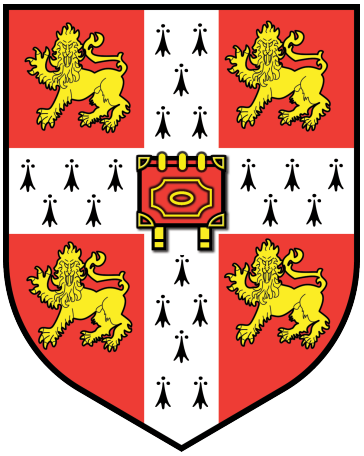
Common genetic variants contribute to risk of rare severe neurodevelopmental disorders



Mari Elisa Katariina Niemi

Queens' College

University of Cambridge

September 2018

Dissertation submitted for the degree of Doctor of Philosophy

Common genetic variants contribute to risk of rare severe neurodevelopmental disorders

Mari Elisa Katariina Niemi, Queens' College, University of Cambridge

Most known genetic causes of severe childhood developmental disorders are rare, deleterious, protein-coding changes that cause Mendelian disorders. Children with these disorders typically show early-onset impairment in growth, learning and adaptive behaviours. Linkage and whole exome sequencing studies on these patients have previously focused on identifying diagnostic rare variants that are solely responsible for the patient's phenotype. In this thesis, I investigate whether common, inherited genetic variation also plays a modifying role in severe, presumably Mendelian neurodevelopmental disorders. In addition, I study the effects of common variants on the cognitive functioning of healthy individuals, who carry rare deleterious variants in genes that are intolerant to such variants in the general population.

To test whether common variants contribute to neurodevelopmental disorders that are expected to be almost entirely monogenic, I conduct a genome-wide association study (GWAS) in nearly 7,000 patients from the Deciphering Developmental Disorders (DDD) Study and ancestry-matched controls. I show that common genetic variants explain almost 8% of variation in risk for these severe disorders. I also find genetic overlap between our study and GWAS for other cognitive and neuropsychiatric traits. This suggests that common variants individually have a small effect on brain development and functioning, influencing both risk for common diseases in the population and risk for severe disorders that affect only a small number of individuals. This polygenic burden in the DDD is also not confined to only patients who do not have diagnostic rare variants. Altogether, these results may have important implications for understanding variable clinical presentation of neurodevelopmental disorders and searching for secondary genetic modifiers.

Finally, I assess the interplay between common and rare variants on the cognitive functioning of seemingly healthy individuals. Using data from the INTERVAL Study, I test whether common variants are protective of the deleterious rare variants in these individuals. Whilst these analyses are potentially currently underpowered,

with additional samples in the future, we may be able to shed more light on expressivity and penetrance of deleterious variants in the general population.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the contributions section of each chapter and/or specified in the text. It is not being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words.

Mari E. K. Niemi
September 2018

Acknowledgements

I would like to start by thanking my supervisor Jeffrey Barrett. This thanks extends all the way back to when you gave me the opportunity to complete my Masters project as part of your team. That year was a turning point in my scientific journey in many ways, and I wouldn't have gone down this path had it not been for the hugely exciting science and working environment that I got to experience. It has been a privilege to work with you, and to learn from and be part of that environment at the Sanger. Most of all, thank you for believing in me. I am hugely grateful to Hilary Martin, who tirelessly devoted time to helping me with the challenges that I faced over the last couple of years, and for supervising me during the final months. You have been a huge inspiration, and I look forward to continuing to learn from you. I would also like to thank Matthew Hurles for his feedback and discussions over the years, and for his supervision during the final year of my PhD. I would like to thank my Cambridge supervisor Lucy Raymond, and Gosia Trynka who together with Matt formed my thesis committee, and provided useful insights. I would like to thank the Wellcome Trust who funded this PhD, and Queens' College Cambridge who supported my work. I am also grateful to the study participants and families of patients who donated the samples used in this work.

A special thanks to Dan, Alex and Katie for helping me out with everything computer, and also to the rest of the Barrett, Anderson and Hurles teams for the useful conversations, feedback, bioinformatics help (and interesting lunchtime discussions). To Elena, especially for the support during all the writing times we had over these last four years. To my fellow Sanger PhD friends, with whom I've shared so many experiences, including the all-important work club. To Ala, for being there from day one. To Gaby, for always being there even if not in the same country. Katariinalle, kiitos kaikesta. To Saana, Rikke and Becky thank you for being there for me and for all the fun times we had. To friends from Queens', from rowing, from Panton, and old friends from times before Cambridge, thank you for being part of this experience. To Maggie & Erkki, thank you for your support - Laguna is definitely the best place to "work from home".

