudate volume was observed at prior 0.01 (β = −0.08, s.e. = 0.03, $P_{uncorrected} = 7.49 \times 10^{-4}$) and remained significant after FDR correction.

PGS for ASD showed positive associations with TBV at all priors except at prior 0.01, which did not reach significance but did show the same direction of effect. The largest magnitude of the association was observed at prior 1 (β = 0.07, s.e. = 0.03; $P_{uncorrected} = 7.75 \times 10^{-3}$). These associations did not surpass FDR correction.

The EA PGSs were consistently associated with larger TBV (strongest prior 0.5: β = 0.14, s.e. = 0.03, $P_{uncorrected} = 7.12 \times 10^{-8}$) and remained significant after FDR correction. Associations at prior 0.05 did not reach significance but showed the same direction of effect. Greater genetic susceptibility for EA also was associated with larger volumes of subcortical GM (prior = 0.05: β = 0.04, s.e. = 0.02, $P_{uncorrected}$ = 0.046), cerebellum (prior = 0.1: β = 0.05, s.e. = 0.02, $P_{uncorrected}$ = 0.047), putamen (prior = 0.05: β = 0.06, s.e. = 0.03, $P_{uncorrected}$ = 0.016), and thalamus at multiples priors (strongest prior = 1: β = 0.05, s.e. = 0.02, $P_{uncorrected}$ = 0.012).

Greater genetic susceptibility for intelligence was significantly related to larger TBV for most priors, even after FDR correction (strongest prior 0.5: β = 0.12, s.e. = 0.03, $P_{uncorrected} = 6.87 \times 10^{-7}$). Other associations not surviving FDR correction included a positive association with subcortical GM (infinitesimal prior: β = 0.04, s.e. = 0.02, $P_{uncorrected}$ = 0.024) and positive associations with cerebellum volume (priors 0.01 and 0.05: β = 0.07, s.e. = 0.02, $P_{uncorrected}$ = 0.003).

Mediation Analysis

Only the association between polygenic risk for ADHD and caudate volume survived FDR correction; therefore, we tested whether caudate volume mediated the association between polygenic risk for ADHD and the attention problems CBCL syndrome scale. The caudate nucleus met the conditions to act as a mediator, because it showed a negative significant association with attention problems (β = −0.06, s.e. = 0.00, P = 0.029). Similarly, polygenic risk for ADHD was significantly associated with attention problems (β = 0.12, s.e. = 0.00, $P = 5.36 \times 10^{-5}$). However, mediation was 4.6% and not significant within the entire sample (**Figure 2**).

In analyses stratified by sex, mediation was significant only in boys, indicating that 11% of the association between polygenic risk for ADHD (prior = 0.01) and attention problems might be mediated by differences in caudate volume (**Figure 2**).

Sensitivity Analysis

Analyses mutually adjusting for polygenic susceptibility for EA and intelligence at prior 0.05 showed that the PGSs

for these 2 traits were independently associated with TBV (PGS for EA: β = 0.10, s.e. = 0.03. $P = 2.6 \times 10^{-4}$; PGS for intelligence: β = 0.08, SE = 0.03, $P = 0.003$).

Discussion

We examined whether polygenic susceptibility for psychiatric disorders and cognition was associated with brain morphology in children.

We found a consistent pattern of results across priors, indicating that the polygenic risk for ADHD was negatively associated with caudate volume, with the finding of a prior of 0.01 surviving multiple testing correction.

Polygenic susceptibility for intelligence and EA showed a positive relation with TBV that was consistent across all priors used, although generally not significant for the more stringent priors (i.e., 0.05 and 0.01). Polygenic risk for SCZ and BD did not show significant associations with brain morphology; however, several brain measurements were related to PGSs for MDD and ASD, although none of these associations survived multiple testing correction. These findings indicate the neurobiological manifestation of polygenic susceptibility for ADHD, intelligence, and EA involves early morphologic differences in caudate volume and TBV during development. Whole-brain and caudate volume reductions have been related to ADHD in a recent mega-analysis[14]. Given the high heritability of ADHD[30], we expected that regions previously associated with the disorder also would be associated with polygenic risk for ADHD. To the best of our knowledge, this is the first study providing evidence indicating that polygenic risk for ADHD might be, at least in part, underlying TBV and caudate reductions in childhood. These findings are particularly relevant for caudate volume reduction, one of the most replicated findings in ADHD[31]. Interestingly, our results suggest that reduced caudate volume might be mediating the association between polygenic risk for ADHD and attention problems in boys. ADHD is 2 to 9 times more prevalent in boys during childhood and adolescence[32].

Sex differences in brain morphology have been used to investigate whether ADHD-related brain abnormalities are more pronounced in male versus female individuals. Although caudate volume did not show sex effects in the mega-analysis conducted by Hoogman et al.[14], another study examining the volume and shape of basal ganglia observed smaller caudate volumes in boys with ADHD compared with male controls and no differences among girls[33]. Similarly, smaller caudate volumes have been found in adult male patients with ADHD compared with male controls, whereas no differences were observed in women[34]. Our findings are in line with these studies supporting that different genetically influenced neurobiological mechanisms might be operating in male and female individuals in the context of ADHD.

The EA PGSs were associated with larger TBV. Intracranial volume has been previously related to EA genetic variants by applying linkage disequilibrium score regression methodology[35]. Genetic variants for EA or other traits can affect TBV directly, through direct gene expression, through gene-environment interaction

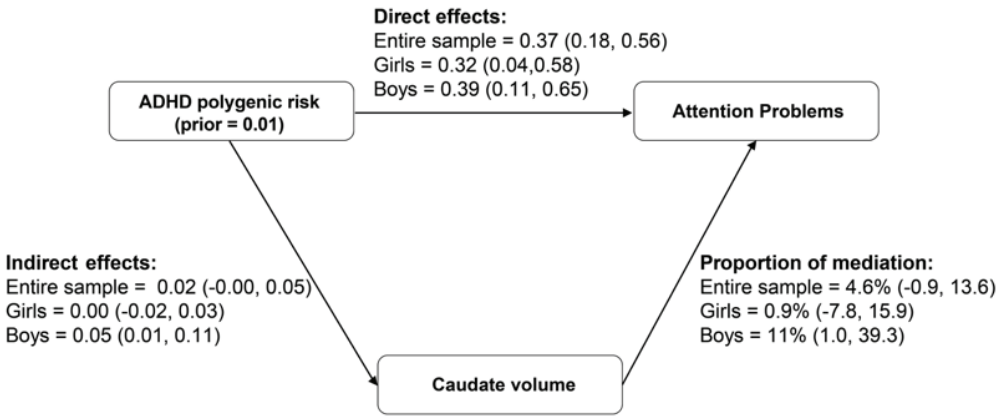


Fig. 2 | Mediation analysis of estimated effect (95% CI) of polygenic risk for attention-deficit/hyperactivity disorder (ADHD; prior = 0.01) on attention problems (Child Behavior Checklist Syndrome Scale) through caudate volume in the entire sample (n = 1,053) and stratified by sex (n = 535 boys; n = 518 girls)

or correlation mechanisms, or through intermediate phenotypes. Remarkably, an important number of single-nucleotide polymorphisms related to EA are located within genomic regions regulating gene expression in the fetal brain and genes mainly expressed in neural tissue[35]. These genes are especially active during the prenatal period and enriched for biological pathways involved in neural development[35]. Thus, it is likely that polygenic susceptibility for EA includes variants that directly promote optimal brain development.

Another possibility would be that EA genetic variants could influence brain morphology through environmental exposures that positively affect brain development, which would imply gene-environment correlation effects. In fact, children with higher genetic loading for EA tend to be raised in socioeconomically advantaged environments[36], which positively affects brain development[37]. It also is important to note that genetic associations with EA can be mediated by other phenotypes such as intelligence or personality traits, which are considered intermediate phenotypes for EA[38].

In addition, higher genetic loading for EA was nominally associated with larger thalamus volumes at multiple priors. The thalamus is a major hub in the brain, relaying multimodal information covering a wide range of cognitive functions, including learning, memory, inhibitory control, decision making, control of visual orienting responses, and attention[39]. Thus, a relation between polygenic susceptibility for cognitive functions relevant for EA with increased volume of the thalamus is neurobiologically plausible.

Not surprisingly, our findings on polygenic susceptibility for intelligence and EA largely overlap in the strength of the association and variance explained by TBV. Similarly to EA, genetic variants related to intelligence were identified in genes predominantly expressed in brain tissue[40]. Interestingly, polygenic susceptibility for EA and for intelligence influenced TBV independently of each other. Because the correlation between the PGSs for EA and intelligence was not extremely strong (**Figure S3,** available online), we speculate that genetic variants related to these traits might act through different pathways. Studies have shown that TBV is positively correlated with intelligence, accounting for approximately 16% of the variance in IQ[41]. Furthermore, our results indicate a shared genetic overlap between IQ and brain size, which is in line with twin studies suggesting that the association between these phenotypes is mainly of genetic origin[42].

Contrary to our hypothesis, polygenic risk for SCZ was not associated with brain morphologic variation at 9 to 11 years of age. This is in line with previous research in adults[17,19]. However, this null finding was surprising, because we found an association between the PGSs for SCZ and internalizing symptoms, and especially thought problems[4]. Behavioral effects of polygenic risk for SCZ must have neural correlates that we could not detect for several potential reasons. First, the neural correlates of SCZ PGS might be related to other neurobiological phenotypes not quantified in our study. This would not be the case for white matter measurements, including global and tract-specific fractional anisotropy and mean diffusivity, that were tested for an association with polygenic risk for SCZ in this sample and showed negative results[43]. Also, polygenic risk for SCZ has been associated with functional brain parameters, such as brain activation patterns detectable with functional MRI during cognitive tasks in adolescents[44,45]. Second, brain structural abnormalities related to genetic risk for SCZ might be detectable only in young individuals beginning in the prodromal phase, when the illness has begun to show clinical manifestations. These finding is "unmasked" as the illness progresses, making it very difficult to observe in general population samples, especially early in life. Third, genetic risk for SCZ has been related to nonparticipation in a large longitudinal population-based cohort study[46], implying that individuals at high genetic risk might be underrepresented. This would lead to underestimating effects of these genetic variants on neurodevelopmental outcomes. However, PGSs for SCZ were very similar among Generation R participants with European ancestry when comparing those included with those excluded in the present study (**Table S3,** available online).

Other interesting findings, albeit not surpassing multiple testing correction, include positive relations between PGS for ASD and TBV and negative associations between MDD PGS and TBV. Converging evidence points to an increased brain size as a characteristic brain abnormality of young children with ASD[47]. Our results suggest that this association could be accounted for by common genetic variants increasing the risk for ASD. Although it might seem counterintuitive that polygenic risk for ASD shows the same direction of effects on TBV as PGSs for EA and intelligence, it has been shown that polygenic risk for these traits is highly correlated and that genetic risk for ASD might act through different etiologic pathways[48]. For MDD PGS, widespread GM and subcortical volume reductions have been reported in individuals affected by MDD[49]. In contrast, less research has been conducted on global structural brain measures such as TBV. Overall, further research is needed to confirm these potential associations.

Our results should be interpreted in the context of several strengths and limitations. The strengths of the present study include the large sample and homogeneity with respect to recruitment, exclusion criteria, scanner, image acquisition, and preprocessing methods, which are especially valuable in imaging genetics. That said, the present sample is adequate for detecting significant effect sizes larger than 0.08 at 80% power; thus, reported smaller effect sizes, which correspond to negative findings, should be interpreted with caution. The main limitation of the study is the cross-sectional design. Studies including brain morphologic measurements at multiple time points are needed to examine whether polygenic risk for psychiatric disorders and cognition contributes to changes in developmental trajectories. Another limitation is that the PGSs typically explain only a small proportion of the total phenotypic variance of complex traits[1,2]. Moreover, it is important to note that the predictive accuracy of the PGS is related to sample size in the discovery sample, which substantially varies among different traits for the PGSs examined in the present study[50]. This should be considered when comparing results for the different traits examined. Nevertheless, we used summary statistics from the most recent, and thus more powerful, GWASs conducted on psychiatric disorders to date, which represents an advantage over previous studies using PGSs based on GWASs conducted on smaller samples.

To conclude, we found a relation between polygenic susceptibility for intelligence and EA and TBV in school-age children. We also found effects of ADHD polygenic risk for caudate volume. Interestingly, we found evidence for mediation only in boys, in whom differences in caudate volume accounted for 11% of the association between polygenic risk for ADHD and attention problems at 9 years of age. Overall, our findings provide molecular genetic evidence for the relation between polygenic susceptibility for cognition and ADHD with early differences in brain morphology.

References

1. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Wray, N. R. *et al.* Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
3. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The Lancet Psychiatry* **4**, 57–62 (2017).
4. Jansen, P. R. *et al.* Polygenic scores for schizophrenia and educational attainment are associated with behavioural problems in early childhood in the general population. *J. Child Psychol. Psychiatry* **59**, 39–47 (2018).
5. Nivard, M. G. *et al.* Genetic overlap between schizophrenia and developmental psychopathology: longitudinal and multivariate polygenic risk prediction of common psychiatric traits during development. *Schizophr. Bull.* **43**, 1197–1207 (2017).
6. Riglin, L. *et al.* The impact of schizophrenia and mood disorder risk alleles on emotional problems: investigating change from childhood to middle age. *Psychol. Med.* **48**, 2153–2158 (2018).
7. Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V. & Zammit, S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. *Schizophr. Res.* **197**, 2–8 (2018).
8. Martin, J., Hamshere, M. L., Stergiakouli, E., O'Donovan, M. C. & Thapar, A. Genetic risk for attention-deficit/hyperactivity disorder contributes to neurodevelopmental traits in the general population. *Biol. Psychiatry* **76**, 664–671 (2014).
9. Stergiakouli, E. *et al.* Association between polygenic risk scores for attention-deficit hyperactivity disorder and educational and cognitive outcomes in the general population. *Int. J. Epidemiol.* **46**, 421–428 (2016).
10. Jansen, A. G., Mous, S. E., White, T., Posthuma, D. & Polderman, T. J. C. What twin studies tell us about the heritability of brain development, morphology, and function: a review. *Neuropsychol. Rev.* **25**, 27–46 (2015).
11. Van Erp, T. G. M. *et al.* Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium. *Biol. Psychiatry* **84**, 644–654 (2018).
12. Hibar, D. P. *et al.* Subcortical volumetric abnormalities in bipolar disorder. *Mol. Psychiatry* **21**, 1710–1716 (2016).
13. Schmaal, L. *et al.* Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol. Psychiatry* **21**, 806–812 (2016).
14. Hoogman, M. *et al.* Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis. *The Lancet Psychiatry* **4**, 310–319 (2017).
15. Van Rooij, D. *et al.* Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD working group. *Am. J. Psychiatry* **175**, 359–369 (2017).
16. Van Scheltinga, A. F. T. *et al.* Genetic schizophrenia risk variants jointly modulate total brain and white matter volume. *Biol. Psychiatry* **73**, 525–531 (2013).
17. Van der Auwera, S. *et al.* No association between polygenic risk for schizophrenia and brain volume in the general population. *Biol. Psychiatry* **78**, e41–e42 (2015).
18. Caseras, X., Tansey, K. E., Foley, S. & Linden, D. Association between genetic risk scoring for schizophrenia and bipolar disorder with regional subcortical volumes. *Transl. Psychiatry* **5**, e692 (2015).
19. Reus, L. M. *et al.* Association of polygenic risk for major psychiatric illness with subcortical volumes and white matter integrity in UK Biobank. *Sci. Rep.* **7**, 42140 (2017).
20. Tiemeier, H. *et al.* The Generation R Study: a review of design, findings to date, and a study of the 5-HTTLPR by environmental interaction from fetal life onward. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 1119–1135 (2012).
21. White, T. *et al.* Paediatric population neuroimaging and the Generation R Study: the second wave. *Eur. J. Epidemiol.* **33**, 99–125 (2018).
22. Jansen, P. R. *et al.* Incidental findings on brain imaging in the general pediatric population. *N. Engl. J. Med.* **377**, 1593–1595 (2017).

23. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
24. White, T. *et al*. Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction. *Hum. Brain Mapp.* **39**, 1218–1231 (2018).
25. Medina-Gomez, C. *et al*. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol*. **30**, 317–330 (2015).
26. Vilhjálmsson, B. J. *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
27. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser.* B 289–300 (1995).
28. Valeri, L. & VanderWeele, T. J. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* **18**, 137–150 (2013).
29. Achenbach, T. M. & Rescorla, L. A. Manual for the ASEBA preschool forms and profiles. 30, (Burlington, VT: University of Vermont, Research center for children, youth, 2000).
30. Thapar, A., Cooper, M., Eyre, O. & Langley, K. What have we learnt about the causes of ADHD? *J. Child. Psychol. Psychiatry* **54**, 3–16 (2013).
31. Nakao, T., Radua, J., Rubia, K. & Mataix-Cols, D. Gray matter volume abnormalities in ADHD: voxel-based meta-analysis exploring the effects of age and stimulant medication. *Am. J. Psychiatry* **168**, 1154–1163 (2011).
32. Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J. & Rohde, L. A. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am. J. Psychiatry* **164**, 942–948 (2007).
33. Qiu, A. *et al*. Basal ganglia volume and shape in children with attention deficit hyperactivity disorder. *Am. J. Psychiatry* **166**, 74–82 (2009).
34. Onnink, A. M. H. *et al*. Brain alterations in adult ADHD: effects of gender, treatment and comorbid depression. *Eur. Neuropsychopharmacol*. **24**, 397–409 (2014).
35. Okbay, A. *et al*. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
36. Belsky, D. W. *et al*. The genetics of success: How single-nucleotide polymorphisms associated with educational attainment relate to life-course development. *Psychol. Sci.* **27**, 957–972 (2016).
37. Noble, K. G. *et al*. Family income, parental education and brain structure in children and adolescents. *Nat. Neurosci.* **18**, 773-778 (2015).
38. Krapohl, E. *et al*. The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proc. Natl. Acad. Sci.* **111**, 15273–15278 (2014).
39. Mitchell, A. S. *et al*. Advances in Understanding Mechanisms of Thalamic Relays in Cognition and Behavior. *J. Neurosci.* **34**, 15340-15346 (2014).
40. Savage, J. E. *et al*. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–917 (2018).
41. Haier, R. J., Jung, R. E., Yeo, R. A., Head, K. & Alkire, M. T. Structural brain variation and general intelligence. *Neuroimage* **23**, 425–433(2004).
42. Posthuma, D. *et al*. The association between brain volume and intelligence is of genetic origin. *Nat. Neurosci.* **5**, 83–84 (2002).
43. Jansen, P. R. *et al*. Polygenic Scores for Neuropsychiatric Traits and White Matter Microstructure in the Pediatric Population. *Biol. Psych. Cogn. Neurosci. Neuroimaging* **4**, 243–250 (2018).
44. Lancaster, T. M. *et al*. Associations between polygenic risk for schizophrenia and brain function during probabilistic learning in healthy individuals. *Hum. Brain Mapp.* **37**, 491–500 (2016).
45. 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)
46. Martin, J. *et al.* Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am. J. Epidemiol.* **183**, 1149–1158 (2016).
47. Riddle, K., Cascio, C. J. & Woodward, N. D. Brain structure in autism: a voxel-based morphometry analysis of the Autism Brain Imaging Database Exchange (ABIDE). *Brain Imaging Behav.* **11**, 541–551(2017).
48. Weiner, D. J. *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).
49. Grieve, S. M., Korgaonkar, M. S., Koslow, S. H., Gordon, E. & Williams, L. M. Widespread reductions in gray matter volume in depression. *NeuroImage Clin.* **3**, 332–339 (2013).
50. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* (2013)

Supplementary information

Supplementary information, figures (1-3) and tables (1-3) can be found in the online version of the manuscript:

https://jaacap.org/article/S0890-8567(19)30007-3/fulltext

Chapter 9

Polygenic Scores for Neuropsychiatric Traits and White Matter Microstructure in the Pediatric Population

Philip R. Jansen, Ryan L. Muetzel, Tinca J.C. Polderman, Vincent W. Jaddoe, Frank C. Verhulst, Aad van der Lugt, Henning Tiemeier, Danielle Posthuma, Tonya White

Published in *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* 2019 Mar 1;4(3):243-50

Background: Genome-wide association studies (GWAS) have identified numerous genetic variants that predispose to neuropsychiatric traits. Identification of mechanisms in the brain that underlie these associations is essential for understanding manifestations of genetic predisposition within the general population. Here, we investigate the association between polygenic scores (PGS) for seven neuropsychiatric traits and white matter microstructure of the brain on diffusion tensor imaging (DTI) in the pediatric population.
Methods: Participants from the Generation R Study that had genotype and DTI data available (n = 1,138, mean age = 0.2 years, range = 8.7–12.0) were included. PGS were calculated for five psychiatric disorders (ADHD, bipolar disorder, autism, major depressive disorder, schiszophrenia) and two cognitive traits (intelligence and educational attainment), and tested for associatios with global and tract-specific fractiona anisotropy (FA) and mean diffusivity (MD).
Results: Significant positive associations with global FA were observed for the PGS of intelligence (β=0.109, SE = 0.029, P < 0.001, ΔR^2 = 0.012) and educational attaiment (β = 0.118, SE = 0.029, P < 0.001, ΔR^2 = 0.014). No significant associations were observed with FA for the PGSs for psychiatric disorders. Tract-specific analysis showed that the PGS for intelligence and educational attainment were associated with FA of several association and projection fibers of the brain.
Conclusions: Our results show that genetic prediposition for cognition-related traits, but not psychiatric disorders, is associated with microstructural diffusion measures of white matter tracts at an early age. These results suggest a shared genetic etiology among structural connectivity, intelligence and educational attainment.

Introduction

Recent genome-wide association studies (GWAS) have improved insight into the highly complex polygenic architecture of human behavioral traits, including psychiatric disorders[1-3] and cognitive ability[4,5]. The rapid discovery of genetic variants has created the need for identification of downstream mechanisms in order to understand the biological impact of genetic risk on a system level[6-8].

Recent studies have utilized polygenic scoring analyses (PGS) to estimate overall genetic risk for psychiatric disorders and test the combined effects of thousands on SNPs on brain-imaging derived phenotypes using magnetic resonance imaging (MRI)[9]. Indeed, structural brain imaging studies in the general population have shown associations with disease-related alterations in healthy individuals carrying a high polygenic score for psychiatric illness, including differences in gyrification patterns[10] and cortical thickness[11]. Functional imaging studies have shown that polygenic risk for schizophrenia can be linked to different brain activity during tasks[12,13], and during rest[14], illustrating the complex combined downstream effects on brain functioning. In addition, evidence of brain differences in healthy subjects at high genetic risk has also been suggested by imaging studies in high risk individuals with a first-degree relative with a psychiatric disorder, which showed abnormalities in a variety of structural[15-17] and functional measures of the brain[17-19].

However, so far only few studies have investigated associations of polygenic risk with white matter fibers of the brain[20,21], even though the structural connectivity of the brain is known to be related to major psychiatric disorders, including schizophrenia[22] and bipolar disorder[23,24], as well as in normal cognitive functioning[25,26], and white matter changes have been observed in healthy relatives of psychiatric patients[27,28]. In addition, most prior genetic studies only included GWAS significant SNPs ($P < 5 \times 10^{-8}$) in the polygenic score, and do not take the contribution in genetic signal of subthreshold SNPs into account[29]. Moreover, prior studies have almost exclusively focused on adolescents or adults, while deviation from normal brain development may be present much earlier in life.

Here, we investigate whether genome-wide polygenic scores for psychiatric traits and cognitive ability are associated with white mater microstructure on diffusion tensor imaging (DTI) of the brain in a large population-based cohort of children between nine and twelve years of age. Insight into a possible shared genetic etiology between psychiatric disorders, cognitive ability and white matter microstructure provides further understanding neurobiological manifestations of genetic predisposition for psychopathology and cognition at early age in the general population

Methods and materials

Study Sample

The current study was conducted within the Generation R Study, a population-based cohort studying multifaceted aspects of child development[30]. Between March 2013 and November 2015, participants were enrolled in the cohort's MRI study with the aim of studying brain development in the general population by collecting, high quality, single scanner MRI data of the brain[31]. The current study included unrelated participants of European ancestry that had

good quality MRI data available and from whom genotype data had been collected previously. The Medical Ethics Committee of the Erasmus University Medical Center approved the study protocol, and the legal representative of the participants provided written informed consent.

Diffusion Tensor Imaging

Diffusion Tensor Imaging (DTI) of the brain was performed on a single study-dedicated 3 Tesla MR750w Discovery MRI scanner (General Electric, Milwaukee, WI, USA). Twelve major WM tracts were identified using probabilistic tractography. Diffusion characteristics within these tracts were used to quantify mean fractional anisotropy (FA) and mean diffusivity (MD). A detailed description of the imaging procedures, scan protocol and subsequent processing of the DTI data is provided in **Supplementary Information 1.1-1.5**. Confirmatory factor analysis (CFA) was applied using the Lavaan R package[32] to model a single latent factor of global fractional anisotropy (FA) and mean diffusivity (MD), as described by Muetzel et al.[25]. White matter tracts included in the model and standardized factor loadings on the global factor are shown in **Figure 1** and **Supplementary Table 1-2**. The global factors were tested for association with the PGS in univariate analyses.

Genotype data

Genotype data was collected at birth or during a visit to the research center using Illumina 610K and 660K genotype arrays (Illumina, San Diego, CA, USA). Data collection and subsequent processing procedures have been described previously[33]. Additional quality control procedures of the genotype data and genotype imputation are described in the **Supplementary Information 2.1-2.4**.

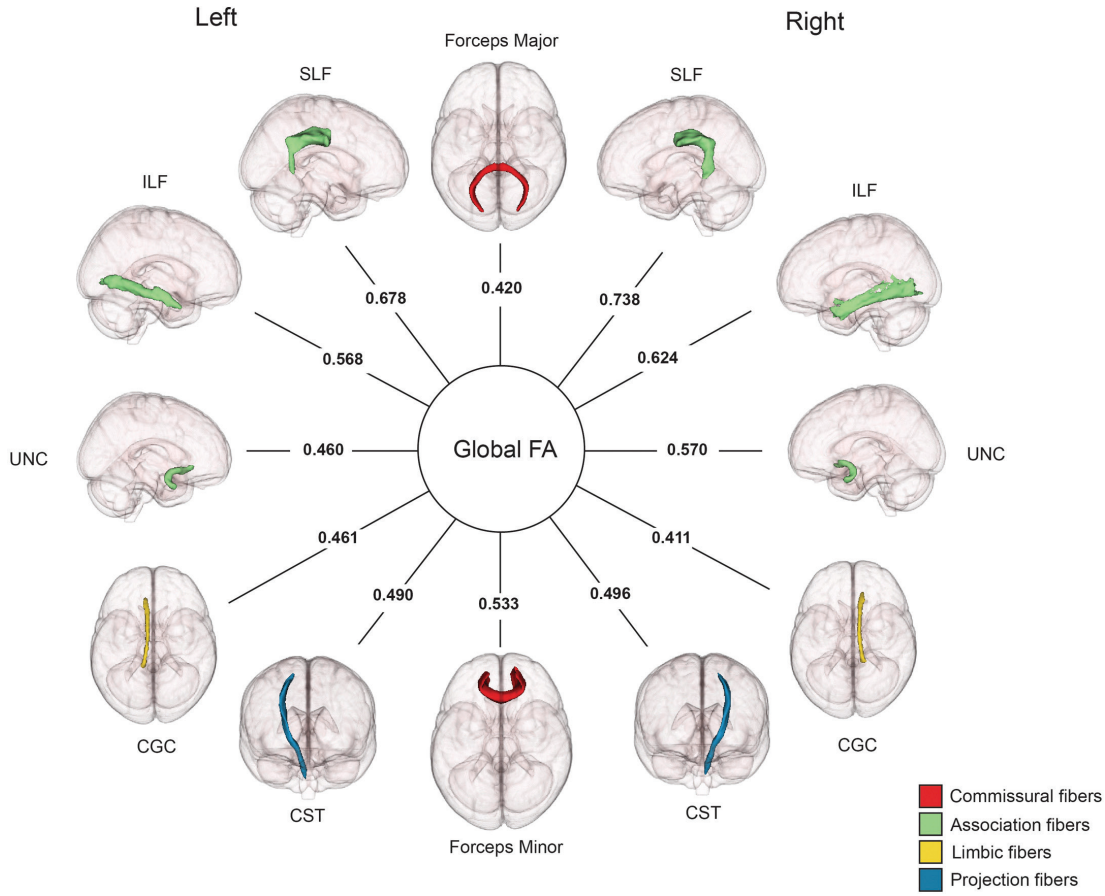


Fig. 1 | Standardized factor loadings of white matter tracts included in the global factor of fractional anisotropy (FA). Global factors for FA were estimated using confirmatory factor analysis (cfa). White matter tracts are color-coded according to subcategory. SLF: superior longitudinal fasciculus; ILF: interior longitudinal fasciculus; UNC: uncinate fasciculus; CGC: cingulum bundle; CST: corticospinal tract.

Polygenic Scoring

Polygenic scores (PGS) were calculated on imputed genotype data using publicly available GWAS results for five psychiatric disorders and two cognitive traits, including attention-deficit hyperactive disorder (ADHD), autism spectrum disorder, bipolar disorder, major depressive disorder, schizophrenia, intelligence, and educational attainment. An overview of the discovery GWAS studies is provided in **Supplementary Table 3**. As the Generation R cohort was included in the GWAS of intelligence, the GWAS was repeated after exclusion of the Generation R cohort (sample size after exclusion n = 267,938). Generation R was not included in any of the other six GWAS studies. PGS were calculated using PRSice[34], a script for calculation of PGS in PLINK[35]. We calculated PGS based on several *P*-value thresholds (pT) for inclusions of SNPs in the score (pT <0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1)

We tested multiple thresholds to find the optimal threshold that has the strongest association

with the outcome. PGS were subsequently standardized to a mean of 0 and standard deviation of 1 for interpretability. The number of SNPs that were included in each PGS and threshold is shown in **Supplementary Table 4**.

Statistical analysis

Statistical analyses were performed using the R statistical software[36] (version 3.2.1). Association testing was performed in a hierarchical approach. First, a global factor of white matter microstructure was predicted from the CFA model and regressed on the PGS. Next, in secondary analyses we studied tract-specific associations by regressing the individual white matter tracts on the PGS *P*-value threshold that showed the strongest association with the global factor in the primary analysis (lowest *P*-value). All analyses were corrected for age, gender and four genetic principal components as covariates. False Discovery Rate was used to correct for multiple testing (FDR)[37]. Correction was applied to the total number of statistical tests for each risk score, *P*-value threshold, and global and tract-specific diffusion measures. An FDR-corrected significance threshold was applied and *P*-values below 0.004, were considered statistically significant.

Results

Sample characteristics

A total number of 3,992 participants underwent MR imaging of the brain. DTI was completed in 3,786 of these participants. After DTI image quality control procedures, 3,279 participants remained. Of these participants, 1,920 individuals had genotype data available. Subsequent filtering based on European ancestry, relatedness and genotype quality resulted in 1,138 participants that were included in the study (flowchart in **Supplementary Figure 1**). The mean age of the sample was 10.2 years (range: 8.72 – 11.99), with a balanced distribution of sex (50.6% boys). The mean standardized polygenic scores for educational attainment and intelligence were slightly higher compared to the genotyped participants of European ancestry that did not participate in the MRI study (Educational attainment: 0.058 vs. –0.039, t = 2.56, P = 0.01; intelligence: 0.099 vs. –0.067, t = 4.39, $P = 1.17 \times 10^{-5}$) (**Supplementary Table 5**), and lower for ADHD (–0.055 vs. 0.036, t = –2.39, P=0.02) and depression (–0.071 vs. 0.047, t = –3.09, P = 0.002) There was a moderate correlation between several polygenic scores (correlation heatmap shown in **Supplementary Figure 2**), showing the largest correlation between the educational attainment and intelligence polygenic scores (r^2 = 0.38 to 0.47 between different *P*-value thresholds).

Associations with IQ

We tested whether the PGS of intelligence and educational attainment were associated with non-verbal IQ, measured in a subsample of 982 participants around the age of six years. The PGS of intelligence and educational attainment were strongly associated with non-verbal IQ, explaining approximately 5% by the PGS of intelligence (β = 0.222, s.e. = 0.032, $P = 1.87 \times 10^{-12}$, ΔR^2 = 0.050) (**Supplementary Table 6**).

Global FA / MD

Explained variance (ΔR^2) in the global factor of FA and MD by the PGS are shown in **Figure 2** and **Figure 3** respectively, full regression results are shown in **Supplementary Tables 7-8**. The PGS of intelligence showed positive associations with global FA across different *P*-value thresholds and was the strongest for the PGS based on a *P*-value threshold of pT (β = 0.118, s.e. = 0.029, $P < 0.001$, ΔR^2 = 0.014). We did not observe significant associations were between the global factor of FA and the PGS of the five psychiatric traits after correcting for multiple testing. In addition, none of the seven PGS showed associations with the global factor MD that survived multiple testing correction (**Figure 3**).

Tract-specific analysis

To test whether associations with specific white matter tracts could explain the association between the PGS and global FA, we performed univariate associations with diffusion measures FA and MD of individual white matter tracts. PGS based on the *P*-value threshold that showed the strongest association with the global factor of FA and

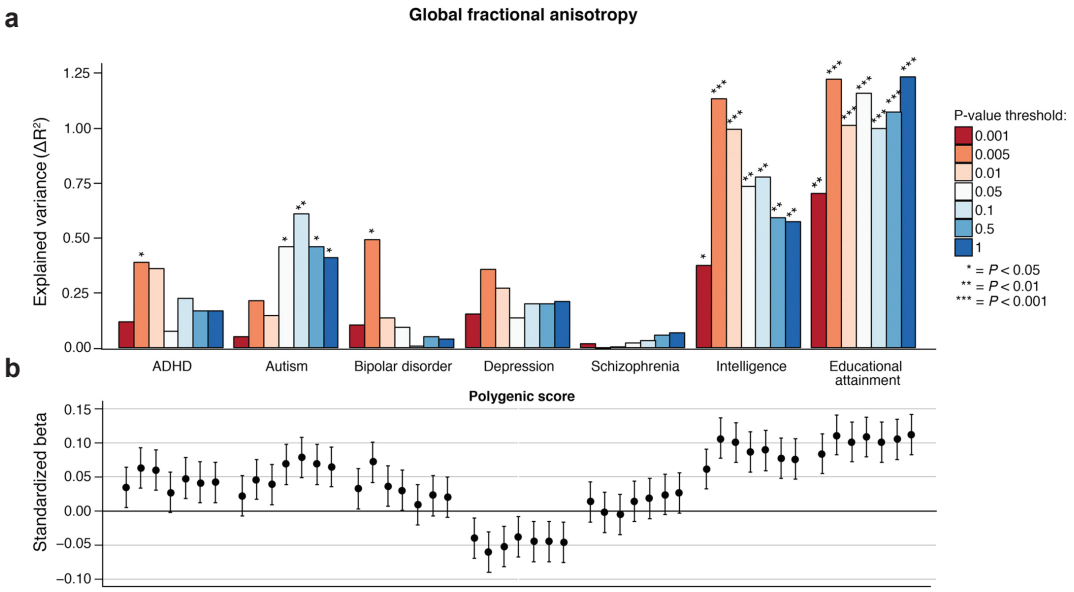


Fig. 2 | Variance explained in global fractional anisotropy by polygenic scores. (a) Variance explained (ΔR^2) in global fractional anisotropy (FA) by the polygenic score (PGS). (**b**) Standardized regression coefficients of associations between the different PGS and global FA for each *P*-value threshold, corrected for age, gender and ten genetic principal components.

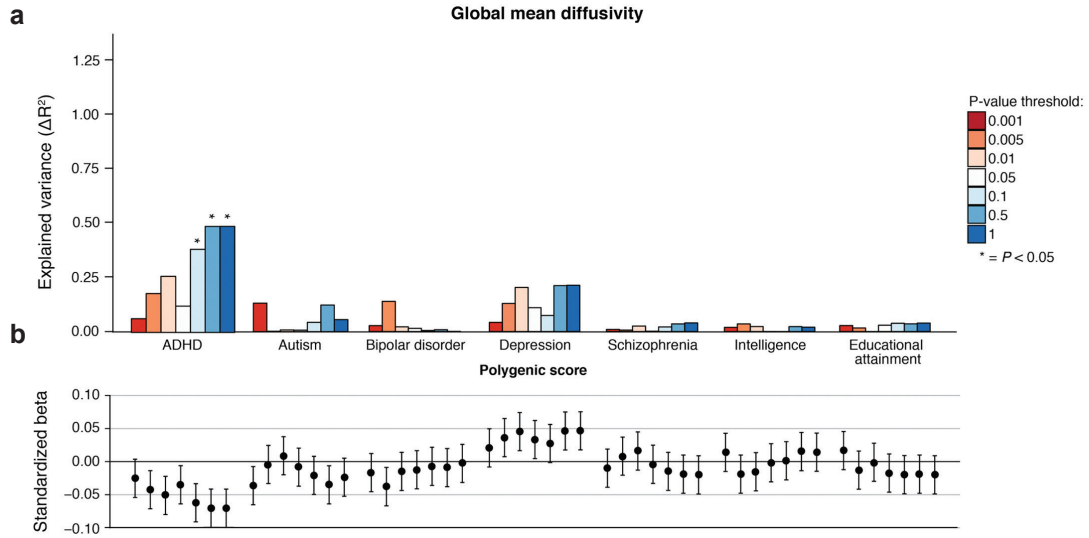


Fig. 3 | Variance explained in global mean diffusivity by polygenic scores. (a) Variance explained (ΔR^2) in the global factor of mean diffusivity (MD) by the polygenic score (PGS). (**b**) Standardized regression coefficients of the PGS on global MD for each individual *P*-value threshold, corrected for age, gender and four genetic principal components.

MD in the primary analysis (lowest *P*-value) were tested for tract-specific associations. **Figure 4** shows the association results between the PGS and FA and MD in each white matter tract, a full overview of the regression results is provided in **Supplementary Table 9-10**. Effect sizes for intelligence and educational attainment are represented visually in **Figure 5**. The PGS of intelligence showed positive associations with tract-specific FA in four major white matter tracts: the right superior longitudinal fasciculus (SLF; $\beta = 0.125$, s.e. = 0.029, $P < 0.001$), the left inferior longitudinal fasciculus (ILF; $\beta = 0.087$, s.e. = 0.029, $P < 0.001$), and both the left and right corticospinal tract (CST; left: $\beta = 0.132$, s.e. = 0.029, $P < 0.001$; right: $\beta = 0.148$, s.e. = 0.029, $P < 0.001$). Associations between educational attainment PGS and white matter tract partially overlapped with results of intelligence PGS, and showed

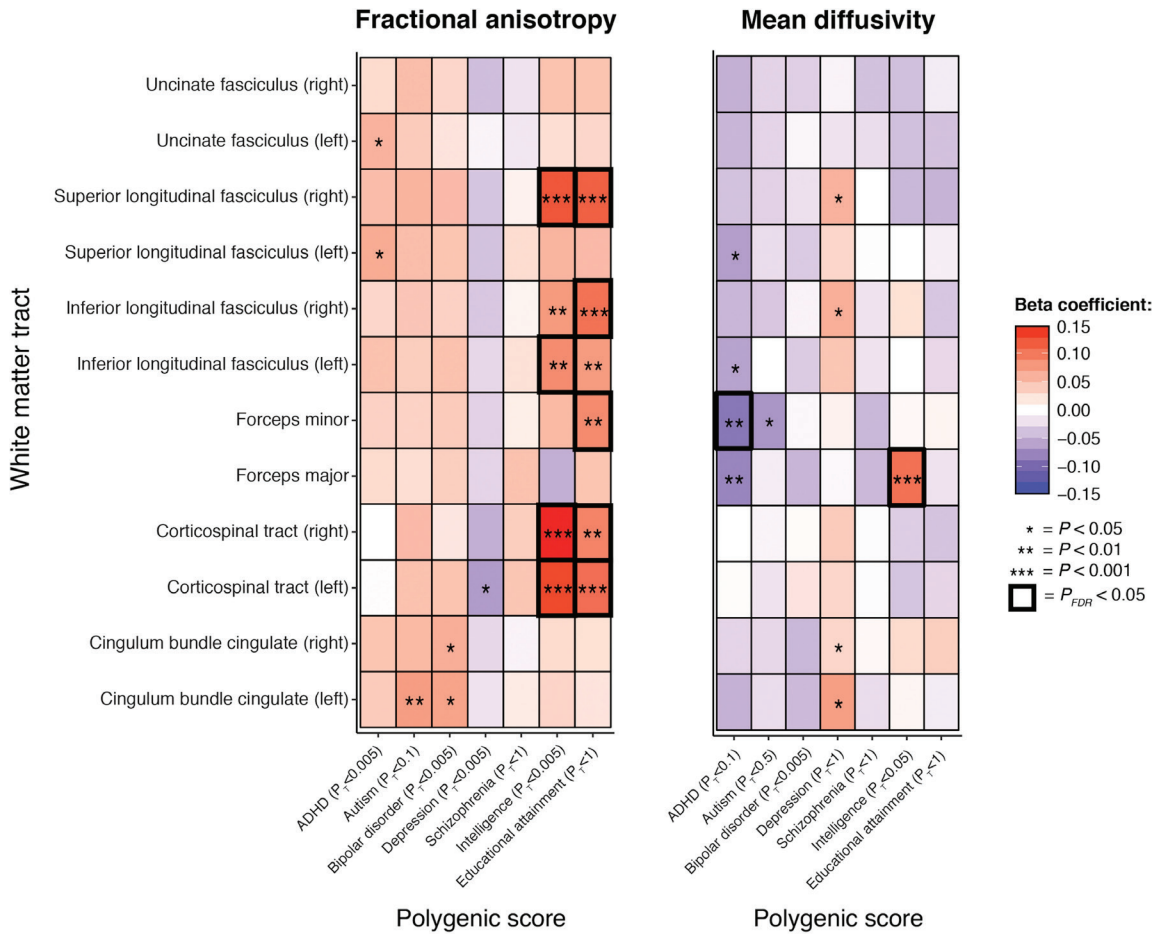


Fig. 4 | Tract-specific associations between polygenic scores and white matter tracts. Associations between polygenic scores and tract-specific fractional anisotropy (FA) and mean diffusivity (MD). Coefficients are standardized regression coefficients, corrected for age, gender and four genetic principal components.

similar positive associations with the right SLF (β= 0.118, s.e. = 0.029, $P < 0.001$) and the left and right CST (left: β = 0.107, s.e. =0.029, $P<0.001$; right: β = 0.092, s.e. = 0.029, $P < 0.001$). In addition, significant associations were observed with the right ILF (β = 0.105, s.e. = 0.029, $P < 0.001$) and the forceps minor (FMI; β = 0.088, s.e. = 0.029, $P < 0.001$). Tract-specific FA was not associated with the psychiatric PGS. For tract-specific MD values, we observed a significant positive association between the intelligence PGS and the forceps major (β = 0.105, s.e. = 0.029, $P < 0.001$) whereas a negative association was observed between the ADHD PGS and MD of the FMI (β = –0.088, s.e. = 0.029, $P < 0.001$).