Suurimmat kiitokset perheelleni: Minnalle, Maijalle, Johannalle & Jukalle. Kiitos tuestanne naiden mahtavien, mutta myos ajoittain erittain vaikeiden vuosien aikana. Olette rakkaimmat.

Contents

Chapter 1

Introduction

In human medical genetics there exists a dichotomy between rare diseases, thought to be caused by rare variants with large deleterious effects, and common diseases explained by common genetic variants that individually have a small effect on the phenotype (Bamshad et al., 2011), combined with environmental factors. With improved technology and understanding of genetic architecture of diseases, more examples have emerged of common, more complex, phenotypes having a contribution from both common variants and rare deleterious ones. However, in the field of rare genetic diseases the contribution of common variants is less well studied.

1.1 Genetics of Mendelian traits

1.1.1 Mendelian inheritance

Mendelian traits and diseases follow clear patterns of inheritance within family pedigrees, and affect a single gene or locus. The term Mendelian trait comes from Gregor Mendel, whose studies in the 1800s on inheritance of traits in peas led to the formulation of Mendel's laws. These are the law of segregation, where Mendel hypothesized that alleles at a locus separate from each other randomly during

gamete production, and the law of independent assortment which describes how a pair of alleles separate independently of other pairs of alleles. The five Mendelian inheritance patterns of inheritance are autosomal recessive, autosomal dominant, X-linked recessive, X-linked dominant, and Y-linked patterns (Strachan and Read, 2011). Although most Mendelian diseases follow a clear pattern of inheritance, some can be inherited in more than one way. For example, deafness associated with gene *CX26* can be caused by dominant and recessive variants in the gene (Kemperman et al., 2002).

1.1.2 Genetic variants in Mendelian traits

We now know that Mendelian diseases are typically single gene diseases, in which deleterious genotypes at one locus are enough to cause the disease phenotype. Some genes are known to be associated with multiple Mendelian diseases (Zhu et al., 2014; Singh et al., 2016), and sometimes the same severe phenotype can be caused by variants in multiple different genes, e.g. there are over 120 genes in which mutations have been found to cause deafness (Nance, 2003). It is thought that there are many thousands of human disorders that are Mendelian (Antonarakis and Beckmann, 2006), but a causal gene for all these has not yet been described (Bamshad et al., 2011; Botstein and Risch, 2003). Mendelian diseases are typically rare on the population level, and collectively affect a small number of individuals. However, the burden on Mendelian diseases on health services is substantial, and account for an estimated 10% of paediatric hospital admittances and 20% of infant deaths in North America (Karczewski and Snyder, 2018).

In order for a variant to have a large deleterious effect enough to cause disease , it typically has to reside within the protein-coding region of the genome (exons of genes) (MacArthur and Tyler-Smith, 2010; Garcia-Alonso et al., 2014). Mendelian disease variants therefore often cause loss-of-function effects, by truncating the protein or disturbing its functional domains. This happens by the introduction of an early stop codon either directly through nonsense mutations, disruption of the reading frame (insertion or deletions), or the alteration the splice sites. Truncation of the peptide sequence often results in depletion or altered function of the protein.

In addition, damaging missense variants can effectively result in loss-of-function effects or lead to a dominant-negative or gain of function effect. Genes associated with Mendelian disorders are typically intolerant of loss-of-function mutations (Lek et al., 2016), and deleterious variants in these genes are under negative selection (Bustamante et al., 2005). Therefore Mendelian disease variants tend to be very rare in the general population.

1.1.3 Penetrance and expressivity of Mendelian traits

Some Mendelian diseases do not always show the typical inheritance patterns. Exceptions to these patterns may be introduced through incomplete (or reduced) penetrance, particularly in the case of dominant traits (Strachan and Read, 2011). Penetrance describes the probability that the phenotype is observed given the individual has the associated genotype. When a proportion of individuals with the disease genotype do not show signs of the disease, it is said to be incompletely penetrant. An example of such a disease is phenylketonuria, which is caused by loss-of-function mutations in the *PAH* gene encoding for an enzyme involved in breakdown of phenylalanine. Without intervention, the disease causes severe intellectual disability. However, if phenylalanine is restricted from birth, the child will grow relatively healthy (Cooper et al., 2013), and thus the disease only manifests depending on phenylalanine intake. Diseases such as Huntington's disease are fully penetrant; although there are differences in the age of onset between patients, essentially everyone who has more than 40 repeats of the CAG-triplet repeats in their *HTT* gene will develop disease (Myers, 2004). As an example of variable penetrance, the overall 57% of women with *BRCA1* variants develop breast cancer by the age of 70, and 40% develop ovarian cancer (Chen and Parmigiani, 2007). However, it is also known that there are differences in penetrance between different variants within the same gene: for example, *BRCA1* the mutation 185delAG in exon 2 of the gene has a much lower penentrance which is also age-dependent than does one of the more penetrant variants, the duplication of exon 13. The median age of breast cancer affliction for 185delAG carriers was 55 years, whereas for exon 13 duplication carriers this was 41 years (Al-Mulla et al., 2009).