Discussion

In this study, we observed positive associations between genetic predisposition for cognition-related traits and white matter microstructure on MRI in the pediatric population, with the PGS of intelligence and educational attainment explaining approximately 1% of the variance in global fractional anisotropy. Tract-specific analyses showed that these associations driven by several association and project fibers of the brain. These results may suggest a shared genetic etiology between global white matter integrity, general cognitive functioning and predicted later-life educational achievement.

Previous research showed that the PGS of educational attainment is associated with general intelligence, but has also been associated with socio-economic status[38], and later-life outcomes, including reproductive behavior[39] and longevity[40]. To date, genetic variants related to cognitive traits have only been linked to total intracranial volume on MRI based on GWAS summary statistics using LD Score regression[41]. Our study is the first to report significant as-

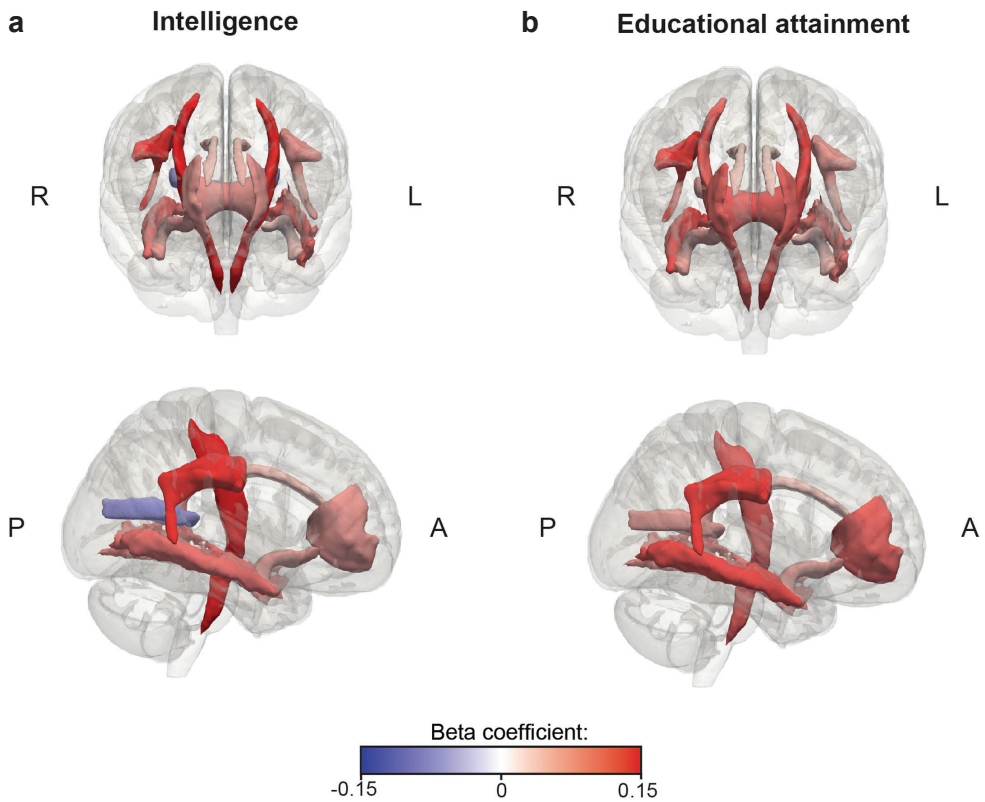


Fig. 5 | Visual representation of tract-specific associations been PGS and white matter tracts. (a) Associations between the PGS for intelligence and tract weighted-average fractional anisotropy (FA), (**b**) PGS of educational attainment and tract weighted-average FA. Coefficients are standardized regression coefficients, corrected for age, gender and four genetic principal components. Regression results are shown in **Supplementary Table S8-S9.**

sociations between PGS for intelligence and educational attainment, and structural connectivity of the brain, emphasizing the important role of white matter microstructure in cognitive functioning. This finding is in line with previous work from our group that reported associations between non-verbal IQ and global fractional anisotropy[25], and specific associations with the superior longitudinal fasciculus. Our study adds to these findings that cognition and white matter microstructure are likely to share a common genetic architecture. We hypothesize that two underlying mechanisms may explain these observed associations. First, the discovery GWAS of educational attainment by Okbay et al.[5] reported that candidate genes near the 74 genome-wide significant variants showed elevated expression in the central nervous system. Moreover, these candidate genes were highly enriched for gene-sets related to neurodevelopment, such as sprouting of dendrites and synaptic plasticity. Similar gene-set results were observed by Savage et al.[42] in the GWAS of intelligence, which highlighted that genes related to several cellular processes in neurons influence cognitive functioning. Given the associations between PGS of intelligence and educational attainment and white matter microstructure in our study, it may be possible that similar molecular pathways and neurobiological processes lead to higher developed states of microstructural organization, which subsequently leads to a higher fractional anisotropy on DTI. Genetic studies of white matter integrity on DTI indeed confirmed that genes involved in synaptic processes, such as neuronal transmission and cell adhesion, are important contributors to white matter microstructure[43]. Second, given previously described associations between the educational attainment PGS of the child and parental socioeconomic status[44], gene-environment correlations with environmental factors that positively impact white matter microstructure, including prenatal factors[45], parenting strategies[46] and a healthy lifestyle[47] may amplify the observed associations. Considering that educational achievement is correlated with a broad range of environmental factors, it is possible that the educational attain-

ment PGS captures the combined effect of a diverse array of factors that impact white matter development.

Interestingly, we did not observe associations between the schizophrenia PGS and white matter microstructure, which is surprising given extensive literature on white matter abnormalities in schizophrenia patients[22], individuals at high genetic risk for schizophrenia, as defined by family history[48,49], and associations between the schizophrenia PGS and behavioral problems in our sample as previously reported[50]. We argue that several factors may explain this negative finding. First, at the age of our study sample (mean age of 10.2 years), white matter abnormalities described in schizophrenia patients may be not yet present, not present on a scale detectable by DTI, or present but obscured by increased variance associated with different rates of white matter maturation between individuals. Moreover, MR imaging modalities examining structural phenotypes and activation patterns of the brain may be more sensitive to developmental changes related to the genetic risk for schizophrenia. Previous studies in healthy individuals indeed have shown associations between schizophrenia PGS and cortical morphology on structural imaging[10,11], and activation patterns during cognitive tasks on functional MRI (fMRI)[13,51]. Second, the PGS in this study only captures genetic signal from common variants (MAF >0.01) of typically low individual effect sizes[52]. White matter alterations found in schizophrenia patients may follow from more deleterious rare variants with comparatively larger effects and higher penetrance. Compelling evidence exists that these rare mutations contribute substantially to schizophrenia risk[53,54] and commonly disrupt neurodevelopmental processes[55,56], which could potentially underlie the observed microstructural abnormalities. Third, nonparticipation among high-risk compared to low-risk individuals in population-based research have been previously described[57]. Subsequent underrepresentation of individuals with the highest risk of schizophrenia may further explain this null result. In addition, no associations were observed for the PGS of four other psychiatric traits. The absence of association for these traits may be partially explained by the GWAS small sample sizes (autism, depression, bipolar), the later onset of these disorders (depression, bipolar disorder), or an absent relation between white matter and these psychiatric disorders.

The current study has several strengths. First, the sample is large for imaging standards, especially in pediatric populations. Second, the sample comprised a narrow age range, and was performed in a population-based cohort, which can minimize, but certainly not remove, age-related differences in white matter development. Third, all subjects were scanned on a single, research-dedicated MRI scanner using the same software version, removing possible noise from inter-scanner differences or changes associated with scanner upgrades. Fourth, PGS for multiple traits were simultaneously tested, allowing for comparisons across traits in a single study sample.

Some limitations are also present. First, the associations between PGS and white matter microstructure were tested using a cross-sectional design. Prospectively collected brain imaging data could provide evidence on whether polygenic scores are associated with variation in trajectories of white matter development in children over time. Second, the current largest discovery GWAS studies used for calculating the PGS of ADHD, autism and bipolar disorder are less powered compared to other traits that were tested. As discovery sample sizes increase rapidly, we expect that PGS studies based on well-powered GWAS results will lead to more robust associations with brain imaging phenotypes. Lastly, polygenic risk scores do not provide insights into which SNPs contribute most to the observed associations with structural connectivity. Future genome-wide studies of structural connectivity in large DTI imaging samples may further aid in identifying SNPs linked to both traits, and in estimating genetic overlap between cognitive functioning, psychiatric disorders and structural connectivity.

In conclusion, we report evidence that genetic predisposition for cognitive traits is associated with higher white matter microstructural integrity in children, whereas no associations were found for five major psychiatric disorders. Future studies are necessary to explore associations with longitudinal developmental trajectories of white matter microstructure over time.

References:

1. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
2. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
3. Magel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
4. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
5. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
6. Breen, G. *et al.* Translating genome-wide association findings into new therapeutics for psychiatry. *Nat. Neurosci.* **19**, 1392–1396 (2016).
7. Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N. & Geschwind,

D. H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat. Neurosci.* **19**, 1397–1407 (2016).
8. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50** 322–328 (2018).
9. Dima, D. & Breen, G. Polygenic risk scores in imaging genetics: usefulness and applications. *J. Psychopharmacol.* **29**, 867–871 (2015).
10. Liu, B. *et al*. Polygenic risk for schizophrenia influences cortical gyrification in 2 independent general populations. *Schizophr. Bull.* **43**, 673–680 (2017).
11. French, L. *et al*. Early cannabis use, polygenic risk score for schizophrenia and brain maturation in adolescence. *JAMA psychiatry* **72**, 1002–1011 (2015).
12. Miller, J. A. *et al*. Effects of Schizophrenia Polygenic Risk Scores on Brain Activity and Performance During Working Memory Subprocesses in Healthy Young Adults. *Schizophr. Bull.* **44**, 844–853 (2017).
13. Lancaster, T. M. *et al*. Associations between polygenic risk for schizophrenia and brain function during probabilistic learning in healthy individuals. *Hum. Brain Mapp.* **37**, 491–500 (2016).
14. Wang, T. *et al*. Polygenic risk for five psychiatric disorders and cross-disorder and disorder-specific neural connectivity in two independent populations. *NeuroImage Clin.* **14**, 441–449 (2017).
15. McIntosh, A. M. *et al*. Longitudinal volume reductions in people at high genetic risk of schizophrenia as they develop psychosis. *Biol. Psychiatry* **69**, 953–958 (2011).
16. Lawrie, S. M. *et al*. Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. *Lancet* **353**, 30–33 (1999).
17. Cooper, D., Barker, V., Radua, J., Fusar-Poli, P. & Lawrie, S. M. Multimodal voxel-based meta-analysis of structural and functional magnetic resonance imaging studies in those at elevated genetic risk of developing schizophrenia. *Psychiatry Res. Neuroimaging* **221**, 69–77 (2014).
18. Zhang, R., Picchioni, M., Allen, P. & Toulopoulou, T. Working memory in unaffected relatives of patients with schizophrenia: a meta-analysis of functional magnetic resonance imaging studies. *Schizophr. Bull.* **42**, 1068–1077 (2016).
19. Lui, S. *et al*. Resting-state brain function in schizophrenia and psychotic bipolar probands and their first-degree relatives. *Psychol. Med.* **45**, 97–108 (2015).
20. Foley, S. F. *et al*. Multimodal Brain Imaging Reveals Structural Differences in Alzheimer's Disease Polygenic Risk Carriers: A Study in Healthy Young Adults. *Biol. Psychiatry* **81**, 154–161 (2017).
21. Reus, L. M. *et al*. Association of polygenic risk for major psychiatric illness with subcortical volumes and white matter integrity in UK Biobank. *Sci. Rep*. **7**, 42140 (2017).
22. Kelly, S. *et al*. Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group. *Mol. Psychiatry* **23**, 1261–1269 (2017).
23. Lin, F., Weng, S., Xie, B., Wu, G. & Lei, H. Abnormal frontal cortex white matter connections in bipolar disorder: a DTI tractography study. *J. Affect. Disord.* **131**, 299–306 (2011).
24. Wise, T. *et al*. Voxel-based meta-analytical evidence of structural disconnectivity in major depression and bipolar disorder. *Biol. Psychiatry* **79**, 293–302 (2016).
25. Muetzel, R. L. *et al*. White matter integrity and cognitive performance in school-age children: a population-based neuroimaging study. *Neuroimage* **119**, 119–128 (2015).
26. Deary, I. J. *et al*. White matter integrity and cognition in childhood and old age. Neurology **66**, 505–512 (2006).
27. Sprooten, E. *et al*. White matter integrity in individuals at high genetic risk of bipolar disorder. *Biol. Psychiatry* **70**, 350–356 (2011).
28. Skudlarski, P. *et al*. Diffusion tensor imaging white matter endophenotypes in patients with schizophrenia or psychotic bipolar disorder and their relatives. *Am. J. Psychiatry* **170**, 886–898 (2013).
29. International Schizophrenia Consortium, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
30. Kooijman, M. N. *et al*. The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31**, 1243–1264 (2016).
31. White, T. *et al*. Paediatric population neuroimaging and the Generation R Study: the second wave. *Eur. J. Epidemiol.* **33**, 99–125 (2018).
32. Rosseel, Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Stat. Softw.* **48**, 1–36 (2012).
33. Medina-Gomez, C. *et al*. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.* **30**, 317–330 (2015).
34. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, 1466–1468 (2014).
35. 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).
36. R Core Team,: A language and environment for statistical computing. (2013).
37. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser.* B 289–300 (1995).
38. Selzam, S. *et al*. Predicting educational achievement from DNA. *Mol. Psychiatry* **22**, 267–272 (2017).
39. Barban, N. *et al*. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* **48**, 1462–1472 (2016).
40. Marioni, R. E. *et al*. Genetic variants linked to education predict longevity. *Proc. Natl. Acad. Sci.* **113**, 13366–13371 (2016).
41. Bulik-Sullivan, B. *et al*. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
42. Savage, J. E. *et al*. GWAS meta-analysis (N= 279,930) identifies new genes and functional links to intelligence. *bioRxiv* 184853 (2017).
43. Lopez, L. M. *et al*. A genome-wide search for genetic influences and biological pathways related to the brain's white matter integrity. Neurobiol. Aging **33**, 1847-e1 (2012).
44. Krapohl, E. & Plomin, R. Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. *Mol. Psychiatry* **21**, 437-443 (2016).
45. Taylor, P. A. *et al*. A DTI-based tractography study of effects on brain structure associated with prenatal alcohol exposure in newborns. *Hum. Brain Mapp.* **36**, 170–186 (2015).
46. Puetz, V. B. *et al*. Altered brain network integrity after childhood maltreatment: A structural connectomic DTI-study. *Hum. Brain Mapp.* **38**, 855–868 (2017).
47. Chaddock-Heyman, L. *et al*. Aerobic fitness is associated with greater white matter integrity in children. *Front. Hum. Neurosci.* **8**, 584 (2014).
48. Hoptman, M. J. *et al*. A DTI study of white matter microstructure in individuals at high genetic risk for schizophrenia. *Schizophr. Res.* **106**, 115–124 (2008).
49. Maniega, S. M. *et al*. A diffusion tensor MRI study of white

matter integrity in subjects at high genetic risk of schizophrenia. *Schizophr. Res.* **106**, 132–139 (2008).
50. Jansen, P. R. *et al.* Polygenic scores for schizophrenia and educational attainment are associated with behavioural problems in early childhood in the general population. *J. Child Psychol. Psychiatry* **59**, 39–47 (2018).
51. 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).
52. Wray, N. R. *et al.* Research review: polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
53. Gratten, J. Rare variants are common in schizophrenia. *Nat. Neurosci.* **19**, 1426–1428 (2016).
54. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
55. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
56. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
57. Martin, J. *et al.* Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am. J. Epidemiol.* **183**, 1149–1158 (2016).

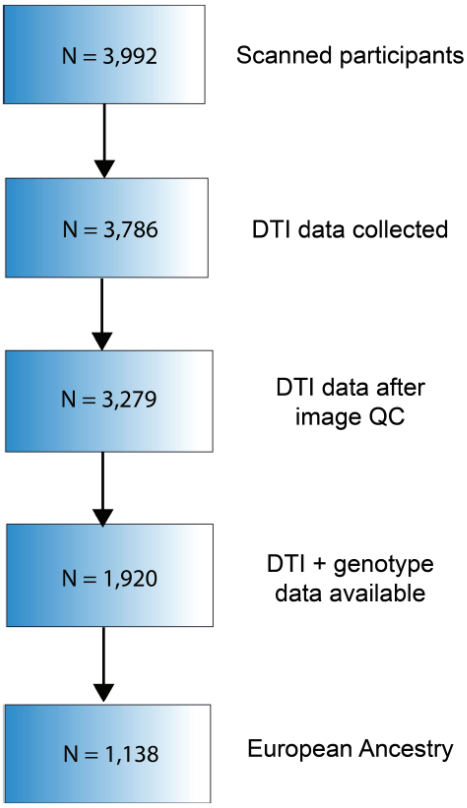
Supplementary information

Supplementary information, figures (1-2) and tables (1-10) can be found in the online version of the manuscript:

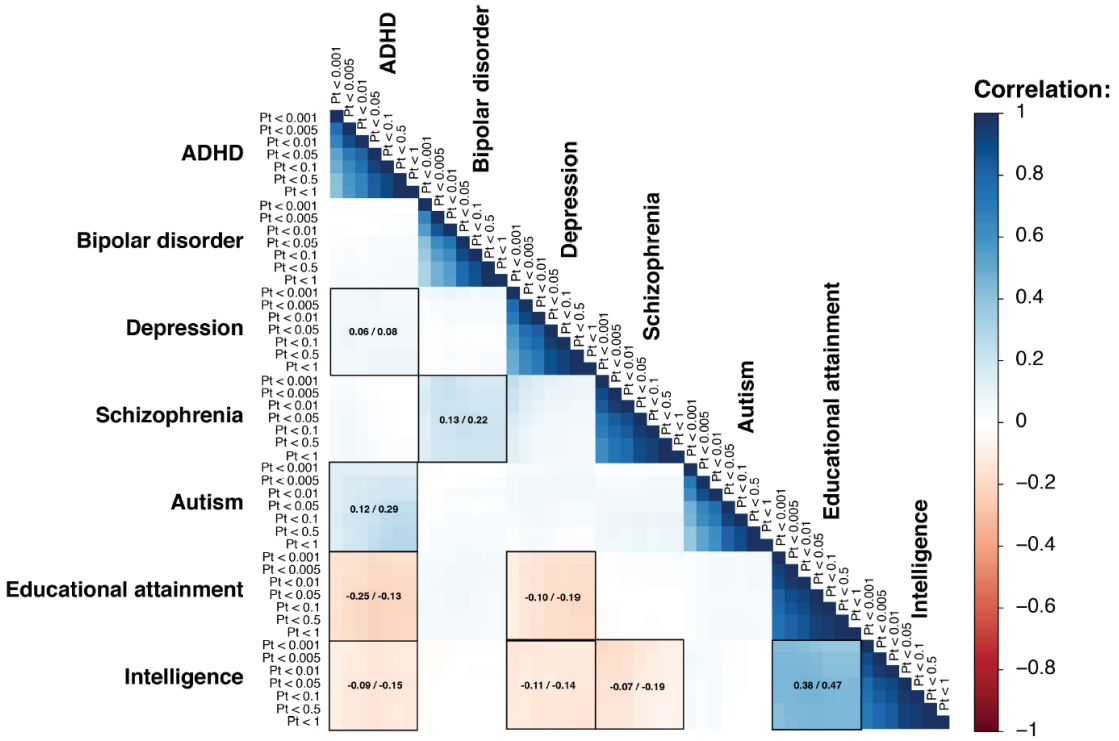
https://www.biologicalpsychiatrycnni.org/article/S2451-022(18)30185-X/fulltext

Supplementary information

Supplementary Fig. S1 | Participant selection. Flowchart showing the selection of participants from the Generation R Study for the current analysis.



Supplementary Fig. S2 | Heatmap showing correlations between the different polygenic scores and P-value thresholds. The highlighted squares show the strongest correlations between polygenic scores. The correlation estimates in the square show the range of correlations between different *P*-value thresholds for these traits.



Part III:
Appendix

Chapter 10: General discussion and summary

After more than a century of scientific milestones in genetics, from experiments in a small number of peas by Gregor Mendel to populations larger than one million individuals (see introduction, Fig. 1a), the identification of single genes that ae related to human behavior has only recently become possible. The current generation of scientists are the first to discover the human genome and find genes related to human characteristics, just as previous generations of scientific researchers in neuroscience were the first to look at the living human brain through new brain imaging techniques such as MRI and CT. Prior scientific steps that led to the discovery of the DNA structure, the first complete map of the human genome and subsequent mapping of variations in single base pairs between individuals ultimately paved the way to start searching for single genetic variants in genes that predispose to many different outcomes. The current thesis aimed to build forth on these steps by identifying novel genes and pathways that underlie human behavior through unprecedented sample sizes and investigate how these genes relate to differences in the brain. An important factor driving the discoveries in this thesis is the availability of large-scale datasets and improved statistical software that can handle analyses on a big data scale. Around the start of the thesis in 2014, the shift towards larger and larger studies led to important milestones in psychiatric genetics, among which was the identification of 108 genomic loci for schizophrenia that were discovered in a genome-wide association study (GWAS) in over 100,000 individuals[1]. This large-scale GWAS pointed towards novel interesting disease biology and new promises for pharmacological treatment options. Importantly, this study evidently demonstrated that, despite disappointing results in the smaller earlier studies, increases in study size unlocked the potential of GWAS for understanding the complexity of psychiatric disorders[2].

1. Summary of research findings

The research projects in this thesis were centered around two main research aims:

1) The first aim was to understand genetic factors that explain the heritability of complex human behavior in the population. The most important starting point for these analyses is a well-powered GWAS to obtain associations for each SNP in the genome. These SNP effects can be used as input for a wide variety of follow-up analyses to find genes and neurobiological mechanisms related to these traits, or to estimate genetic risk for these traits through the use of a polygenic risk score (PRS). The results related to this objective are described in the first part of the thesis.

2) The second aim was to use magnetic resonance imaging (MRI) of the brain collected in thousands to hundreds of thousands of individuals, including adults and children, to study how genetic variation between individuals relates to differences in the morphology of the brain. To this end we performed GWAS of brain MRI data and used PRS to investigate whether differences in polygenic risk for psychiatric disorders and cognition associate with deviations in brain morphology in a population-based sample of over a thousand children. These findings are reported in the second part of the thesis.

In the first part of the thesis (*aim 1*), we investigated the genetic architecture of several behavioral traits in the general population.

In **Chapter 2**, we described our study in which genetic data and behavioral measures collected in over 2,000 children from the age of 3 years onwards, were used to investigate whether genetic risk for psychiatric disorders is associated with problem behavior during early childhood. By calculating a PRS for five psychiatric disorders and educational attainment, we were able to show that variation in polygenic predisposition for schizophrenia based on common genetic variants is associated with higher levels of behavioral problems already at the age of three years, as reported by their mother. This was in particular reflected in higher levels of broad-scale internalizing problems. These associations persisted during follow-up reports at the age of 6 and 10 years of age. Interestingly, when examining specific problem scales, we found that the schizophrenia PRS was most strongly associated with the Thought Problem scale of the Child Behavior Checklist (CBCL), which may reflect positive symptoms such as that are associated with schizophrenia. In contrast, the polygenic score for educational attainment was negatively correlated with most scales on the CBCL and showed the strongest negative association with the Attention Problems scale at the age of 10. The reported results show clear evidence of behavioral manifestations of genetic risk

in early childhood and may suggest a potential role of PRS for these disorders to identify those children most at risk of developing behavioral problems at an early age, in combination with traditional environmental risk factors for this disease.

In **Chapter 3,** we report our findings investigating the genetic architecture of insomnia and sleep-related traits by carrying out the largest GWAS to date in a population aged 18-80 years of over one million individuals in the UK Biobank and 23andMe datasets, both which contained completed sleep questionnaires and genotype data. We identified more than 200 loci and nearly 1,000 genes for insomnia and showed that much of the genetic signal is shared with a variety of psychiatric disorders, including anxiety and depression. Based on Mendelian Randomization using these GWAS summary statistics, our results showed that insomnia shows a causal association with obesity, type 2 diabetes, and coronary artery disease, but not vice versa, emphasizing that sleep problems are an important risk factor for many negative health outcomes. RNA sequencing data from single neurons showed that there was enrichment of genetic signal of insomnia in neurons located in the hypothalamus and claustrum of the brain, confirming a suggested role of these brain structures in sleep regulation. These results provide genes and cell-types as novel targets for functional follow-up studies and illustrate how large-scale studies can greatly improve our knowledge of complex traits

To further uncover the genetic architecture of neuroticism and depression, **Chapter 4** describes a large-scale meta-analysis of GWAS studies on neuroticism and depression in over 400,000 adult individuals. These analyses showed nearly 200 loci to be involved in neuroticism. By carrying out cluster analysis, we showed that the total neuroticism questionnaire score is genetically heterogenous and that the questionnaire items can be broadly subdivided into subgroups of depressed affect and worrying with differential SNP and gene associations, and genetic correlations with previous GWAS studies. By analyzing gene-expression in single-cells, we showed evidence of the involvement of serotonergic neurons and dopaminergic neuroblasts in the brain, confirming the serotonergic and dopaminergic pathways as a target for antidepressant therapy. Neuronal involvement was further evidenced by enrichment of genetic signal in a number of neuronal processes, such as neuron differentiation, neurogenesis, neurodevelopment and neuron differentiation. In addition, we observed that many genes identified for neuroticism and depression may form targets for pharmacological treatment as indicated by a large drug-gene interaction database.

In **Chapter 5,** we report findings of our combined GWAS data on intelligence test scores of 14 cohorts and over 200,000 individuals, including both children and adults, to find genes associated with intelligence. We were able to detect over 200 loci of which most were new findings. We showed that several brain areas are significantly enriched for genetic signals, in particular in the frontal cortical areas of the brain. In addition, significant enrichment was observed for several types of pyramidal neurons, a major cell-type in the cortex and hippocampus of the brain, and neurons located in the CA1 region of the hippocampus. These findings are in line with observed significant enrichment in gene-sets and pathways associated with neurogenesis, a key process that takes place in the hippocampus of the brain. This large scale genetic analysis of intelligence provides many links between genetic variation and variation in intelligence scores and suggested functional pathways in neuronal cell-types.

In the second part of the thesis, we used magnetic resonance imaging (MRI) of the brain in two large population-based studies of both children and adults to study genetic influences on brain morphology and white matter microstructure.

In **Chapter 6,** we described the prevalence of incidental findings in brain MRI scans that were observed during the large-scale data collection of MRI brain imaging data in over 4,000 children of the Generation R cohort between the age of 8 and 12 years old. Our results clearly demonstrate that incidental findings are highly prevalent in children, which warrant careful screening of collected research data and streamlined protocols for clinical follow-up of the findings. In total, there were seven children where a suspected brain tumor was found, of which two children underwent surgery as a direct consequence of study participation shortly after participating in the study. These results are the most precise estimate of incidental findings prevalence in brain MRI scans in children to date.

In **Chapter 7**, we reported the results of GWAS meta-analysis of brain volume (BV) using brain scan data of over 40,000 individuals. We observed many novel loci related to BV and found that associated genes play a role in several interesting signaling pathways related to cell division, differentiation and apoptosis regulation, such as the *Erbb2/Erbb4*-signaling pathways. Interestingly, rare mutations in many of these genes are known to cause severe monogenic disorders characterized by intellectual disability and abnormal brain development, including micro-, macro- and megalocephaly. Prior evidence suggested that the phenotypic correlation between brain

volume and intelligence is completely due to overlap in genetic factors[3]. To further highlight which genes are involved in BV and intelligence, we carried out extensive cross-trait analyses of shared loci and genes between the BV GWAS results, and the GWAS of intelligence that we performed previously (**Chapter 5**). In total, we observed that 64 genes are implicated in both brain volume and our GWAS study of intelligence, and that the function of these genes is located in important signaling pathways of the brain involved in cell division and differentiation. These results are a step forward in understanding the genes associated with BV and illustrate how GWAS summary statistics alone can be used to find novel interesting overlapping biological mechanisms between genetically correlated traits.

In **Chapter 8**, we used structural brain imaging data and PRS of psychiatric disorders and cognition-related traits to study whether structural differences in brain morphology can be explained by variation in these polygenic scores. There were strong positive associations between PRS for cognition-related scores and total brain volume and suggested negative associations between the PRS of ADHD and brain volume. More specifically, a higher ADHD PRS was associated with smaller caudate nucleus volumes. This latter effect was most pronounced in boys. Using data on behavioral measures of ADHD symptoms in the same population, we observed that the caudate nucleus volume mediated the association between the ADHD PRS and ADHD symptoms in the population. These results suggest that caudate nucleus volume may be an endophenotype between polygenic predisposition for ADHD and ADHD symptoms in the general pediatric population. In contrast, the PRS for cognition was mostly linked to global differences in brain morphology, as evidenced by significant associations with total brain volume.

In the following **Chapter 9**, we used PRS for five psychiatric disorders (ADHD, autism, bipolar disorder, major depression, schizophrenia) and for cognition-related traits (educational attainment and intelligence) to study whether genetic predisposition for these traits can be translated to abnormal white matter tract connections in children. We used diffusion tensor imaging (DTI) data, collected in a large group of children in the Generation R Study to investigate associations between these PRSs and white matter microstructure. Although schizophrenia has repeatedly been linked to abnormal white matter microstructure in vivo[4], there were no associations between the schizophrenia PRS and white matter integrity. In contrast, we showed clear evidence that genetic predisposition for cognition-related traits such as educational attainment and intelligence is associated with higher fractional anisotropy. This may suggest a more optimal myelinization of the white matter tracts, a healthier state of white matter, in the brains of children who have a higher genetic predisposition to intelligence and educational achievement later in life. We demonstrate a clear link between genetic predisposition for cognition and white matter tracts of the brain, but not for PRS of psychiatric disorders. These results show the promising role of PRS to find associations with imaging biomarkers that correspond with the genetic predisposition for certain behavioral traits.

2. Interpretation of research finding

2.1 GWAS

In the GWAS that were carried out in this thesis, we have identified over 500 genomic loci in total for insomnia, neuroticism and intelligence, showing a non-uniform distribution across the genome. We observed genetic overlap between these traits (genetic correlations r_g = -0.19 to 0.44) that clusters in certain genomic regions (See **Fig. 1**), as evidenced by overlap in the number of loci (range $n_{overlap}$=18 to 20 loci) and several loci that are found in all three traits (e.g. loci on chromosome 22, where the genetic signal for all three traits is localized on the long q-arm of the chromosome only). This considerable pleiotropy of multiple genomic loci clearly shows that the boundaries between the genetic architectures of seemingly distinct behavioral traits are diffuse, in line with what has been observed in psychiatric and behavioral traits in general. The interpretation of this pleiotropy of shared loci is not straightforward, as these statistical pleiotropic effects can be explained by true genetic effects of a locus on multiple traits (biological pleiotropy), by a mediating effect of one phenotype on the other (mediated pleiotropy), or pleiotropy caused by different causal variants tagged by the same locus though LD (spurious pleiotropy)[5]. Recent analysis based on over 500 GWAS studies showed that this pleiotropy across the genome is highly present across several domains of human characteristics and estimated that over 90% of the loci and 80.9% of the genes are associated with multiple traits. Particularly in psychiatric traits, the pleiotropy tends to be high, whereas neurological diseases are known to be more genetically distinct[6], possibly explained by more objective diagnostic criteria (e.g., brain imaging, blood tests) compared to psychiatric disorders and more robust neurobiological differences.

Comparing the genetic architectures of the three studied traits, it is interesting to note that there is clear variation in the number of loci that were discovered for each trait

given the sample size of each study. In the GWAS of insomnia, the largest study in over 1 million individuals, we observed a comparable number of loci (n=202) as the much smaller GWAS of intelligence in almost 300,000 individuals (n=205 loci). This may be explained by a much higher SNP heritability for intelligence (h^2_{snp}=0.19, SE=0.01, **Chapter 5**) than insomnia (h^2_{snp}=0.07, SE=0.002, **Chapter 3**), differences in discoverability (effect size variants) and polygenicity (the proportion of causal variants) of these traits[7], and possibly more reliability in the measurement of sleep problems through questionnaire compared to cognitive test scores. Interestingly, the genetic signal in all three studies was associated with expression in the medium spiny neuron, a neuron cell-type that was significant in all three studies. These GABA-ergic inhibitory neurons make up the majority of the cells within the human striatum[8]. Prior evidence indeed suggests that the striatium is involved in a large number of brain functions, including emotion regulation[9], including cognition[10] and sleep-wake activity[11]. This finding of a

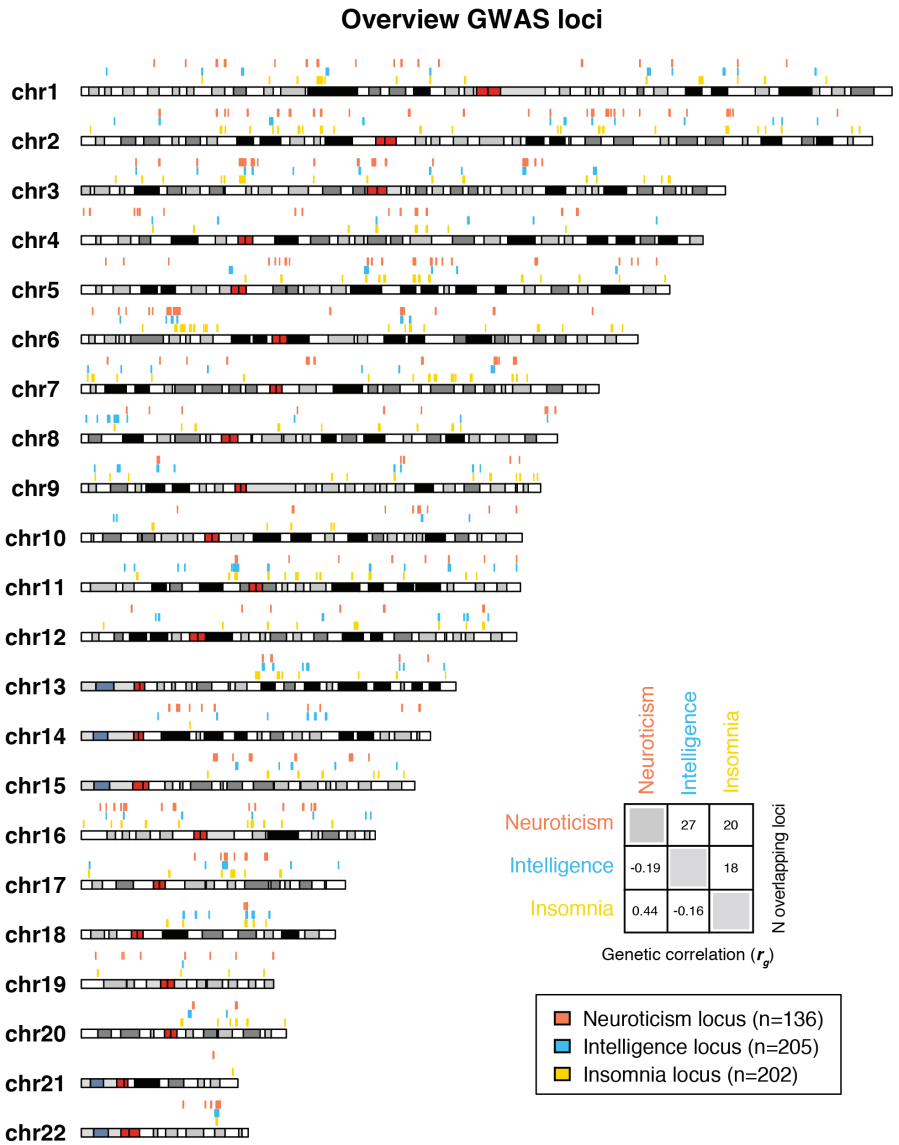


Fig. 1 | Overview of the independent genomic loci identified in the three GWAS studies of insomnia, neuroticism and intelligence, as presented in this thesis. Genomic positions of genome-wide significant loci are shown for insomnia (yellow, **Chapter 3**), intelligence (blue, **Chapter 5**) and neuroticism (pink, **Chapter 4**). The 3-by-3 matrix highlights the genetic correlations between these traits (lower side of the matrix) and number of overlapping loci (upper side of the matrix).

shared implicated neuronal cell-type may explain the multitude of neurobiological functions of the striatum.

In addition to the large number of significant loci, gene-mapping efforts in these GWAS studies led to over 1,000 genes (i.e., more than 5% of the genome) that were observed through different gene-mapping methods (see introduction, **Fig. 3**). These results are a step forward in understanding variation in these traits compared to previous GWAS studies carried out in smaller samples[12–14]. The wealth of novel loci and genes in the reported analyses illustrate that GWAS studies are reaching a critical size where novel findings are discovered more rapidly with increase in sample size. In addition, these results show that larger samples are not only necessary for detecting novel loci and genes, but that sufficient power is necessary for robustly identifying pathways and gene-sets, which can only be achieved when reaching sample sizes of well over 100,000 individuals for these three traits.

Future research is able to build forth on results reported in this thesis in a variety of ways. First, the large number of genes that are implicated in our work provide many previously undiscovered targets for follow-up studies. The functional effects of variants in individual genes can be studied on a cellular level by manipulating genes in induced-pluripotent stem cell (IPSC) cultures that may be differentiated to neuronal cell types to observe how these variants alter cellular characteristics. More recently, the emergence of high-throughput genome editing with CRISPR-CAS has made it possible to highly efficiently assert the functional consequences of single variants by introducing these in specific cell-lines[15]. On a larger scale, knock-out models of these genes using animal models may help understand the exact role of the identified gene on the level of a complete organism. These potential functional follow-up studies based on our GWAS results fulfill a strong need to fully grasp the functional role of the identified genes in cellular mechanisms and underlying biology that explain the association of these variants with observed behavioral differences. Second, we show that GWAS can be used to point towards specific tissue and cell-types through the analysis of RNA sequencing data, for example claustrum neurons and medium spiny neurons in insomnia in **Chapter 3**, serotonergic neurons in neuroticism and depression in **Chapter 4**, and CA1 hippocampal neurons in intelligence in **Chapter 5**. Experimental studies aimed at specifically manipulating these individual neuronal cell-types can be carried out to confirm the role of these cells in the biology of the trait. Third, PRS calculations based on these large GWAS results can be used to study trait-associated outcomes or find novel associations between PRS and endophenotypes, including biomarkers and brain imaging-derived parameters[16]. This is illustrated in **Chapter 8** and **9**, where polygenic scores based on the GWAS results of intelligence reported in **Chapter 5** was studied in association with brain MRI phenotypes including brain white matter tract integrity and brain volumes. Fourth, by making GWAS summary statistics of our GWAS studies available, future studies will be able to use our results by combining the GWAS results with novel bioinformatics data for insights into the studied traits, by performing meta-analysis with our results and GWAS performed in newly collected datasets to expand the number of identified variants, or by using our results for GWAS on correlated traits using multivariate GWAS approaches[17].