Another characteristic of Mendelian phenotypes is the expressivity of the trait. Variable expression of a disease refers to individuals with the same genotype showing different symptoms or different severity of these, even within families (Strachan and Read, 2011). In some diseases, expressivity has been shown to be explained at least partly due to mutations in different functional regions of the gene (Zhu et al., 2014). However, for some Mendelian diseases, so-called modifier genes have been found to contribute to differential expressivity. Examples of this include cystic fibrosis, where recessive variants in the gene *CFTR* cause cystic fibrosis, a disease that obstructs the lungs and affects organs. Variants in several genes have been reported to possibly alter the disease course and different organ system symptoms associated with the disease (Cutting, 2010).

1.1.4 Linkage Studies for gene discovery

Early gene discovery studies focused on finding causal loci and genes for Mendelian diseases and traits using genome-wide linkage (Botstein et al., 1980). These studies, before the availability of human reference genomes, assessed the co-segregation of the trait and known genetic markers, along family pedigrees (Ardlie et al., 2002). Linkage is the physical relation between loci (Strachan and Read, 2011). When two loci that are in physical proximity of each other, they tend to be transmitted together and are thus termed 'linked'. Recombination during meiosis between the loci is more likely if the loci are further apart, or less tightly linked (Strachan and Read, 2011). Markers that were linked to the causal genetic locus would be shared among affected family members, and not observed in unaffected family members. Linkage studies assessing Mendelian diseases used statistical approaches that assumed a specific disease model of inheritance, and in literature these studies are called parametric linkage studies (Strachan and Read, 2011).

Because finding the causal loci for a disease using linkage studies did not require base pair resolution of chromosomal sequences, many Mendelian disease loci and genes were identified using linkage in the 90s when sequencing was still very expensive. The markers initially used were microsatellites, small repeat sequences of DNA, as these were highly polymorphic sites in the genome (Strachan and Read, 2011).

Using a few hundred markers spread across the genome, the transmission of tagged regions could be traced at a megabase resolution. Later on, studies switched more to using single nucleotide polymorphisms (SNPs). An example of one of the early successes was a study by Hästbacka et al. (1992), where the authors traced the causal dominant-acting variant for diastrophic dysplasia, a cartilage and bone disorder, to ∼60kb region from the gene *CSFR1* gene.

1.1.5 Sequencing for diagnosis and discovery

With an increasing number of genes and chromosomal loci associated with Mendelian diseases, sequencing relatively small lists of candidate genes for a given disease became more popular in clinical genetics for Mendelian disorder diagnosis in clinical genetics (Zelst-Stams et al., 2014). In these studies a list of potential genes were drawn based on what was thought could be the underlying biology behind the disease. This best guess approach was also used for genetic discovery for disorders that resembled those for which a causal gene had already been found (Bamshad et al., 2011). However, sequencing technologies at the time were expensive (Petersen et al., 2017), and often this approach did not in fact identify the causal genes. Characteristics such as disease penetrance and expressivity, and the small sample sizes of family studies also limited the power for gene discovery using these methods (Bamshad et al., 2011).

With the emergence of next-generation sequencing technologies around a decade ago, larger scale sequencing studies for Mendelian traits have since become possible for both diagnosis and new gene discovery (Bamshad et al., 2011). Now, sequencing of all protein-coding regions in the genome (the exome) has become the preferred framework for targeted sequencing studies in Mendelian diseases (Petersen et al., 2017). It is also likely that exome sequencing will be incorporated more into clinical practice in the future even for early stages of clinical diagnostic investigations (Zelst-Stams et al., 2014; Wright et al., 2018c). Whole exome sequencing (WES) uses targeted capture of protein-coding regions, which amounts to ∼2% of the genome (Sazonovs and Barrett, 2018). Since most variants contributing to Mendelian diseases are located within the exome, this approach is justified for searching

for genes associated with these rare diseases. Whole genome sequencing (WGS) can also be used for Mendelian disease analyses, but the relative cost of WES is still around a third of the cost of a WGS genome (Sazonovs and Barrett, 2018). The downside of WES compared to WGS is that non-coding regions including regulatory elements are not captured, the coverage is more variable, and it is harder to accurately call structural variants.

Exome sequencing is particularly useful in finding diagnoses for patients with Mendelian disease phenotypes but for whom the causal variant has not been identified through other means (Bamshad et al., 2011). WES has been particularly useful for sequencing of trios, enabling the identification of *de novo* mutations in genes that cause such severe phenotypes that these result in severe reduction in reproductive capacity (meaning the patient is sterile, that carriers do not reach reproductive age, or that they do not produce progeny due to the severity of the clinical symptoms). WES of patients with previously undiagnosed neurodevelopmental or neurological Mendelian diseases has resulted in a genetic diagnosis of up to ∼40% in some cohorts (Yang et al., 2013; Gilissen et al., 2014; Deciphering Developmental Disorders Study, 2017). In addition, trio exome sequencing in a large number of patients has proved useful for discovery of many new genes associated with these Mendelian disorders (Deciphering Developmental Disorders Study, 2017).

However, utilising WES and WGS comes with some drawbacks. Generating and analysing sequence data is still computationally expensive, and requires particular expertise (Sazonovs and Barrett, 2018). It may also be difficult to identify clinically relevant variants given the large number of variants in the exome, particularly when data is not available for parents. In some studies, candidate variants have been fed back to clinicians, who have in turn assessed whether the variants identified are likely to be diagnostic to the patient (Beaulieu et al., 2014), but this approach requires careful study planning and involvement of clinical geneticists.

1.2 Genetics of complex traits

1.2.1 Non-Mendelian inheritance

As early as the late 1800s, scientists already noted that not all human traits seemed to be inherited according to Mendelian patterns, and instead were passed down in a blended manner from parents to offspring (Visscher and Bruce Walsh, 2017). These diseases or traits appeared to aggregate in families without following a distinct recessive or dominant pattern. Additionally, non-categorical (continuous) traits such as human height were also correlated between family members. Debate over the inheritance of these complex or quantitative traits was eventually resolved by Ronald Fisher, who in 1918 published his seminal work on the genetics of complex traits (Fisher, 1918). This work included introducing the concept of variance, and categorisation of genetic effects into additive and dominance effects (more in section 1.2.3), without the need for prior knowledge of the genes underlying the trait (Visscher and Bruce Walsh, 2017). Fisher proposed that some traits were multifactorial, meaning many factors contribute to the trait, and that random sampling of alleles in a population results in a continuous trait when multiple alleles contributed to it. He also proposed that the individual contributions of each allele becomes smaller when more alleles contribute to the trait (Fisher, 1918). These ideas formed the basis of the study of complex traits and diseases.

Complex diseases are often relatively common in the population (Becker, 2004), and therefore are regularly referred to as common diseases. Complex diseases have a contribution from both environmental and genetic factors, but typically the genetic component is a combination of multiple genetic variants, hundreds to potentially tens of thousands (Lee et al., 2018). Some rare variants can cause the same or a similar phenotype as the complex trait, but this constitutes a small fraction (typically less than 10%) of the disease cases (Scheuner et al., 2004). It is also thought that late-onset diseases are more likely to be polygenic instead of Mendelian (Wright et al., 2003). Complex diseases also collectively affect a large number of individuals, e.g. prevalence for type 2 diabetes is 8% (Morris et al., 2012) and 15% for major depressive disorder (Major Depressive Disorder Working

Group of the Psychiatric GWAS Consortium et al., 2013), which results in major burden on health services.

1.2.2 Genetic studies in complex traits

Early attempts with linkage studies

Attempts to discover genes associated with complex diseases initially used many of the same methods as studies for Mendelian diseases. This was because complex diseases also cluster in families. However, without a known model of inheritance, the statistical methods used had to be non-parametric (Strachan and Read, 2011). Infrequent successes of linkage family-based linkage studies in mapping complex disease loci, include the identification a susceptibility locus on chromosome 16 for inflammatory bowel disease (Hugot et al., 1996). The locus, which we now know contains the gene *NOD2*, has since remained one of the strongest known effect loci for Crohn's disease, a subtype of inflammatory bowel disease. Eventually though not many complex disease loci were mapped using linkage studies, because the genetic variants contributing to complex diseases did not confer high enough susceptibility in a population for it to be detected using family data, and at the resolution that genetic markers at the time provided (Strachan and Read, 2011).

Improving the resolution for association

Over the years, it became apparent that mapping loci for complex traits was difficult, likely due to the small effect sizes of the loci involved. The International Hapmap Project was launched in 2002 to characterise the patterns of genetic variation across different populations (International HapMap Consortium, 2003). The HapMap project showed that the finer scale structure of human chromosome haplotypes (blocks sequences that were inherited together) was more complex than previously thought (Strachan and Read, 2011). The HapMap project facilitated genetic association studies, by allowing researchers to better design DNA chips with a limited number of markers that optimally tagged (i.e. were correlated with) most

of the other common variation in the genome (mainly in European populations). As it turned out, approximately 500,000 SNPs were sufficiently informative to tag the remaining common variants (MAF>5%) in a European ancestry genome, due to linkage disequilibrium (Consortium, 2007).