2.2 Brain imaging studies

In the second part of the thesis, we report results of the population-based studies that utilized brain imaging techniques.

In parallel to the upscaling in population size in genetics, larger studies that contain brain imaging data are essential for understanding the link between genes and the brain and how the mechanisms in the brain mediate gene-behavior associations (see introduction, **Fig. 2a**). However, the ambition to obtain more and more brain images through brain imaging research is not entirely without consequences. The reported incidental findings in the Generation R cohort in **Chapter 6**, caution that adequate follow-up protocols need to be in place prior to the collection of imaging data in large numbers of individuals given the reportedly high rate of incidental findings in children that warrant follow-up. The rate of incidental findings, especially suspected brain tumors, was higher than expected based on previous studies, and shows that incidental findings are not only an important issue to consider in adults but are highly relevant in pediatric neuroimaging as well. This is especially important in the advent of multiple large neuroimaging efforts that are ongoing or planned in the near future. In the field of genetics, incidental findings are also more and more frequently encountered, in both the clinical[18,19] and research[20] setting, where more accurate whole-exome and whole-genome sequencing increases the detection rate of unsuspected findings. The abundance of genetic variants across the human genome also in healthy participants, requires well-documented a priori agreement on which variants to report or not. Exchanging experience in reporting incidental findings in both scientific fields may improve protocols for handling these findings, which should include opinions and experiences from research participants and patients, as well as input from different

medical disciplines, including medical specialists, psychologists and medical ethicists.

In **Chapter 7**, we used these large imaging datasets to perform extensive genetic analyses on brain MRI data. These analyses showed that brain imaging phenotypes are highly polygenic, influenced by potentially hundreds to thousands of loci, in line with previous observations in complex traits[21]. In this chapter, we showed overlap between genes identified through GWAS, and genes that are known to lead to monogenic disorders characterized by abnormal brain growth. This is a particular interesting finding, as this suggests that both low-penetrant common variants present in the general population and highly penetrant rare mutations that cause monogenic (Mendelian) disorders in patients may converge on similar neurobiological mechanisms and pathways. Moreover, this overlap implies that the two apparently distinct areas of genetics, namely those concerned with monogenic causes of rare diseases and common variation in common polygenic traits and disorders, may not be as separate as often described. Findings in both fields may be translated between disciplines and support each other in generating hypothesis about genes and pathways that may be involved. Analysis of the overlap between rare and common variation in these same genes and the pathways is an exciting new area of research[22] that will lead to novel insights now that GWAS is bringing forth larger numbers of common genetic variants genes for a variety of traits.

In follow-up analyses of the BV GWAS, we showed several significant gene-sets to be involved in signaling pathways that are involved in brain developmental processes. Interestingly, we showed that specific genes within these pathways appear in the gene-mapping based on the GWAS results, either through mapping in FUMA[23] or by gene-based association testing in MAGMA[24], whereas for other genes that are part of the same signaling pathway, we found no evidence of being implicated in the trait. This observation is interesting, as it suggests that variation along several steps of the pathway has differential effects on the phenotype at the end of the pathway. The use of pathway analysis indicates that GWAS on macroscale brain-imaging phenotypes can lead to exciting new conclusions and hypothesis about which processes are driving variation on the microscale level in the brain.

In addition to GWAS analyses, we used brain imaging to study the link between PRS and brain morphology and microstructure in **Chapter 8** and **9**, respectively. In both chapters, we observed a much stronger association for polygenic scores based on cognitive trait GWAS (intelligence and educational attainment) than for those based on GWAS of psychiatric disorders. Interestingly, these associations were mostly found with global differences in brain structure rather than being confined to specific brain areas or tracts, including associations with total brain volume in **Chapter 8**, and global white matter microstructure in **Chapter 9**.

The contrast between associations based on polygenic scores for cognitive traits and psychiatric disorders suggests that cognitive polygenic scores may lead to measurable differences even at an early age. In contrast, the largely negative findings for PRS of psychiatric disorders may suggest that the underlying neurobiological differences as a product of these PRSs may not (yet) be present in this specific age range, or not at a scale that could be detected by the imaging modalities that were used for this thesis. An exception to this was the significant association we found for the ADHD polygenic score and caudate volumes in boys, where we showed that brain structure mediates the association between PRS and behavior. This association confirms prior evidence that found smaller caudate nucleus volumes in ADHD patients compared to controls[25,26].

The caudate nucleus is a C-shaped structure located along the lateral ventricles of the brain and forms a part of the human striatum and the large network of connections with the cortex and thalamus (cortico-basal ganglia-thalamic-cortical network). Among important brain functions of the caudate nucleus are cognitive processing[27], learning[28], attention[29] and response inhibition[30], several of which are impaired in ADHD[31]. A role of the caudate nucleus in ADHD was recently demonstrated in a large meta-analysis of children and adults carried out by the ENIGMA consortium, that found several gray matter structures to be smaller in ADHD cases[32]. Interestingly, the largest pooled effect size in the meta-analysis was observed for volumes of the amygdala and putamen, structures that are also part of the basal ganglia, for which we did not observe significant associations in our analyses. A possible explanation may be that our analyses included children only, whereas the meta-analysis included both children and adults, covering an age range of 4 to 63 years[32]. An earlier meta-analysis that included children-only samples indeed showed smaller volumes in the caudate nucleus[33], and no evidence of other sub-cortical structures being involved. Whereas meta-analyses including also older patients may be more sensitive to detecting differences related to the disease over a longer time period, the use of PRS in children samples may find associations with brain structures that are involved in an earlier stage of the disease[16]. However, it needs to be taken into account that reported associations with the caudate nucleus reflects the link between predisposition for ADHD and brain

morphology in a general population sample (in contrast to a case-control sample in these meta-analyses), and it is yet unclear whether these findings can be extrapolated to clinical samples.

3. Consideration in genetic research

Several important methodological considerations need to be taken into account when interpreting the results from this thesis.

In **Chapter 3, 4** and **5** of this thesis, we performed three GWAS studies that combined a number of large datasets, where we report a number of discussion points.

First, the enormous scale of the UK Biobank and 23andMe cohorts requires cost- and time-efficient data collection in large sets of individuals. The large sample size of these studies comes at the expense of a reduction in the depth and dimensionality of the collected phenotypes, which are often reduced to a limited set of questions. For example, the single insomnia question that was used in UK Biobank in the insomnia GWAS in **Chapter 3** is a proxy for more complex correlated phenotypes that may be present in the population. Although we report a high accuracy of this question for the detection of insomnia disorder by validation of this question in an external sample containing insomnia disorder patients, phenotyping based on a single question that is correlated with the phenotype of interest may be more susceptible to noise compared to extensive questionnaires, requiring larger samples to distill the same genetic signal. It is yet unclear whether the same genetic signal would be observed in a more extensive phenotype definition of insomnia. More extensive phenotyping or longitudinal data collection of sleep-measures would provide a more precise estimate of genetic effects on insomnia.

Second, our GWAS results reported in **Chapter 3, 4** and **5,** show that the contribution of individual risk loci is limited, with typical odds ratios below 1.1, and absolute standardized effect estimates below 0.02. GWAS has indeed frequently been criticized for discovering only (very) small associations of genetic variants with disease which are often difficult to interpret[34]. While we report exceptions (e.g. *MEIS1* (OR=1.1-1.2) locus in insomnia in **Chapter 3**), these results clearly emphasize that these traits are highly polygenic and that small individual effect sizes are part of their genetic architecture. In this thesis we show that, although focusing on the interpretation single risk loci does not lead to valuable new insights because of their small effect, performing extensive follow-up analyses based on the GWAS results by gene mapping, gene-set and gene-expression does lead to new insights into the genetics of the phenotype. GWAS studies should thus be regarded as a starting point of further follow-up analyses using methods that take the combined effects of multiple SNPs into account.

Third, research projects in this thesis took place in large population-based studies that sample from the general population, including the Generation R cohort, the UK Biobank, and 23andMe. These studies aim to include individuals that are a representative sample of the general population in order to find associations that can be extrapolated to the population at large. However, non-participation and attrition over time are more common in individuals with psychiatric disorders[35], or at high genetic risk for psychiatric disorders[36], and those with a lower educational level[37]. In this regard, the population-based studies in this thesis oversample for higher education and intellectual level, and under sample genetic risk of psychiatric disorders. However, with the large size of these population-based studies, there is a high probability of also including participants with clinical and subclinical disorders.

Forth, in the reported GWAS, we made use of meta-analyses and combined separate datasets to maximize statistical power. Due to variation in the genetic background of these cohorts, differences in phenotype definitions and data collection, in addition to random variation, the genetic correlations between meta-analyzed samples is not perfect (although high in **Chapter 3**, **4,** and **7**). This may lead to a loss of information and has the effect of lowering the estimate of heritability and the variation that can be explained by the GWAS in a PRS. In **Chapter 3**, we performed a meta-analysis of UK Biobank and 23andMe. Although this was the largest GWAS study performed to date, the explained variance in holdout samples did not exceed an estimated 3% explained variance, possibly due to imperfect correlations between these samples[38](r_g=0.69). Higher homogeneity between samples in data collection, study population and phenotype definition would have led to more homogeneity in the genetic signal, and subsequent better-powered PRS based on the meta-analysis. At the same time, meta-analysis provides an internal replication between samples, and loci that are shared between samples are likely to remain significant in the meta-analysis. For the same reason, we chose to meta-analyze several different test score domains in **Chapter 6** instead of performing GWAS on individual cognitive domains, in order to boost statistical power by combining correlated domain scores and find loci that are likely to be shared across different domains.

Fifth, we reported several gene-mapping strategies in these GWAS studies (see introduction, **Fig. 3**). Here, we chose an inclusive approach by mapping genes through

different mechanisms, including by gene-expression by eQTL, and by physical interactions through the 3D folding of the genome, showing considerable overlap in **Chapter 3**, **4**, and **5**. Because of this combination of different methods, we were able to map a large number of genes, which is likely an upper estimate of the number of genes that can be mapped based on these GWASs. We argue that none of these methods is superior to others and they contribute equal evidence, albeit through different mechanisms, that a gene may be implicated in a certain trait.

With regard to the PRS studies that were reported in **Chapter 2, 8** and **9,** it is important to consider that the usefulness of PRS for finding novel associations with behavior or brain phenotypes strongly depends on the availability of a sufficiently powered GWAS study. Although increasing much in size since the start of this thesis, increase in scale of GWAS studies of psychiatric disorders tend to be slow due to more hurdles in collecting patient data for genetic studies compared to healthy research participants. The asymmetry of sample sizes of GWAS studies used in the reported PRS studies may partly distort comparison of associations for psychiatric disorder PRS and PRS for cognitive tracts, since the educational attainment and intelligence GWAS studies were performed in much larger samples (educational attainment[39]: N=293,723, intelligence[40]: 269,867) than those in psychiatric disorders. In this light, several reported associations between PRS and the outcomes studied in this thesis, such as the link between the ADHD PRS and caudate nucleus volume, may be just the tip of the ice-berg, and may show much stronger associations with these outcomes or several new associations when the discovery sample size of GWAS in psychiatry increases.

In addition, there are considerations when interpreting the findings from brain imaging studies carried out in the second part of the thesis. In **Chapter 8**, we found significant negative associations between genetic risk for ADHD and the volume of the caudate nucleus. Although these results are interesting given prior evidence of ADHD symptoms and their association with volumes of this brain structure, the results do not generate hypotheses about the microstructural changes or mechanisms that can explain these smaller volumes and how they relate to genetic risk. Structural brain imaging, including T_1 and T_2-weighted sequences have traditionally been used to diagnose neurological disorders that alter brain anatomy, including brain tumors, white matter lesions and cerebrovascular disease. For these diseases, it is quite clear that the pathophysiological mechanism results in an abnormal appearance of the brain on CT or MRI. In psychiatric disorders, these pathophysiological mechanisms are much less clear, and future research is necessary to investigate how microstructural changes lead to smaller macrostructural volumes quantifiable on MRI. Considering that accepted hypotheses about disease mechanisms in psychiatry are on the microscopic scale (NMDA theory in schizophrenia[41], dopamine theory in ADHD[42]) it remains to be investigated how findings on the micro- and macrostructural scale can be reconciled.

Along these lines, the reported positive associations between polygenic scores and higher fractional anisotropy on DTI in **Chapter 9** need to be interpreted in the context of the underlying biological substrate of DTI measures. DTI models the primary diffusion directions of hydrogen within individual voxels, expressed in diffusion metrics such as fractional anisotropy (FA) and mean diffusivity (MD), which are thought to represent primarily myelinization and fiber density[43]. The interpretation of anisotropic diffusion in white matter varies widely in the neuroimaging literature[44], including 'white matter microstructure' and 'microstructural integrity'. Although the links between anisotropic diffusion and myelinization are evidenced by lower FA in demyelinating diseases such as multiple sclerosis[45,46] and clear overlap between DTI tracts and postmortem white matter bundles overlap quite elegantly, the knowledge of the correlation between diffusion measures and histological features of brain tissue is still lacking in the current literature. It remains to be seen whether a clear neurobiological substrate for DTI measures will be found for these measures, as it is currently not possible to study these in a living human brain. More detailed MR sequences of microscopic features in living tissue may further aid in understanding the exact neurobiological background of DTI.

4. Future directions

4.1 Big data and genetics

During the preceding four years of this thesis, GWAS discovery samples have seen a remarkable expansion in scale that reached the 'one-million-milestone' in 2018. The wealth of results that come from these studies demonstrates that this *big data* approach is not merely a 'buzz-term' but forms the basis for novel insights about human genetic architecture, now and in the years to come. Given the major steps forward in GWAS, it is often mentioned that we have already arrived in the *post-GWAS era*[47–49]. However, for many traits larger GWAS studies are just starting to uncover its genetic architecture. In this thesis, we have shown that larger studies are indeed leading to many more genes being associated with a trait: In our GWAS meta-analysis of insomnia in **Chapter 3**, we

show that an approximate tenfold increase in sample size lead to a hundred-fold increase in number of discovered independent loci compared to a previous GWAS[13]. These enormous samples raise the obvious question to what extent costly increases in study sample size will still return additional insights into the genetics of human behavior.

Next to these datasets containing a larger number of individuals (i.e. more rows in the data set), the improved coverage of new genotyping arrays and larger reference panels will lead to many more variants that can be tested (i.e. more columns in the data set), leading to large datasets that are even more difficult to handle[50]. This will require research groups to have access to more computational resources such as high performance and parallel computing, and necessitates the use of cloud-operated computing systems[51] such as the Hadoop distributed file system[52].

The collection of larger datasets (even well beyond 1 million individuals) in future studies is still going to be important for a number of reasons. First, many phenotype have not yet reached the tipping point where the number of discovered variants increases exponentially with larger samples, which strongly depend on the polygenicity and discoverability that make up the genetic architecture of a trait or disorder[7]. For many phenotypes, the first few genomic loci are just starting to be found, including ADHD[53], autism[54] and depression[55]. Second, the genetic signal within subgroups that constitute a phenotype definition have been shown to be heterogeneous[56]. Larger samples allow sufficient power to investigate genetic effects that are shared and specific to subgroups within broadly defined phenotypes. Third, larger datasets allow more complex statistical modelling including interactions between SNPs (epistasis[57,58]) and between SNPs and environmental factors (gene-environment interactions)[59,60]. Fourth, although the yield of novel loci may reduce with increasing sample size, an increase in statistical power of follow-up analysis based on these even better powered GWAS studies will be beneficial for detecting enrichment in novel pathways[61], tissue types and neuronal cell types. Fifth, polygenic scores based on even larger discovery samples will be a more powerful tool for a wide range of applications, including prediction of the same phenotype, and finding novel associations with related phenotypes. Especially for polygenic scores for disease-related phenotypes that may become clinically useful for risk stratification of patient groups, even small increases in explained variation may still be desirable. Whether researchers, consortia, funding agencies and scientific journals will indeed support the endeavor of even larger genetic studies beyond 1 million individuals remains an open question.

Several major sources of big data will lead to even larger genetic studies in the near future. First, in the coming year the UK Biobank is expected to release whole exome data in approximately 100,000 individuals, which will possibly lead to another wave of discoveries by fine-scale analyses of exonic regions in the genome. Second, following an example of the UK Biobank study, there are several biobank initiatives that gather data in the general population, including large-scale genotyping. These include the Kadoori Biobank[62] in China that includes 500,000 million individuals enrolled between 2004 and 2008 (which is also regulated open-access to researchers), the FinnGen biobank in Finland in up to 500,000 individuals by 2023, and the ambitious 'All of Us' research program carried out by the National Institute of Health (NIH)[63] that was initiated in 2016 and aims to include 1 million individuals throughout the US. Third, a wealth of high dimensional studies is already stored in electronic health record systems. Linking these records to genotype data forms a great opportunity to study genetic variation in relation to a massive number of outcomes[64,65]. An example of such an ongoing initiative is carried out by the Geisinger health care system[66] in the US, a data base that includes health care record data and exome chip data in over 100,000 patients. Fourth, direct-to-consumer genotyping companies such as 23andMe will continue to be a major source of data for genetic analysis, as this company is aiming to reach 10 million customers in the near future[67], with many opting in for use of their data in genetic research. These companies have clear advantages compared to the coordination of large cohort studies: the data set will continue to grow as new customers are being included to the customer database, and the online data collection is much more flexible as it can easily be updated by adding new questions to existing questionnaires. At the same time, phenotype operationalizations in online data collection like in 23andMe is still limited to online questionnaire data that is often quite general (referred to as "minimal phenotyping"[68]) and may not reach the level of phenotyping that is achieved at research centers of smaller population cohorts[69].

4.2 Precision phenotyping

In these ongoing increases in study scale, it becomes more challenging to measure phenotypes at a sufficient depth due to increasing costs and time constraints. To fully understand genomic risk loci related to psychiatric disorders, smaller-scale studies that contain extensively phenotyped individuals, including detailed disease biomarkers and omics data collected in longitudinal

designs, can provide more information about associations between risk loci and a multitude of (endo)phenotypes. With ongoing technological advances in the use of portable devices and biosensors, there are opportunities for continuous data collection in participants that are able to capture variation and fluctuation in measures longitudinally in contrast to single cross-sectional measures. In future studies, more data will be collected from electronic devices that can be carried by participants and continuously measure health-related data[70]. This digital phenotyping holds promise to be able to better capture behavioral determinants and outcomes through the use of sensory devices and continuous monitoring[71]. In combination with advanced artificial intelligence and deep learning algorithms, a myriad of complex phenotypes may be assessed or predicted from the collected data. Combining these fine-grain resolution data with previously collected genotype data, a much more detailed understanding of the link between genetic and phenotypic variation may be obtained.

4.3 Polygenic risk prediction

In this thesis, we showed the usefulness of PRSs to investigate early childhood reflections of genetic risk for psychiatric disorders using questionnaire (**Chapter 2**) and brain imaging data (**Chapter 8** and **9**).

Population-based studies have greatly benefited from using polygenic scores as a valuable estimate of overall genetic risk, leading to a variety of novel insights into associations between genetic predisposition and behavioral and health-related measures[72–75]. In the clinical setting, the utility of risk stratification for common diseases (cancer, cardiovascular and metabolic health) based on polygenic scores is expected to be even higher. Recently, many novel risk loci of common variants have been discovered for breast cancer[76] and colon cancer[77], and cardiovascular disease[78]. Polygenic scores based on these GWAS are shown to accurately separate high risk from low-risk groups, with individuals in the highest polygenic risk groups having a comparable risk (relative risk of 2 to 3 compared to the lowest quintile) as those that carry a rare monogenic mutation with a strong effect on the disease[79]. These relative risks for diseases that are already common in the population (breast cancer incidence in the Netherlands is roughly 1 in 8 or 13%[80]) means a very strong change in absolute risk of disease. By including polygenic scores in risk stratification, certain groups that were not identified as high risk may be shifted towards earlier screening risk categories, while others may be classified as at low risk may have less need for early screening[81].