Linkage disequilibrium is a statistical association between two alleles that are genetically linked, and therefore observed together on the same haplotype more often than expected (Strachan and Read, 2011). A variant can tag the surrounding variants within its haplotype block, as they are likely to be observed together given the other surrounding markers. This also means the surrounding SNPs can be statistically inferred (imputed) and subsequently tested for association (Marchini and Howie, 2010). Linkage disequilibrium patterns also differ between populations because of recombination events that happened in previous generations. Even though for some diseases e.g. inflammatory bowel disease it has been shown that many risk loci are shared between populations (Liu et al., 2015), the variants that tag underlying causal variants in different population may not be the same. Because of differences in LD structure, genetic association studies need to be carried out within a fairly homogeneous population to avoid spurious associations (Anderson et al., 2010) arising from the inclusion of individuals with differential haplotype structure.

Genome-wide association studies

With the new information on haplotype structure, complex disease association testing moved on to testing hundreds of thousands of SNPs at a time. Data for these were generated using DNA chips, which were a relatively cheap (Sazonovs and Barrett, 2018), enabling studies to recruit increasingly larger cohorts - and consequently gained more statistical power for finding susceptibility loci with smaller effects. Currently, genotyping one sample costs ∼$20 on the Illumina Global Screening chip, and sample sizes for association analyses have moved from few hundred to tens of thousands or up to a million (Lee et al., 2018). The reduced cost per sample and the improved resolution at which disease associations could be detected using chips caused the interest in genome-wide association studies

(GWAS) to grow rapidly in the mid 2000's. GWAS use statistical methods to compare allele frequencies between cases and controls (e.g. logistic regression) (Clarke et al., 2011) and to find variants that are associated with continuous traits (e.g. linear regression). GWAS was the name given to these large association scans that typically used chip data. At the time, it was understood that more power could be gained for complex disease association analysis by genotyping unrelated individuals to use as cases and controls (or for continuous traits) (Teng and Risch, 1999). It was also more straightforward to recruit these individuals rather than genotype whole families where the trait clustered.

Modern GWAS typically test several million SNPs, of which the majority will have been imputed in order to boost genome-wide coverage. This large number of variants to test introduces the issue of false positives. Each test on a single variant is treated as an independent test, and therefore setting a threshold of significance at the typical P-value <0.05 would result in potentially millions of false associations. To account for the number of tests performed in a genome-wide scan, multiple testing correction (e.g. Bonferroni correction) is applied to GWAS data. Typically, the P-value cut off for a GWAS in Europeans is set to $< 5 \times 10^{-8}$ for 1M SNPs (Risch and Merikangas, 1996; Pe'er et al., 2008). The problem with this however is that many variants that probably do confer risk, do not pass this genome-wide significance threshold, and studies have to recruit large numbers of samples to gain sufficient power to detect the association. For example, autism spectrum disorder (ASD) has for long been suspected to have a polygenic component to disease aetiology, but researchers reported the first significant GWAS loci only after reaching sample sizes of 18,000 cases and 28,000 controls (Grove et al., 2017).

It is important to note that association does not necessarily mean causation, and the most highly associated variant is not necessarily the causal one. Finding the causal variant requires fine-mapping of the region around the associated variants (Schaid et al., 2018) and functional follow up using cell lines, animal models or other methods. Most GWAS hits are in non-coding regions which makes it difficult to figure out the causal variant and the biological pathways affected (Zhu et al., 2017). One of the main aims of GWAS, as with family-based studies, is to find potential drug targets for patients suffering from disease.

One the biggest caveats of GWAS is that these only measure common variants. It has been argued that rare variants may contribute substantially to many complex traits (Lee et al., 2011), and there is now evidence for this for several traits and diseases (Luo et al., 2016; Ganna et al., 2016). However, the challenge in studying this hypothesis through GWAS is that rare variants are not well tagged by surrounding common variants, so need to be ascertained directly through sequencing. Power once again becomes an issue when investigating rare variant associations (Sazonovs and Barrett, 2018). Some hope that by finding rare variant associations at loci that likely lead to larger effects on the phenotype can point more quickly towards causal genes. If we were able to find these, this could help with developing drugs more quickly.

1.2.3 Heritability of complex traits

Broad sense heritability

Both Mendelian and complex diseases and traits are heritable, which means that genetic variants contribute to phenotypic variance between individuals (Visscher et al., 2008). The term broad-sense heritability (H^2) is used to describe the proportion of variation in the phenotype that is attributable to all genetic variation (Visscher et al., 2008). Genetic effects contributing to the phenotypic variance can originate from additive genetic effects (combined genetic effects are equal to the sum of individual allele effects), dominance effects (where interactions between alleles at the same locus affect the outcome), and epistatic effects (where interactions between alleles at different loci affect each other) (Strachan and Read, 2011). Fully penetrant Mendelian diseases are in principle fully heritable with a H^2 of 1 (Visscher et al., 2008). Complex diseases will have a H^2 less than 1, as part of the variation in the phenotype typically comes from non-genetic effects for these diseases. Importantly, heritability for a given trait can differ between populations and timepoints. As an example of this, in European populations, a trait such reading ability would have been far less heritable a few centuries ago than it is now. This is because back then, one's schooling depended mainly on socio-economic status (although we note that socioeconomic status has now been shown to be at

least partly heritable today (Trzaskowski et al., 2014), whereas with the modern standardised schooling system the environment is less variable (Strachan and Read, 2011). Therefore the background (including genetic) characteristics enabling one to learn to read, such as cognitive ability, will now contribute proportionally more to the trait. The genetic architecture underlying heritability can also vary greatly. In the case of fully penetrant Mendelian diseases, a single deleterious variant with large effect size will explain all of the H^2 (Visscher et al., 2008). But for complex diseases, the number of genetic variants explaining H^2 can be thousands to tens of thousands (or more) spread across the genome, each with individually small effect sizes (Lee et al., 2018).

It is important to quantify heritability in order to understand how much genetics contributes to the trait. For complex diseases, H^2 has traditionally been estimated through family studies. These studies used data collected especially on twins. Twin pairs typically share most of their environment, resulting in minimal bias from different environments, and the average shared proportion of alleles is known. By comparing the phenotypic concordance between monozygotic twins (who share all their alleles) to the concordance between dizygotic twins (who share on average half their genetic content) gives an estimate of the overall genetic contribution to the trait (H^2).

Narrow-sense heritability

The additive genetic component of broad sense heritability is termed narrow-sense heritability or h^2. It represents the total proportion of variance in the trait that can be explained by summing the additive effects of all variants. Narrow-sense heritability is of interest in population genetics, because it is relatively straightforward to to measure within a population study design. The definition of additive effects excludes factors such as interactions between variants, and it is often thought to be a large contributor to the overall trait heritability on a population level (Visscher et al., 2008). The additive model implies a linear relationship between how much of their genome a pair of individuals share (i.e. how related they are to each other) and how closely they resemble each for the trait in question.

Narrow sense heritability could also be measured using family studies. The simple twin studies advanced into more complicated family designs, in which the inclusion of multiple relative types made it possible to tease apart additive from dominance variance, and from the effects of shared environment and unique environment. Dominance variance does not contribute much to heritability of complex traits (Visscher et al., 2008).

SNP heritability

Family studies in complex disease had for decades shown that many diseases and traits had substantial genetic contributions. Therefore when SNP genotyping platforms became cheaper and more widely used in the early 2000s, it was expected that the GWAS approach would finally find the additive genetic effects and pinpoint causative variants to human traits and diseases. However, it was quickly realised that the genome-wide significant SNPs discovered by GWAS failed to explain much of the expected narrow sense heritability h^2 that had been estimated through twin and family studies (Lee et al., 2011). The heritability captured by GWAS is termed SNP heritability or sometimes chip heritability and represents the additive genetic effects tagged by common SNPs (Yang et al., 2010).

The remaining gap between family-study based estimates of H^2 ($\sim h^2$) and the proportion of variance explained by population studies was termed missing heritability (Maher, 2008). At the time, possible sources of missing heritability were thought to be the thousands of common variants with very small effect sizes hard to detect using available study sample sizes, or in rarer variants with intermediate effects that were not well tagged by the common SNPs (Eichler et al., 2010). Some have also argued that twin and family studies have over-estimated the true narrow-sense heritability due to flaws in their assumptions. The current predominant view is that most of the missing heritability lies in thousands of common variants, which we have increasingly better power to detect (Yang et al., 2010; Lee et al., 2018).

Reporting heritability for dichotomous traits

An important consideration for reporting SNP heritability for categorical traits such as disease status is to make the distinction between the discontinuous observed and continuous liability scale heritability (Lee et al., 2011; Visscher et al., 2008). For categorical traits, although the underlying polygenic liability may follow a normal distribution in the population, the trait only has categorical outcomes (e.g. no disease or disease). In this model, it is thought that the underlying polygenic liability (together with other factors) pushes certain individuals past a threshold, after which the individual's phenotype will change categories (e.g. from no disease to disease) (Lee et al., 2011). This becomes a problem for estimating SNP heritability when a trait has low prevalence in the population. If the proportions of cases and controls were to be kept the same as the population prevalence, a study would have to recruit potentially hundreds of thousands of controls to gain enough cases to study the trait with substantial power. Instead, genotyping or sequencing studies tend to be over-represented for cases with respect to the population prevalence of the trait (which itself can also differ between populations). This is because after a certain point, the addition of more controls in effort to retain the population proportions of cases and controls does not gain much power for the analysis, and there is a considerable financial cost associated with such an endeavour. However, it is possible to do a mathematical adjustment to scale the observed h^2 estimate for dichotomous traits in a GWAS, to account for the proportion of cases in the study sample and the population prevalence of the trait (Lee et al., 2011).