For cardiovascular and coronary artery disease, individuals in certain unfavorable high genetic risk groups may undergo regular check-up, start cholesterol-lowering drugs earlier and may even more strongly advised to take up a healthier lifestyle[82]. In psychiatric disorders, a higher PRS in patients with schizophrenia has been linked to a higher likelihood of a more chronic disease course and one that is often more difficult to treat[83,84]. Based on genetic predisposition and symptoms during presentation, polygenic scores may eventually be used in presymptomatic healthy individuals to modulate well-known risk factors for the disease or patients with a high genetic load for the disease to treat certain risk groups with higher dose of medication or according to a more intensive follow-up routine. Given the lack of reliable biomarkers or diagnostic tests compared to diseases such as heart disease and cancer types, the implementation and validation of these scores in psychiatric patient groups will be much more difficult.

Several limitations exist that preclude the implementation of PRSs in routine clinical practice. First, most GWAS study samples exclude non-European subjects to improve external validity and prevent confounding due to population stratification. PRSs building forth on these results thus tend to have a poor predictive value in non-European populations[85]. Unless large-scale GWAS are carried out in populations of different ethnic backgrounds, the PRS will only be useful for a specific group of patients mostly from European ancestry. Second, studies reporting the value of combing PRSs with traditional modifiable and non-modifiable risk factors are scarce, and are necessary to estimate the value of genetic risk profiles in addition to risk estimation based on e.g. family history[86]. Third, validation studies are necessary to define the optimal threshold for defining high and low risk of disease. Current PRS studies estimate predictive accuracy in defining cases based on cross-sectional data. The optimal implementation and cut-off of genetic risk based on PRS requires long-term follow-up studies to estimate whether the PRS actually contributes to better treatment strategies and prognosis. Fourth, for many diseases, data collection has notoriously been slow due to difficulties in patient enrollment compared to the study of healthy individuals. PRSs based on smaller GWAS do not yet provide useful information, and use of these low-predictive scores leads to overdiagnosis or underestimation of risk[87]. Fifth, implementation of novel genetic tests requires patients, clinical geneticists, and health professionals to be well-informed about what PRSs exactly capture, before they can be considered in clinical practice.

4.4 Artificial intelligence in genetics

Another limitation in the predictive accuracy of PRS is the fact that GWAS and PRS have so far applied a linear model for estimation of SNP associations (GWAS) or prediction of the phenotype (PRS). Although linear models are highly interpretable and have excellent computational tractability for large datasets, it may not capture the complex non-linear associations of individual genetic variants that may be found in nature. In addition, PRS includes a weighted combination of thousands of SNP effects based on their linear association with the phenotype in GWAS which does not take the concurrent effects of other SNPs into account. Next to the classical statistical modelling techniques, there has been an emergence of algorithmic modelling in the last decade[88]. Machine learning and artificial intelligence using deep learning methods such as artificial neural networks[89] have shown promising results and improved prediction in almost all areas of scientific research, owing to its superior ability to capture non-linear relationships between predictor and outcome[90]. In the field of genetics, highly flexible algorithms such as neural networks are used to predict regulatory functions and pathogenicity of genetic variants[91,92]. While deep learning has shown to be promising in predicting the relevance of functional categories of variants in disease risk[93], prediction of phenotypes based on just genotype data has proven to be challenging due to the high computational demand, the high dimensional genetic data (i.e. many more variables than subjects), and complex level of layers through which variants act upon a phenotype[94]. This task of predicting phenotype from genotype is referred to as a 'supervised problem' in machine learning, as the training data of the algorithm is labeled with the true outcome that needs to be predicted in the unlabeled test dataset[95]. Moreover, given that genetic constitution and ethnic background are strongly related, prediction algorithms may be prone to predicting ancestry differences instead of actual disease status between groups[96].

Although several barriers still exist, machine learning and artificial intelligence are a promising and sophisticated alternative to the linear risk prediction models in genetics now that technological developments allow sufficient computing power for using these complex algorithms. Whether this will lead to better accuracy of targeting genetic risk predictors compared to standard linear models needs to be further explored in future research.

4.5 Clinical applications of GWAS results

There are several ways how the great abundance of knowledge brought forth by GWAS studies will contribute to improved clinical care of patients. First, GWAS has had an impact on nosology, the study of how we classify disease, by showing overlap in genetic factors and neurobiological mechanisms between seemingly distinct disorders[6], which may aid in future disease classification systems. Second, GWAS studies are pointing towards a large number of genes that can potentially be targeted with pharmacological intervention[97]. This opportunity is particularly needed for psychiatric disorders, where progress in pharmaceutical treatment over the last decades has been notoriously slow[98]. In addition, gene-targets may be identified that are currently targeted for other medical indications[97] (i.e. 'drug repositioning'), such as the calcium-channel genes (*CACNA1C*) that are implicated in schizophrenia[1], and are targeted with calcium-channel blockers to treat hypertension[99]. It is hoped that targeting these same gene products may indeed lead to alleviation of symptoms in schizophrenic patients in clinical trials. In **Chapter 4**, we observe a large number of genes implicated in neuroticism and depression that are known to interact with existing drugs, or genes that code for a protein or pathway that could potentially be targeted by novel drug therapies. In the clinical setting, genetic screening for drug susceptibility (e.g. cytochrome-related genes, or CYP) in patient that receive pharmaceutical may find opportunities for distinguishing 'rapid metabolizers' that are insensitive to the drug from those from those that are more sensitive and can be treated with lower doses. Third, as mentioned in previous sections, PRSs may contribute to diagnostic information for personalised screening, treatment strategies and prognosis, which may improve patient care in the near future. The results in **Chapter 2** illustrate that the PRS of schizophrenia may eventually be lead to additional information in distinguishing those at a low and high risk of developing internalizing (emotional) and externalizing (behavioral) problems. Future clinical studies are necessary to investigate whether the PRS may eventually have a potential role in more accurate diagnosis and ultimately in individual-tailored treatment plans in psychiatric care.

4.6 Neuroimaging

In the coming years, neuroimaging techniques will be an increasingly useful tool for studying relationships between genetic variance and structural or functional brain imaging-derived phenotypes. By using ultra-high-field imaging (field strength of 7 Tesla and higher[100]) MRI scanners will be able to capture more fine-grained images of the brain well below millimeter scale[101] at a faster pace. Also, the ongoing trend towards quantitative MRI will take a prominent place in neuroimaging[102], with strong

emphasis on novel sequences that assess microstructural characteristics of brain tissue, including MR spectroscopy for measuring neurotransmitter concentrations in the brain[103], and perfusion MRI to quantify the small vasculature of the brain. These fine-scale measurements of the brain will lead to a better visualization of variation in brain structure related to genomic variation that take place at a scale closer to the microscopic scale of genetic mechanisms.

In addition to further development of single MRI sequences, future brain imaging studies will increasingly use a combination (or integration) of different MRI sequences in a multimodal and multivariate brain imaging approach[104], which broadens the searchlight of finding neuroanatomical differences related to genetic variation. Also, novel approaches that integrate structural and functional sequences in network approaches may be better able to model the inner workings on the brain in connectome-based analyses compared to traditional structural imaging of the brain[105].

5. Conclusions

To conclude, the reported research findings contribute to our understanding of the genetic architecture of human behavior and the neurobiological tissues, cell-types and mechanisms through which they act. Genetic studies will continue to be central to lifting the veil of the infinitely complex biological background of human characteristics and will hopefully improve clinical care and human health and wellbeing. Despite the importance of larger sample sizes, big data and more advanced statistical modelling and software, the last decade of breakthroughs in genetics has above all taught us that these discoveries can only be achieved through teamwork and scientific collaboration.

References

1. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
2. Sullivan, P. Don't give up on GWAS. *Mol. Psychiatry* **17,** 2–3 (2012).
3. Posthuma, D. *et al.* The association between brain volume and intelligence is of genetic origin. *Nat. Neurosci.* **5,** 83–84 (2002).
4. Kelly, S. *et al.* Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group. *Mol. Psychiatry* **23**, 1261–1269 (2017).
5. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14,** 483–495 (2013).
6. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360,** eaap8757 (2018).
7. Holland, D. *et al.* Beyond SNP Heritability: Polygenicity and Discoverability Estimated for Multiple Phenotypes with a Univariate Gaussian Mixture Model. *bioRxiv* (2018).
8. Gertler, T. S., Chan, C. S. & Surmeier, D. J. Dichotomous anatomical properties of adult striatal medium spiny neurons. *J. Neurosci.* **28,** 10814–10824 (2008).
9. Cardinal, R. N., Parkinson, J. A., Hall, J. & Everitt, B. J. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci. Biobehav. Rev.* **26,** 321–352 (2002).
10. Cools, R. Dopaminergic control of the striatum for high-level cognition. *Curr. Opin. Neurobiol.* **21,** 402–407 (2011).
11. Pennartz, C. M. A. *et al.* The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *J. Neurosci.* **24,** 6446–6456 (2004).
12. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624-633 (2016).
13. Hammerschlag, A. R. *et al.* Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49,** 1584–1592 (2017).
14. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49,** 1107–1112 (2017).
15. Smith, A. J. P., Deloukas, P. & Munroe, P. B. Emerging applications of genome-editing technology to examine functionality of GWAS-associated variants for complex traits. *Physiol. Genomics* **50**, 510–522 (2018).
16. Dima, D. & Breen, G. Polygenic risk scores in imaging genetics: usefulness and applications. *J. Psychopharmacol.* **29,** 867–871 (2015).
17. Galesloot, T. E., Van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* **9,** e95923 (2014).
18. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369,** 1502–1511 (2013).
19. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312,** 1870–1879 (2014).
20. Middleton, A. *et al.* Attitudes of nearly 7000 health professionals, genomic researchers and publics toward the return of incidental results from sequencing research. *Eur. J. Hum. Genet.* **24,** 21–29 (2016).
21. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169,** 1177–1186 (2017).
22. Freund, M. K. *et al.* Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *bioRxiv* 324558 (2018).
23. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. FUMA: Functional mapping and annotation of genetic associations. *Nat. Commun.* **8,** 1826 (2017).
24. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11,** 1–19 (2015).
25. Dang, L. C. *et al.* Caudate asymmetry is related to attentional impulsivity and an objective measure of ADHD-like attentional problems in healthy adults. *Brain Struct. Funct.* **221,** 277–286 (2016).
26. Mataro, M., Garcia-Sanchez, C., Junque, C., Estevez-Gonzalez, A. & Pujol, J. Magnetic resonance imaging measurement of the caudate nucleus in adolescents with attention-deficit hyperactivity disorder and its relationship with neuropsychological and behavioral measures. *Arch. Neurol.* **54,** 963–968 (1997).
27. Grahn, J. A., Parkinson, J. A. & Owen, A. M. The cognitive functions of the caudate nucleus. *Prog. Neurobiol.* **86,** 141–155 (2008).
28. Brovelli, A., Nazarian, B., Meunier, M. & Boussaoud, D.

Differential roles of caudate nucleus and putamen during instrumental learning. *Neuroimage* **57,** 1580–1590 (2011).
29. Anderson, B. A. *et al.* Linking dopaminergic reward signals to the development of attentional bias: A positron emission tomographic study. *Neuroimage* **157,** 27–33 (2017).
30. Li, C. R., Yan, P., Sinha, R. & Lee, T.-W. Subcortical processes of motor response inhibition during a stop signal task. *Neuroimage* **41,** 1352–1363 (2008).
31. Tarver, J., Daley, D. & Sayal, K. Attention-deficit hyperactivity disorder (ADHD): an updated review of the essential facts. *Child. Care. Health Dev.* **40,** 762–774 (2014).
32. Hoogman, M. *et al.* Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults: a cross-sectional mega-analysis. *The Lancet Psychiatry* **4,** 310–319 (2017).
33. Valera, E. M., Faraone, S. V, Murray, K. E. & Seidman, L. J. Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **61,** 1361–1369 (2007).
34. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
35. Wolke, D. *et al.* Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *Br. J. Psychiatry* **195,** 249–256 (2009).
36. Martin, J. *et al.* Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am. J. Epidemiol.* **183,** 1149–1158 (2016).
37. Gustavson, K., von Soest, T., Karevold, E. & Røysamb, E. Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC Public Health* **12,** 918 (2012).
38. de Vlaming, R. *et al.* Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLoS Genet.* **13,** e1006495 (2017).
39. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
40. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
41. Coyle, J. T. NMDA receptor and schizophrenia: a brief history. *Schizophr. Bull.* **38**, 920–926 (2012).
42. Tripp, G. & Wickens, J. R. Neurobiology of ADHD. *Neuropharmacology* **57**, 579–589 (2009).
43. Beaulieu, C. The basis of anisotropic water diffusion in the nervous system–a technical review. *NMR Biomed.* **15**, 435–455 (2002).
44. Jones, D. K., Knösche, T. R. & Turner, R. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. *Neuroimage* **73**, 239–254 (2013).
45. Roosendaal, S. D. *et al.* Regional DTI differences in multiple sclerosis patients. *Neuroimage* **44**, 1397–1403 (2009).
46. Werring, D. J., Clark, C. A., Barker, G. J., Thompson, A. J. & Miller, D. H. Diffusion tensor imaging of lesions and normal-appearing white matter in multiple sclerosis. *Neurology* **52**, 1626–1632 (1999).
47. Wijmenga, C. & Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **50**, 322–328 (2018).
48. Marjoram, P., Zubair, A. & Nuzhdin, S. V. Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity (Edinb).* **112**, 79–88 (2014).
49. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS Era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
50. Stephens, Z. D. *et al.* Big data: astronomical or genomical? *PLoS Biol.* **13,** e1002195 (2015).
51. Agrawal, D., Das, S. & El Abbadi, A. Big data and cloud computing: current state and future opportunities. in *Proceedings of the 14th International Conference on Extending Database Technology* 530–533 (ACM, 2011).
52. Shvachko, K., Kuang, H., Radia, S. & Chansler, R. The hadoop distributed file system. in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on* 1–10 (Ieee, 2010).
53. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* 145581 (2017).
54. The ASD consortium of the Psychiatric Genetics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Mol. Autism* **8,** 1–17 (2017).
55. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50,** 668–681 (2018).
56. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & van der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun.* **9,** 905 (2018).
57. Carlborg, Ö. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* **5,** 618–625 (2004).
58. Cordell, H. J. Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10,** 392–404 (2009).
59. Thomas, D. Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* **11,** 259–272 (2010).
60. Hunter, D. J. Gene–environment interactions in human diseases. *Nat. Rev. Genet.* **6,** 287-298 (2005).
61. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17,** 353–364 (2016).
62. Li, L. M. *et al.* The China Kadoorie Biobank: related methodology and baseline characteristics of the participants. *Zhonghua liu xing bing xue za zhi* **33,** 249–255 (2012).
63. Sankar, P. L. & Parker, L. S. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet. Med.* **19,** 743–750 (2017).
64. Wolford, B. N., Willer, C. J. & Surakka, I. Electronic health records: the next wave of complex disease genetics. *Hum. Mol. Genet.* **27,** R14–R21 (2018).
65. Lakhani, C. M. *et al.* Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat. Genet.* **51**, 327–334 (2019).
66. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genet. Med.* **18,** 906-913 (2016).
67. Check Hayden, E. The rise and fall and rise again of 23andMe. *Nat. News* **550,** 174–177 (2017).
68. Cai, N., Kendler, K. & Flint, J. Minimal phenotyping yields GWAS hits of low specificity for major depression. *bioRxiv* 440735 (2018).
69. Abbasi, J. 23andMe, big data, and the genetics of depression. *JAMA* **317,** 14–16 (2017).
70. Torous, J., Onnela, J. P. & Keshavan, M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Transl. Psychiatry* **7,** e1053 (2017).
71. Insel, T. R. Digital phenotyping: technology for a new science of behavior. *JAMA* **318,** 1215–1216 (2017).
72. Nivard, M. G. *et al.* Genetic overlap between schizophrenia and

developmental psychopathology: longitudinal and multivariate polygenic risk prediction of common psychiatric traits during development. *Schizophr. Bull.* **43,** 1197–1207 (2017).
73. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The Lancet Psychiatry* **4,** 57–62 (2017).
74. Riglin, L. *et al.* Association of genetic risk variants with attention-deficit/hyperactivity disorder trajectories in the general population. *JAMA psychiatry* **73,** 1285–1292 (2016).
75. Krapohl, E. *et al.* Phenome-wide analysis of genome-wide polygenic scores. *Mol. Psychiatry* **21,** 1188–1193 (2016).
76. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94 (2017).
77. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2018).
78. Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).
79. Khera, A. V *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50,** 1219–1224 (2018).
80. Netherlands Cancer Registry. Comprehensive Cancer Centre the Netherlands (IKNL) (2014).
81. Li, H. *et al.* Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. *Genet. Med.* **19,** 30–35 (2017).
82. Khera, A. V *et al.* Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* **375,** 2349–2358 (2016).
83. Meier, S. M. *et al.* High loading of polygenic risk in cases with chronic schizophrenia. *Mol. Psychiatry* **21,** 969–974 (2016).
84. Zhang, J.-P. *et al.* Schizophrenia polygenic risk score as a predictor of antipsychotic efficacy in first-episode psychosis. *Am. J. Psychiatry* **176**, 21–28 (2018).
85. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nat Genet.* **50**, 1112–1121 (2018).
86. Maas, P. *et al.* Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2,** 1295–1302 (2016).
87. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9,** e1003348 (2013).
88. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16,** 199–231 (2001).
89. Hopfield, J. J. Artificial neural networks. *IEEE Circuits Devices Mag.* **4,** 3–10 (1988).
90. Musib, M. *et al.* Artificial intelligence in research. *Science* **357,** 28–30 (2017).
91. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50,** 1171–1179 (2018).
92. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16,** 321–332 (2015).
93. Zhou, J. *et al.* Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism. *bioRxiv* 319681 (2018).
94. Leung, M. K. K., Delong, A., Alipanahi, B. & Frey, B. J. Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE* **104,** 176–197 (2016).
95. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning.* **112,** (Springer, 2013).
96. Robinson, E. B. *et al.* Response to 'Predicting the diagnosis of autism spectrum disorder using gene pathway analysis'. *Mol. Psychiatry* **19,** 859–861 (2014).
97. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30,** 317–320 (2012).
98. Breen, G. *et al.* Translating genome-wide association findings into new therapeutics for psychiatry. *Nat. Neurosci.* **19,** 1392–1396 (2016).
99. Lencz, T. & Malhotra, A. K. Targeting the schizophrenia genome: a fast track strategy from GWAS to clinic. *Mol. Psychiatry* **20,** 820–826 (2015).
100. Ladd, M. E. *et al.* Pros and cons of ultra-high-field MRI/MRS for human application. *Prog. Nucl. Magn. Reson. Spectrosc.* **109,** 1–50 (2018).
101. Krafft, O., Fischer, A., Nagel, A. M., Mönninghoff, C. & Ladd, M. E. MRI at 7 Tesla and above: demonstrated and potential capabilities. *J. Magn. Reson. Imaging* **41,** 13–33 (2015).
102. Tofts, P. *Quantitative MRI of the brain: measuring changes caused by disease.* (John Wiley & Sons, 2005).
103. Soares, D. P. & Law, M. Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clin. Radiol.* **64,** 12–21 (2009).
104. Sui, J., Adali, T., Yu, Q., Chen, J. & Calhoun, V. D. A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods* **204,** 68–81 (2012).
105. Fox, M. D. Mapping symptoms to brain networks with the human connectome. *N. Engl. J. Med.* **379,** 2237–2245 (2018).

Nederlandse samenvatting

Het doel van mijn thesis was het beantwoorden van een tweetal onderzoeksvragen:
1) Het eerste onderzoeksdoel was om de genetische factoren te begrijpen die de erfelijkheid verklaren van menselijk gedrag in de algemene populatie Het belangrijkste beginpunt van alle analyses die hiernaar zijn uitgevoerd is een genoomwijde associatie studie (GWAS) om voor vele SNPs (genetische varianten) in het genoom de associatie met de uitkomstmaat te verkrijgen. Deze SNP-effecten kunnen vervolgens worden gebruikt als input voor velerlei analyses die daarop volgen, waarmee genen en neurobiologische mechanismen gevonden kunnen worden, en om het totale genetische risico met polygene risicoscores te schatten.
2) Het tweede doel was om magnetic resonance imaging (MRI) van het brein toe te passen, verzameld in een paar duizend individuen, waaronder kinderen en volwassenen, om te verklaren hoe genetische variatie tussen individuen leidt tot verschillen in de morfologie van het brein. Om dit te onderzoeken voerden we een genoomwijde associatie studie uit en gebruikten we polygene risicoscores om vast te stellen of verschillen in genetische predispositie voor psychiatrische ziekten leiden tot een andere hersenontwikkeling in een studiepopulatie van kinderen.

In **Hoofdstuk 2** gebruikten we genetische data en gedragsmaten, verzameld in meer dan 2,000 kinderen vanaf de leeftijd van 3 jaar en hoger, om te onderzoeken of genetische aanleg voor psychiatrische ziektebeelden geassocieerd is met afwijkend gedrag op jonge kinderleeftijd. Door polygene risico scores van vijf psychiatrische ziektebeelden te berekenen, konden we aantonen dat variatie in genetische predispositie voor schizofrenie geassocieerd is met verschillen in door hun moeder gerapporteerde gedragsproblemen vanaf de leeftijd van drie jaar, met name in internaliserende problemen. Deze associaties bleven bestaan tijdens follow up op de leeftijd van zes en 10 jaar. De polygene score voor opleidingsniveau daarentegen was negatief geassocieerd met vrijwel alle gedragsschalen en liet de sterkste negatieve associatie zien met aandachtsproblemen. Deze bevindingen lieten duidelijk zien dat gedragsmanifestaties zijn geassocieerd met genetisch risico in de vroege kindertijd, en suggeren dat polygene risico scores voor deze ziekten mogelijk een rol kunnen spelen in het vroeg identificeren van kinderen met het grootste risico op gedragsproblemen, en een mogelijk risico voor schizofrenie op volwassen leeftijd. Dit suggereert dat verschillen in hersenstructuur en functie gerelateerd aan het genetisch risico mogelijk op vroege leeftijd aanwezig zijn.

In **Hoofdstuk 3** voerden we de grootste genetische studie tot nu toe uit door een meta-analyse van GWAS studies naar slapeloosheid uit te voeren in meer dan 1,000,000 individuen in twee grootschalige cohorten waarin slaapvragenlijsten en genetische data verzameld zijn. We identificeerden meer dan 200 loci en bijna 1,000 genen. De resultaten lieten zien dat veel van het genetische signaal voor slapeloosheid overlapt met een ruim aantal psychiatrische ziektebeelden, waaronder angststoornissen en depressie. Gebaseerd op Mendelian Randomization met GWAS resultaten konden wij laten zien dat slapeloosheid een causale lijkt te spelen in obesitas, type 2 diabetes en coronaire hartziekten, maar niet vice versa. Door RNA sequencing data in individuele neuronen te analyseren, lieten we zien date er verrijking is van genetisch signaal in neuronen gelocaliseerd in onder andere de hypothalamus en het claustrum. Dit bevestigt de rol van deze hersenstructuren in slaapregulatie. Deze resultaten tonen nieuwe genen en celsoorten aan als mogelijk nieuw doel van functioneel vervolgonderzoek.

Om de genetische architectuur van neuroticisme en depressie beter te begrijpen, voerden we in **Hoofdstuk 4** een grootschalige meta-analyse uit van GWAS naar neuroticisme en depressie in meer dan 400,000 individuen. Deze analyses onhulden meer dan 100 loci in het genoom die betrokken zijn bij neuroticisme. Door een clusteranalyse uit te voeren op alle neuroticisme items van de vragenlijst lieten we zien dat de totaalscore van neuroticisme genetisch heterogeen is en deze individuele items kunnen worden onderverdeeld in subgroepen die verschillen in significante SNPs, gen associaties en genetische correlaties met eerdere GWAS studies. Voor depressie toonden we bewijs dat serotonerge neuronen in het brein betrokken zijn. Dit bevestigt de rol van serotonerge mechanismen als doel van serotonerg aangrijpende antidepressiva. Ook vonden we vele nieuwe

genen waarvan bekend is dat ze interacties vertonen in een grote referentiedatabase naar farmacologische interactie met bepaalde medicijnen, en die mogelijk hierdoor een nieuw doel vormen voor behandeling.

In **Hoofdstuk 5** combineerden we GWAS resultaten van cognitieve test scores verzameld in 14 cohorten en meer dan 200,000 individuen om genen te vinden die geassocieerd zijn met intelligentie. We waren in staat om meer dan 200 geassocieerde locaties in het DNA te ontdekken waarin variaties gerelateerd zijn aan variatie in intelligentie. Een groot deel, hiervan waren nieuwe ontdekkingen. We lieten met genexpressie zien dat verschillende hersenregio's significant verrijkt zijn voor genen betrokken bij intelligentie, met name corticale gebieden van het brein. Ook vonden we significante verrijking van genetisch signaal in verschillende soorten pyramidale neuronen, en neuronen gelegen in de CA1 regio van de menselijke hippocampus. Deze bevindingen zijn in overeenstemming met significante verrijking van genetisch signaal in gen-sets en biologische mechanismen betrokken bij neurogenese, de aanmaak van neuronen, een belangrijk process dat in de hippocampus plaats vindt. These grootschalige analyse van intelligentie biedt vele links tussen genetische variatie en variatie in intelligentie scores, en duidt specifieke mechanismen aan die deze link kunnen verklaren.

In het tweede deel van deze thesis, gebruikten we magnetic resonance imaging (MRI) van het brein van zowel kinderen als volwassenen uit grote populatiestudies om de genetische effecten op brain morfologie en witte stof microstructuur te bestuderen.

Het gebruik van grootschalige data van scans van het menselijk lichaam heeft het duidelijke voordeel dat er veel statististische rekenkracht in deze data zit, waardoor subtiele associaties met genetische en omgevingsfactoren gedetecteerd kunnen worden. Een grote dataverzameling gaat echter ook gepaard met een grote kans op het vinden van toevalsbevindingen die gezondheidsconsequenties voor de deelnemer kunnen hebben. In **Hoofdstuk 6** rapporteren we de prevalentie van incidentele bevindingen op hersen MRI scans die werden waargenomen tijdens de grootschalige verzameling van hersen MRI data in meer dan 4,000 kinderen. Onze resultaten lieten duidelijk zien dat incidentele bevindingen zeer prevalent zijn in kinderen, en dat nauwkeurige screening van deze data en een gestreamlined protocol voor de klinische follow-up hiervan zeer belangrijk zijn. De incidentele bevindingen betroffen zeven kinderen waar de klinische verdenking op een hersentumor bestond. Van hen zijn twee kinderen geopereerd als direct gevolg van het meedoen aan het populatie onderzoek.

In **Hoofdstuk 7**, deden we een GWAS meta-analyse van brein volume (BV) op basis van hersen MRI data in meer dan 40,000 individuen. We vonden vele nieuwe genetische loci voor breinvolume en vonden genen betrokken bij verschillende interessante signaleringsmechanismen in het brein gerelateerd aan celdeling, celdifferentiatie en apoptose regulatie. Interessant genoeg vonden dat veel van deze genen gelinkt zijn aan het optreden van monogene ziekte door zeldzame varianten en ziektebeelden waar vaak een abnormale hersenontwikkeling bij optreedt, zoals micro- macro- en megalencephalie. Eerder bewijs suggereert dat de fenotypische correlatie tussen brein volume en intelligentie wordt verklaard door overlap in genetische factoren. Om genen te vinden die betrokken zijn in zowel brein volume als intelligentie, voerden we een uitgebreide genetische analyse uit van de overlap in loci en genen tussen intracranieel volume en onze genetische studie naar intelligentie. De functie van deze genen is met name gerelateerd aan signaalmechanismen betrokken bij celcyclusregulatie. Deze resultaten zijn een grote stap vooruit in het begrip van genen geassocieerd met brein volume, en illustreren hoe GWAS resultaten kunnen worden gebruikt om interessante overlap in genetische factoren te vinden tussen genetisch gecorreleerde eigenschappen.

Voortbouwend op deze beinvingen, gebruikten we in **Hoofdstuk 8** polygene risico scores voor psychiatrische ziektebeelden en cognitie-gerelateerde eigenschappen om te bestuderen of variatie in hersenstructuur verklaard kan worden door verschillen in genetische predispositie. Er waren sterke positieve associaties tussen cognitie-gerelateerde genetische scores en brein volume, en suggestieve associaties in de negatieve richting met ADHD polygene risicoscores en brein volume. Meer specifiek toonden we associaties aan tussen hogere polygene risicoscores voor ADHD en kleinere volumes van de nucleus caudatus, dit effect was met name in jongens waarneembaar. Door gedragsuitkomsten van ADHD symptomen in deze kinderen te gebruiken, konden we aantonen dat het volume van de nucleus caudatus de associatie tussen ADHD polygene scores en ADHD symptomen medieerde, en dat deze structuur de associatie tussen predispositie voor ADHD en ADHD symptomen verklaart.

In het daarop volgende **Hoofdstuk 9**, berekenden

we polygene risico scores voor vijf psychiatrische ziektebeelden (ADHD, autism, bipolaire stoornis, depressie en schizofrenie), en voor cognitie-gerelateerde eigenschappen (opleidingsniveau en intelligentie) om te onderzoeken of genetische aanleg voor deze eigenschappen kan worden vertaald naar een abnormaal ontwikkelingspatroon van de witte stofbanen van het brein. We gebruikten opnieuw DTI, verzameld in een grote groep deelnemers, om de microstructuur van witte stof te onderzoeken. Hoewel schizofrenie herhaaldelijk in verband is gebracht met een abnormale ontwikkeling van witte stof op hersenimaging data, vonden we geen associatie tussen genetisch risico voor schizofrenie en witte stof integriteit. Daarentegen vonden we duidelijk bewijs dat genetische predispositie voor cognitie-gerelateerde eigenschappen zoals opleidingsniveau en intelligentie geassocieerd zijn met hogere fractional anisotropy (FA) van meerdere witte stof banen, dit suggereert een meer optimale myelinizatie van witte stof banen van het brein in kinderen met een hogere genetische aanleg voor intelligentie en opleidingsniveau later in het leven.

List of publications

Accepted / published peer-reviewed articles

1. Eindhoven, J.A., van den Bosch, A.E., **Jansen, P.R.**, Boersma, E., & Roos-Hesselink, J.W., The usefulness of brain natriuretic peptide in complex congenital heart disease: a systematic review. *Journal of the American College of Cardiology*, **60**, 2140-2149 (2012).

2. Sniekers, S., Stringer, S., Watanabe, K., **Jansen, P.R.**, Coleman, J.R., Krapohl, E., Taskesen, E., Hammerschlag, A.R., Okbay, A., Zabaneh, D., Amin, N., Breen, G., Cesarini, D., Chabris, C.F., Iacono, W.G., Ikram, M.A., Johannesson, M., Koellinger, P., Lee, J.J., Magnusson, P.K.E., McGue, M., Miller, M.B., Ollier, W.E.R., Payton, A., Pendleton, N., Plomin, R., Rietveld, C.A., Tiemeier, H., van Duijn, C.M., Posthuma, D., Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics*, **49**, 1107-1112 (2017).

3. Kelly, S., Jahanshad, N., Zalesky, A., Kochunov, P., Agartz, I., Alloza, C., Andreassen, O.A., Arango, C., Banaj, N., Bouix, S., Bousman, C.A., Brouwer, R.M., Bruggemann, J., Bustillo J., Cahn, W., Calhoun, V., Cannon, D., Carr, V., Catts, S., Chen. J., Chen, J-x, Chen, X., Chiapponi, C., Cho, Kl. K., Ciullo, V., Corvin, A.S., Crespo-Facorro, B., Cropley, V., De Rossi, P., Diaz-Caneja, C.M., Dickie, E.W., Ehrlich, S., Fan, F-m., Faskowitz, J., Fatouros-Bergman, H., Flyckt, L., Ford, J.M., Fouche, J-P., Fukunaga, M., Gill, M., Glahn, D.C., Gollub, R., Goudzwaard, E.D., Guo, H., Gur, R.E., Gur, R.C., Gurholt, T.P., Hashimoto, R., Hatton, S.N., Henskens, F.A., Hibar, D.P., Hickie, I.B., Hong, L.E., Horacek, J., Howells, F.M., Hulshoff Pol, H.E., Hyde, C.L., Isaev, D., Jablensky, A., **Jansen, P.R.**, Janssen, J., Jönsson, E.G., Jung, L.A., Kahn, R.S., Kikinis, Z., Liu K., Klauser, P., Knöchel, C., Kubicki, M., Lagopoulos, J., Langen, C., Lawrie, S., Lenroot, R.K., Lim, K.O., Lopez-Jaramillo, C., Lyall, A., Magnotta, V., Mandl, R.C.W., Mathalon, D.H., McCarley, R.W., McCarthy-Jones, S., McDonald, C., McEwen, S., McIntosh, A., Melicher, T., Mesholam-Gately, R.I., Michie, P.T., Mowry, B., Mueller, B.A., Newell, D.T., O'Donnell, P., Oertel-Knöchel, V., Oestreich, L., Paciga, S.A., Pantelis, C., Pasternak, O., Pearlson, G., Pellicano, G.R., Pereira, A., Pineda Zapata, J., Piras, F., Potkin, S.G., Preda, A., Rasser, P.E., Roalf, D.R., Roiz, R., Roos, A., Rotenberg, D., Satterthwaite, T.D., Savadjiev, P., Schall, U., Scott, R.J., Seal, M.L., Seidman, L.J., Shannon Weickert, C., Whelan, C.D., Shenton, M.E., Kwon, J.S., Spalletta, G., Spaniel, F., Sprooten, E., Stäblein, M., Stein, D.J., Sundram, S., Tan, Y., Tan, S., Tang, S., Temmingh, H.S., Westlye, L.T., Tønnesen, S., Tordesillas-Gutierrez, D., Doan, N.T., Vaidya, J., van Haren, N.E.M., Vargas, C.D., Vecchio, D., Velakoulis, D., Voineskos, A., Voyvodic, J.Q., Wang, Z., Wan, P., Wei, D., Weickert, T.W., Whalley, H., White, T., Whitford, T.J., Wojcik, J.D., Xiang, H., Xie, Z., Yamamori, H., Yang, F., Yao, N., Zhang, G., Zhao, J., van Erp, T.G.M., Turner, J., Thompson, P.M. & Donohoe, G. Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group., *Molecular Psychiatry,* **23**, 1261-1269 (2017).

4. **Jansen, P.R.**, Dremmen, M., van den Berg, A., Dekkers, I.A., Blanken, L. M., Muetzel, R. L., Bolhuis K, Mulder, R.M., Kocevska, D., Jansen, T.A., de Wit, M.C.Y., Neuteboom, R.F., Polderman, T.J.C., Posthuma D., Jaddoe, V.W.V., Verhulst, F.C., Tiemeier, H., van der Lugt, A., White, T., Incidental findings on brain imaging in the general pediatric population. *New England Journal of Medicine*, **377**, 1593-1595 (2017).

5. Tielbeek, J. J., Johansson, A., Polderman, T. J., Rautiainen, M. R., **Jansen, P.R.**, Taylor, M., Tong, X., Lu, Q., Burt, A.S., Tiemeier, H., Viding, E., Plomin, R., Martin, N.G., Heath, A.C., Madden, P.A.F., Montgomery, G., Beaver, K.M., Waldman, I., Gelernter, J., Kranzler, H.R., Farrer, L.A., Perry, J.R.B., Munafò, M., LoParo, D., Paunio, T., Tiihonen, J., Mous, S.E., Pappa, I., de Leeuw, C., Watanabe, K., Hammerschlag, A.R., Salvatore, J.E., Aliev,

F., Bigdeli, T.B., Dick, D., Faraone, S.V., Popma, A., Medland, S.E., Posthuma, D., Broad Antisocial Behavior Consortium collaborators. Genome-wide association studies of a broad spectrum of antisocial behavior. *JAMA Psychiatry*, **74**, 1242-1250 (2017).

6. Serdarevic, F., **Jansen, P.R.,** Ghassabian, A., White, T., Jaddoe, V.W., Posthuma, D., Tiemeier, H., Association of genetic risk for schizophrenia and bipolar disorder with infant neuromotor development. *JAMA Psychiatry*, **75**, 96-98 (2018).

7. J**ansen, P.R.**, Polderman, T.J., Bolhuis, K., van der Ende, J., Jaddoe, V.W., Verhulst, F.C., White, T., Posthuma, D., Tiemeier, H., Polygenic scores for schizophrenia and educational attainment are associated with behavioural problems in early childhood in the general population. *Journal of Child Psychology and Psychiatry*, **59**, 39-47 (2018).

8. White, T., Muetzel, R.L., El Marroun, H., Blanken, L.M., **Jansen, P.R**., Bolhuis, K., Kocevska, D., Mous, S.E., Mulder, R., Jaddoe, V.W.V., van der Lugt, A., Verhulst, F.C., Tiemeier, H., Paediatric population neuroimaging and the Generation R Study: the second wave. *European Journal of Epidemiology*, **33**, 99-125, (2018).

9. White, T., **Jansen, P.R.**, Muetzel, R. L., Sudre, G., El Marroun, H., Tiemeier, H., Qiu, A., Shaw, P., Michael, A.M., Verhulst, F.C. Automated quality assessment of structural magnetic resonance images in children: Comparison with visual inspection and surface-based reconstruction. *Human Brain Mapping*, **39**, 1218-1231 (2018).

10. Coleman, J.R., Bryois, J., Gaspar, H.A., **Jansen, P.R.**, Savage, J.E., Skene, N., Plomin R., Muñoz-Manchado, A.B., Linnarsson, S., Crawford, G., Hjerling-Leffler, J., Sullivan, P.F., Posthuma, D., Breen, G., Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87,740 individuals. *Molecular Psychiatry*, **24**, 182-197 (2018).

11. Szekely, E., Schwantes-An, T. H.L., Justice, C.M., Sabourin, J.A., **Jansen, P.R.**, Muetzel, R. L., Sharp., W., Tiemeier, H., Sung, H., White, T., Wilson, A.F., Shaw, P., Genetic associations with childhood brain growth, defined in two longitudinal cohorts. *Genetic Epidemiology*, **42**, 405-414 (2018).

12. Nagel*, M., **Jansen*, P.R.**, Stringer, S., Watanabe, K., de Leeuw, C.A., Bryois, J., Savage, J.E., Hammerschlag, A.R., Skene, N.G., Muñoz-Manchado, A.B., 23andMe Research Team, White, T., Tiemeier, H., Linnarsson, S., Hjerling-Leffler, J., Polderman, T.J.C., Sullivan, P.F., van der Sluis, S., Posthuma, D., Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature Genetics*, **50**, 920-927 (2018).

13. Savage*, J.E., **Jansen*, P.R.**, Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., Nagel, M.,Awasthi, S., Barr, P.B., Coleman, J.R.I., Gasby, K.L., Hammerschlag, A.R., Kaminski, J.A., Karlsson, R., Krapohl, E., Lam, M., Nygaard, M., Reynolds, C.A., Trampush, J.W., Young, H., Zabaneh, D., Hagg, S., Hansell, N.K., Karlsson, I.K., Linnarsson, S., Montgomery, G.W., Munoz-Manchado, A.B., Quinlan, E.B., Schumann, G., Skene, N.G., Webb, B.T., White, T., Arking, D.E., Avramopoulos, D., Bilder, R.M., Bitsios, P., Burdick, K.E., Cannon, T.D., Chiba-Falek, O., Christoforou, A., Cirulli, E.T., Congdon, E., Corvin, A., Davies, G., Deary, I.J., DeRosse, P., Dickinson, D., Djurovic, S., Donohoe, G., Conley, E.D., Eriksson, J.G., Espeseth, T., Freimer, N.A., Giakoumaki, S., Giegling, I., Gill, M., Glahn, D.C., Hariri, A.R., Hatzimanolis, A., Keller, M.C., Knowles, E., Koltai, D., Konte, B., Lahti, J., Le Hellard, S., Lencz, T., Liewald, D.C., London, E., Lundervold, A.J., Malhotra, A.K., Melle, I., Morris, D., Need, A.C., Ollier, W., Palotie, A., Payton, A., Pendleton, N., Poldrack, R.A., Raikkonen, K., Reinvang, I., Roussos, P., Rujescu, D., Sabb, F.W., Scult, M.A., Smeland, O.B., Smyrnis, N., Starr, J.M., Steen, V.M., Stefanis, N.C., Straub, R.E., Sundet, K., Tiemeier, H., Voineskos, A.N.,Weinberger, D.R., Widen, E., Yu, J., Abecasis, G., Andreassen, O.A., Breen, G., Christiansen, L., Debrabant, B., Dick, D.M., Heinz, A., Hjerling-Leffler, J., Ikram, M.A., Kendler, K.S., Martin, N.G., Medland, S.E., Pedersen, N.L., Plomin, R., Polderman, T.J.C., Ripke, S., van der Sluis, S., Sullivan, P.F., Vrieze, S.I., Wright, M.J., Posthuma, D., Genome-wide association meta-analysis in 269,867

individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, **50**, 912-919 (2018).