1.3 Convergence of rare and common variant analyses

Due to the seemingly stark differences in genetic architectures between Mendelian and complex traits, for many decades there existed a dichotomy between rare disease-rare variant and common disease-common variant theories. However, this view has been gradually changing as we learn more about the genetic architectures

of different traits and diseases. This has happened particularly in the case of common complex traits, in which the role of rarer genetic variation has been revealed through new sequencing technologies.

1.3.1 Low frequency variants in common disease

Years of attempts to find low-frequency, intermediate effect variants through GWAS have so far not found many significant associations with complex diseases, with a few exceptions. One study that set out to uncover these low frequency variants was conducted on inflammatory bowel disease. The authors found one example of an intermediate effect, low frequency variant that was significantly associated with the disease. This variant in *ADCY7* conferred risk for Crohn's disease, and had a minor allele frequency of 0.6% in Europeans (Luo et al., 2016). For some traits, SNP heritability analyses partitioning SNP h^2 by MAF bins have shown evidence for intermediate frequency variants contributing to the trait. An example of this, a recent study by Hill et al. (2018) showed that ∼20% more heritability of intelligence can be explained by including variants with low MAF 0.1-1% (although the error margins for this estimate are wide). However, generally it is now thought that much of the missing heritability for complex traits is to be found in common variants with smaller effect sizes (Lee et al., 2011). This view is becoming more popular with larger and larger GWAS explaining more variation using variants with tiny effect sizes.

1.3.2 Rare variants in common disease

Although, collectively rare variants likely do not contribute nearly as much to complex trait heritability as do common variants, it is now evident that rare variants play a role in complex diseases as well (Singh et al., 2016; Genovese et al., 2016a; Luo et al., 2016; Fuchsberger et al., 2016). There is also now more evidence that rare and common variants can affect the same genes or biological pathways in complex diseases. For example, the Crohn's disease susceptibility locus *NOD2* has been implicated in both linkage studies, GWAS (Liu and Anderson, 2014) and

more recently in rare variant burden analyses in sequence studies (Luo et al., 2016). Schizophrenia GWAS have also found associations for calcium channel genes and targets of the *FMRP* gene, and both groups of genes have been implicated in rare variant burden analyses (Purcell et al., 2014). Even so, the effect sizes of rare and intermediate variants in complex disease have not reached similar magnitudes as Mendelian disease variants.

1.3.3 Common variants give rise to extreme phenotypes

It is often expected that rare variants cause more severe phenotypic outcomes than common variants, if the trait is under negative selection. However, some recent work has shown that for some complex diseases, both rare variants and common variants can cause extreme phenotypes. One example of this is a study by Natarajan et al. (2017), which showed that the overall polygenic load (polygenic scores, more in Chapter 2) had a similarly large effect on LDL-C cholesterol levels as did rare variants in known hypercholerolemia genes. In addition, from those individuals who had clinically high LDL-C levels, only 2% had a rare variant where 23% had a high polygenic score. Thee results demonstrate that common variation can play an important role in severe traits. Similarly, a recent study by Khera et al. (2018) found that particularly for coronary artery disease, but also for atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, individuals with high polygenic scores had a similar risk of disease as those with monogenic forms of the diseases. This study particularly has sparked global debate over the usefulness of polygenic scores in predicting disease in the clinic.

1.3.4 Do common variants contribute to rare disease?

Evidence from chromosomal abnormality syndromes

There have not been many studies investigating whether common variants contribute to rare, severe forms of disease. This is despite variable expressivity of Mendelian disease phenotypes between patients being a known phenomenon. It would seem

plausible that that expressivity could be affected by the genetic background, nongenetic factors (e.g. skewed X-inactivation in Duchenne muscular dystrophy (Abbadi et al., 1994)) or a combination of both. Interestingly, in literature on chromosomal abnormalities, the idea of variable expressivity driven by inherited common variants dates back to the 1970s (Moreno-De-Luca et al., 2015). In the CNV disorder and aneuploidy field, it was suspected that background genetics of the patient affected their performance outcome. This was because despite their disease, the IQ of patients with chromosomal abnormalities (Olszewski et al., 2014) (including trisomy-21 (47, XX/XY + 21) (Fraser and Sadovnick, 1976) and height for patients with Turner's (45, X0) and Klinefelter (47, XXY)(Brook et al., 1977)) correlated with parental phenotypes for cognitive performance and height. The same was reported for IQ and height in Prader-Willi Syndrome in 2000 (Malich et al., 2000). These studies implied that background genetic effects could be playing a role in the phenotype outcome.