14. **Jansen, P.R.**, Petrus, N., Venema, A., Posthuma, D., Mannens, M., Sprikkelman, A., Henneman, P. . Higher Polygenetic Predisposition for Asthma in Cow's Milk Allergic Children. *Nutrients*, **10**, 158 (2018).

15. **Jansen, P.R.**, Muetzel, R. L., Polderman, T. J., Jaddoe, V. W., Verhulst, F. C., van der Lugt, A., Tiemeier, H., Posthuma, D., White, T. , Polygenic Scores for Neuropsychiatric Traits and White Matter Microstructure in the Pediatric Population. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, **4**, 243-250 (2019).

16. Alemany, S., **Jansen, P.R.**, Muetzel, R.L., Marqués, N., Marroun, H.E., Jaddoe, V.W.V., Polderman, T.J.C., Tiemeier H., Posthuma, D., White, T. Common polygenic variation for psychiatric disorders and cognitive traits and structural brain imaging in the general pediatric population, *Journal of the American Academy for Child and Adolescent Psychiatry (In Press).*

17. **Jansen, P.R.**, Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A.R., .de Leeuw, C.A., Benjamins, J.S., Munoz-Manchado, A.B., Nagel, M., Savage, J.E., Tiemeier, H., White, T., Tung, J.Y., Hinds, D.A., Vacic, V., Wang, X., Sullivan, P.F., van der Sluis, S., Polderman, T.J.C., Smit, A.B., Hjerling-Leffler, J., Van Someren, E.J.W., Posthuma, D., Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies Novel Loci and Functional Pathways. *Nature Genetics*, **51**, 394-403 (2019).

18. Liu, M., Jiang, Y. , Wedow, R., Li, Y., Brazel, D., Chen, F., Datta, G., Zhan, X., Docherty, A., Faul, J., Foerster, J., Gordon, S., Haessler, J., Hottenga, J., **Jansen, P.R.**, Ling, Y.Y., Palviainen, T., Pandit, A., Smith, J., Taylor, A., Turman, C., Young, H., Zajac, G., Zhao, W., Boardman, J., Boehnke, M., Boomsma, D., Chen, C., Cucca, F., Eaton, C., Ehringer, M., Gillespie, N., Haller, T., Mullan Harris, K., Heath, A., Hewitt, J., Hokanson, J., Hopfer, C., Iacono, W., Johnson, E., Kardia, S., Keller, M., Kellis, M., Kooperberg, C., Kraft, P., Krauter, K., Laakso, M., Lind, P., Loukola, A., Lutz, S., Madden, P., Martin, N., McGue, M., McQueen, M., Medland, S., Mohlke, K., Nielsen, J., Okada, Y., Peters, U., Polderman, T., Posthuma, D., Kaprio, J., Stitzel, J., Reiner, A., Stallings, M., Tindle, H., Wall, T., Weir, D., Whitfield, J., Zucculo, L., Bierut, L., Hveem, K., Lee, J., Munafo, M., Saccone, N., Willer, C., Cornelis, M., David, S., Jorgenson, E., Stefansson, K., Hinds, D., Thorgeirsson, T., Liu, D., Abecasis, G., Davila-Velderrain J., McGuire, D., Tian, C., Choquet, H., Gabrielsen, M., Huang, H., Matoba, N., Mägi, R., McMahon, G., Mulas, A., Orrù, V., Reginsson, G., Skogholt, A., Willemsen, G., Young, K., Bjornsdottir, G., Davies, G., Esko, T., Fiorillo, E., Gudbjartsson, D., Hickie, I. , Hunter, D., Kamatani, Y., Metspalu, A., Rice, J., Rimm, E., Rose, R., Runarsdottir, V., Stančáková, A., Stefansson, H., Thai, K., Tyrfingsson, T., Weisner, C., Winsvold, B., Yin, J. , Fritsche, L., Jang, S. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, **51**, 237-244 (2019).

19. Dekkers, I.A., **Jansen, P.R.**, Lamb, H.J., Obesity is negatively associated with global and subcortical brain volumes: a cross-sectional population-based imaging study in the UK Biobank. *Radiology,* **23**, 181012 (2019).

20. Bolhuis, K., Tiemeier, H., **Jansen, P.R,** Jaddoe, V., Neumann, A., vd Akker, E., van Rossum, E., White, T., Hillegers, M., Muetzel, R.L., Kushner, S., Interaction of schizophrenia polygenic risk and cortisol level on pre-adolescent brain structure. *Psychoneuroendocrinology,* **101**, 295-303 (2019).

21. **Jansen, P.R.**, Broeders, M.J.M., Cornel, M.C., Meijers-Heijboer, H., Polygenetische risico scores voor vaak voorkomende ziekten: van epidemiologie naar klinische toepassing *(Nederlands Tijdschrift voor Geneeskunde, provisionally accepted).*

Manuscripts submitted / in prepatation

22. Serdarevic, F., Tiemeier, H., **Jansen, P.R.,** Alemany, S., Xerxa, Y., Hillegers, M.H.J., Verhulst, F.C., Ghassabian, A., Polygenic risk scores for developmental disorders, neuromotor functioning during infancy, and autistic traits in childhood *(Submitted).*

23. Wei, Y., de Lange, S.C., Scholtens, L.H., Ardesch, D.J., **Jansen, P.R.,** Savage, J.E., Watanabe, K., Li, L., Preuss, T.M. Rilling, J.K., Posthuma, D., van den Heuvel, M.P., Genetic correlates of evolutionary changes in the human default mode network *(Submitted).*

24. **Jansen, P.R.**, Vrijkotte, T.G., Tielbeek, J.J. BroadABC consortium, Ensink, J.B.M., Jaddoe V.W.V., Tiemeier T., White, T., Zafarmand, M.H., Polderman, T.J.C., Externalizing problem behavior in young children: the effect of maternal smoking during pregnancy and genetic risk *(Submitted).*

25. Giusti-Rodríguez, P., Lu, L., Yang, Y., Crowley, C.A., Liu, X., Juric, I., Martin, J.S., Abnousi, A., Allred, S.C. , Ancalade, N., Bray, N.J., Breen G., Bryois, J., Bulik, C.M., Crowley, J.J., Guintivano, J.R., **Jansen, P.R.**, Jurjus, G.J., Li, Y., Mahajan, G., Marzi, S., Mill, J., O'Donovan, M.C., Overholserm, J.C., Owen, M.J., Pardiñas, A.F., Pochareddy, S., Posthuma D., Rajkowska, G., Santpere, G., Savage, J.E., Sestan, N., Shin, Y., Stockmeier, C.A., Walters, J.T.R, Yao, S., Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Eating Disorders Working Group of the Psychiatric Genomics Consortium, Crawford, G.E., Jin, F., Hu, M., Li, Y., Sullivan, P.F., Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to 2 interpret genetic results for psychiatric disorders and cognitive traits *(Submitted).*

26. Jansen, A.G., Dieleman, G.C., **Jansen, P.R**., Verhulst, F.C., Posthuma, D., Polderman, T.J.C., Psychiatric polygenic risk scores as predictor for attention deficit/hyperactivity disorder and autism spectrum disorder in a clinical child and adolescent sample *(Submitted).*

27. De Mol, C.L., **Jansen, P.R.**, Muetzel, R.L., Knol, M.J., Adams, H.H.H., Jaddoe, V.W.V., Hintzen, R.Q., White, T., Neuteboom, R.F., Polygenic risk for multiple sclerosis and multi-modal brain imaging among 8-12 year old children in the general population *(Submitted).*

28. **Jansen*, P.R.**, Nagel*, M., Watanabe, K., Wei, Y., Savage, J.E., de Leeuw, C.A., Polderman, T.J.C., van den Heuvel, M.P., van der Sluis, S., Posthuma, D., Cross-trait analysis of brain volume and intelligence identifies shared genomic loci and genes *(In preparation).*

29. **Jansen, P.R.**, Dekkers, I.A., de Lange, S.C., Watanabe, K., Polderman, T.J.C, van den Heuvel, M.P., Posthuma, D., Novel loci and genes for white matter microstructure (N=15,450) identified through genome-wide analysis *(In preparation).*

* these authors contributed equally

Non-peer-reviewed papers

30. **Jansen, P.R.**, Nagel, M., Savage, J.E., van der Sluis, S., Posthuma, D., Breaking the genetic code of human intelligence and neuroticism. *Amsterdam Science* (2018).

31. **Jansen, P.R.,** Savage, J.E., Nagel, M., van der Sluis, S., Posthuma, D., Over de genetische achtergrond van intelligentie en neuroticisme: resultaten uit genoomwijde associatie studies. *Analyse (Nederlandse Vereniging voor Biomedisch Laboratorium-medewerkers)* (2018).

PhD Portfolio

Name: Philip R. Jansen
Research School: Netherlands Institute of Health (NIHES) / ONWAR Neuroscience graduate school
PhD period: August 2014 – August 2018
Departments: Epidemiology/Child and Adolescent Psychiatry (Erasmus MC), Complex Trait Genetics (Vrije Universiteit),
Promotors: Prof. dr. Frank Verhulst, Prof. dr. Danielle Posthuma,
Copromotors: dr. Tinca Polderman, dr. Tonya White

1. PhD training	**Year**	**ECTS**
Master of Science degree in Clinical Research, NIHES graduate school courses (2010-2012)		
Principles of Research in Medicine and Epidemiology	2010	**0.7**
Introduction to Data-analysis	2010	**1.0**
Regression Analysis	2010	**1.9**
Methods of Clinical Research	2010	**0.7**
Clinical Trials	2010	**0.7**
Topics in Meta-analysis	2010	**0.7**
Pharmaco-epidemiology	2011	**0.7**
Health Economics	2011	**0.7**
Survival Analysis	2011	**1.9**
Cohort Studies	2011	**0.7**
Primary and Secondary Prevention Research	2011	**0.7**
Introduction to Decision-making in Medicine	2011	**0.7**
Study Design	2011	**4.3**
Introduction to Clinical Research	2011	**0.9**
Advanced Topics in Decision-making in Medicine	2011	**1.9**
Intervention Research and Clinical Trials	2011	**0.9**
Diagnostic Research	2012	**1.1**
Advanced Topics in Clinical Trials	2012	**1.9**
Advanced Analysis of Prognosis Studies	2012	**0.9**
Prognosis Research	2012	**0.2**
Principles of Epidemiologic Data-analysis	2012	**0.7**
Working with SPSS for Windows	2012	**0.15**
Scientific Writing in English for Publication	2012	**2.0**
ONWAR graduate school courses (2014 – 2018)		
Introduction to ONWAR	2014	**0.9**
Neuropsychopharmacology	2015	**1.4**
Functional neuroanatomy	2015	**1.4**
General courses		
Scientific integrity	2018	**0.3**

Specific courses		
MRI safety course	2015	**0.3**
International Workshop of Statistical Genetics (Boulder, CO)	2015	**1.8**
Linux for Scientists (NIHES)	2015	**0.6**
R programming (MOLMED)	2015	**1.8**
FSL course 2016 (Giardini, Italy)	2016	**1.8**
Neuroimaging Training Program (NITP) UCLA (LA, CA)	2016	**3.0**
FreeSurfer MRI course (Copenhagen, Denmark)	2016	**1.8**
Brain Connectomics Summer School (UMCU)	2017	**1.8**
Online courses		
R programming (Coursera)	2017	-
Python programming (Coursera)	2017	-
Introduction to Python (SoloLearn)	2017	-
Introduction to Java (SoloLearn)	2017	-
Introduction to C++ (SoloLearn)	2017	-
Scientific presentations, meetings		
Complex Trait Genetics research meetings (oral presentations)	2014-2018	**1.2**
Generation R research meetings (oral presentations)	2014-2018	**1.2**
Sophia Research Day 2015 (oral presentation)	2015	**0.3**
ONWAR Neuroscience meeting (scientific poster)	2016	**0.3**
ONWAR Neuroscience meeting (oral presentation)	2017	**0.3**
Amsterdam Neuroscience meeting 2017 (scientific poster)	2017	**0.3**
VU Science exchange day 2017 (scientific poster)	2017	**0.3**
NIN symposium (oral presentation)	2018	**0.3**
AMC Clinical Genetics lunch lecture (oral presentation)	2018	**0.3**
PCG web seminar (oral presentation)	2018	**0.3**
Invited lunch lecture ISglobal Barcelona (oral presentation)	2018	**0.3**
Amsterdam Neuroscience meeting 2018 (scientific poster)	2017	**0.3**
WEON epidemiology 2018 meeting, de Bilt, Netherlands (scientific poster)	2018	**0.3**
Science day Clinical Genetics VUmc 2018 (scientific poster)	2018	**0.6**
International conferences, meetings		
Human Brain Mapping 2016, Geneva, Switzerland (scientific poster)	2016	**1.2**
Invited lecture COST international working group on malformations of cortical development, Genoa, Italy (oral presentation)	2017	**1.2**
RSNA meeting 2017, Chicago, USA	2017	**2.4**
JOINT Clinical genetics meeting 2018, Utrecht, Netherlands (oral presentation)	2018	**0.6**
World Congress of Psychiatric Genetics, Glasgow, Scotland (oral presentation)	2018	**1.2**

2. Teaching activities	**Year**	**ECTS**
Supervising master's thesis Wouter Deijl (Erasmus MC): *Vitamin D and IQ in the Generation R cohort*	2015	**3.0**
Supervising master's thesis Janine Lavooy (Leiden University): *Polygenic risk for depression and cognition*	2015	**3.0**
Supervising master's thesis Jantien Noordman (Erasmus MC): *Preterm birth and diffusion tensor imaging*	2015	**3.0**
Supervising master's thesis Amin El-Achari (Erasmus MC): *Polygenic risk for schizophrenia and fMRI*	2016	**3.0**
Supervising bachelor's thesis Thirza Dado (Amsterdam University): *Polygenic risk for neuroticism and diffusion tensor imaging*	2016	**3.0**
Supervising honours program Amber Boer (Vrije Universiteit): *Polygenic risk for antisocial behavior and diffusion tensor imaging*	2017	**3.0**
Supervising bachelor's thesis Bo Shan-Go (Vrije Universiteit): *Genome-wide analysis of white matter microstructure on diffusion tensor imaging*	2018	**3.0**

Teaching:		
Child Psychiatry lectures medical students Erasmus MC	2015-2016	**1.0**
Scientific review supervision medical students Erasmus MC	2016	**0.6**
Teaching assistant Genetics lectures VU bachelor students	2016	**0.6**
Honours program: psychiatric genetics VU: GWAS lectures	2017-2018	**0.6**
Minor: psychiatric omics course VU: GWAS lectures	2018	**0.3**
3. Other activities	**Year**	**ECTS**
Incidental findings coordinator Generation R	2014-2016	**6.0**
Research center schedule coordinator Generation R	2015-2016	**3.0**
Peer review: Cerebral Cortex, American Journal of Psychiatry, Erasmus Journal of Medicine	2016-2018	**1.0**
Media appearances:		
Sports genetics, commentary (RTL Late Night, television)	2018	-
Sports genetics, commentary (De Telegraaf, newspaper)	2019	-
Insomnia genetics, commentary (538 Radio, radio)	2019	-

1 ECTS (European Credit Transfer System) is equal to a workload of 28 hours.

Authors and Affiliations

Department of Complex Trait Genetics, Center for Neurogenomics and Cognitve Research, Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, the Netherlands
Philip R. Jansen, Mats Nagel, Kyoko Watanabe, Jeanne E. Savage, Sven Stringer, Anke R. Hammerschlag, Yong-bin Wei, Martijn P. van den Heuvel, Sophie van der Sluis. Tinca J.C. Polderman, Danielle Posthuma

Department of Clinical Genetics, VU Medical Center, Amsterdam University Medical Center, Amsterdam, the Netherlands
Mats Nagel, Sophie van der Sluis, Danielle Posthuma

Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, the Netherlands
August B. Smit

Department of Sleep and Cognition, Netherlands Institute for Neuroscience (an institute of the Royal Netherlands Academy of Arts and Sciences), Amsterdam, the Netherlands
Eus J.W. Van Someren

Departments of Psychiatry and Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam University Medical Center, Amsterdam, the Netherlands
Eus J.W. Van Someren

Department of Child and Adolescent Psychiatry/Psychology, Erasmus MC-Sophia, Rotterdam, the Netherlands
Philip R. Jansen, Ryan L. Muetzel, Desana Kocevska, Koen Bolhuis, Toyah A. Jansen, Rosa H. Mulder, Laura M.E. Blanken, Jan van der Ende, Frank C. Verhulst, Henning Tiemeier, Tonya White

The Generation R Study Group, Erasmus MC, Rotterdam, the Netherlands
Philip R. Jansen, Ryan L. Muetzel, Desana Kocevska, Koen Bolhuis, Toyah A. Jansen, Rosa H. Mulder, Laura M.E. Blanken,Vincent W.V. Jaddoe, Frank C. Verhulst, Henning Tiemeier, Tonya White,

Department of Radiology, Erasmus MC, Rotterdam, the Netherlands
Philip R. Jansen, Ryan Muetzel, Aad van der Lugt, Marjolein Dremmen, Aaike van den Berg, Tonya White

Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands
Henning Tiemeier, Vincent W.V. Jaddoe

Department of Psychiatry, Erasmus MC, Rotterdam, the Netherlands
Henning Tiemeier

Department of Pediatrics, Erasmus MC-Sophia, Rotterdam, the Netherlands
Vincent W.V. Jaddoe

Department of Child Neurology, Erasmus MC-Sophia, Rotterdam, the Netherlands
Marie-Claire Y. de Wit, Rinze F. Neuteboom

Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands
Ilona A. Dekkers

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
Julien Bryois

Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
Nathan Skene, Ana B. Muñoz-Manchado, Jens Hjerling-Leffler, Sten Linnarson

UCL Institute of Neurology, Queen Square, London, UK
Nathan G. Skene, Patrick F. Sullivan

Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
Patrick F. Sullivan

Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA
Patrick F. Sullivan

23andMe, Inc., Mountain View, CA, USA
Joyce Y. Tung, David A. Hinds, Vladimir Vacic, Xin Wang

About the Author

Philip Rombout Jansen (1989, Utrecht) grew up in Houten (nearby Utrecht), the Netherlands. After completing VWO at the College de Heemlanden in Houten, he started medical school at the Erasmus University Medical Center (Erasmus MC) in Rotterdam in 2007, following his older brother Victor who had started medical school one year earlier at the same university. At the start of medical school, Philip already became interested in scientific research and medical imaging. He successfully completed a research master in Clinical Research from the Netherlands Institute of Health Sciences (NIHES) in 2010 and was registered as an epidemiologist (registration A). During his master's, he investigated the use of diffusion-weighted magnetic resonance imaging (MRI) in head and neck cancer, under supervision of prof. dr. Aad van der Lugt. After finishing medical school, he started a joint PhD project in the Generation R Study at the department of Epidemiology of the Erasmus MC (advisor: prof. dr. Frank Verhulst), in collaboration with the department of Complex Trait Genetics of the VU University in Amsterdam (advisor: prof. dr. Danielle Posthuma). His PhD project focused on investigating the complex genetic architecture of psychiatric disorders and behavioral phenotypes, and the use of polygenic risk scores to study associations between genetic risk for psychiatric disorders and MRI brain imaging. From September 2018 onwards, Philip continued his clinical career and started to work as a resident at the department of Clinical Genetics at the VU Medical Center (Amsterdam UMC), where he initiated his training to become a clinical geneticist in February 2019. Philip will continue to combine clinical training in the VU Medical Center with scientific research in the Complex Trait Genetics group at the VU University. By integrating knowledge from statistical genetics with clinical medicine and by translating scientific results to patient care, he hopes that his future research may bring genetic science closer to the clinic, improving clinical care and the lives of patients.


GGCTACGACTAGCGATCTAGCGATG
CCGATGCTGATCGCTAGATCGCTAC
OH
H
N
(R)
HO
OH
NH2
HO
N
H
HO
NH2
HO
N
N
H