More recently, a study by Moreno-De-Luca et al. (2015) described that 16p11.2 deletion patients had a 1-2SD decrease in intelligence, social functioning, motor functioning and body mass index compared to their parents and healthy siblings. However, the correlation of patient-family phenotypes were similar to the correlation between children and parents in the general population. The authors suggest it would be possible to predict a range of social and cognitive performance metrics for the affected child based on the parental phenotypes.

Common variants in monogenic disease

As mentioned, there are only a few examples of modifier effect from common variation to monogenic diseases, with oligogenic modifiers of *CFTR* in cystic fibrosis being one of the best known ones (Cutting, 2010). There are also now some examples of common variants modifying the outcomes of monogenic, but not fully penetrant diseases. Many of these examples are from the cancer and lipid fields, where the effects of polygenic scores have been assessed for patients with familial, severe effect mutations. As an example of this for breast and ovarian cancers, Kuchenbaecker et al. (2017) investigated the effects of polygenic scores

on the cancer risk in carriers of pathogenic *BRCA1* and *BRCA2* variants. Not all carriers of deleterious *BRCA1* and *BRCA2* carriers develop breast or ovarian cancers as these cancers are variably penetrant, but the lifetime risk for carriers of these is extremely high. The authors of the study found that e.g. *BRCA2* carriers with 10th percentile PRS (low polygenic risk) for ovarian cancer had a 13% lower risk of ovarian cancer by 80 years of age, than did those with a polygenic score at the 90th percentile (high polygenic risk). Another example of common variants acting together with rare familial variants is a paper by Talmud et al. (2013), where the authors compared polygenic scores for hypercholesterolemia in patients with and without familial mutations to healthy controls. The authors found that patients with familial mutations had lower polygenic scores for the trait than patients without familial varaints, but still higher scores overall than healthy controls.

Examples of common variants influencing the expressivity of monogenic diseases are very few. Recently though, one of the first exciting examples of this came from a study on Huntington's disease. A study by Hensman Moss et al. (2017) described a GWAS against a measure of Huntington's disease progression. They discovered a significantly associated locus which spanned three genes on chromosome 5. They described how the gene *MSH3* in this locus was a likely modifier of Huntington's disease progression, and found that the modifier effects were independent of the age of onset of disease. The findings from this study are very interesting as they represent one of the first known GWAS modifier associations for fully penetrant monogenic disease. Another GWAS (Bezzina et al., 2013) identified three common variant associations increasing risk of Brugada syndrome, a rare cardiac arrhythmia disorder. The syndrome is thought to have dominant Mendelian inheritance, but it has also been shown to have low penetrance in families with familial mutations, and to affect family members who do not carry these mutations.

1.4 Investigating common variants in rare, likely monogenic disorders

In this thesis, I will focus on investigating common variant effects in the context of rare neurodevelopmental disorders, which have been thought to be almost entirely monogenic. As described above, this work represents one of the first studies looking for common variant effects in presumably monogenic disease. In Chapters 2 and 3, I will focus on analysing data from patients suffering from these disorders. Then in Chapter 4, I describe further analyses looking at whether common variants modify penetrance of rare, deleterious variants that are observed in the general population.

Chapter 2

Common variants contribute to rare neurodevelopmental disorders

2.1 Chapter overview

In this chapter, I address the question of whether inherited common genetic variation plays a modifying role in severe neurodevelopmental disorders, that have been thought to be almost entirely monogenic. I begin by identifying individuals from the Deciphering Developmental Disorders Study (DDD) who had at least one abnormality affecting the central nervous system morphology or physiology. I then perform a discovery GWAS on neurodevelopmental disorder risk using controls from the UK Household Longitudinal Study. Through SNP heritability analysis of the GWAS results, I show that there is a significant contribution to these disorders from common genetic variation. I then replicate this finding in an independent set of proband-parent trios from the DDD Study, by showing over-transmission of neurodevelopmental disorder risk from parents to patients. For this analysis I utilise polygenic scores constructed from the discovery GWAS. I will next introduce the DDD study and discuss background to polygenic scores.

These scores are extensively used in complex trait genetics, and I implement them in several analyses throughout this thesis.

2.2 Background

2.2.1 Severe neurodevelopmental disorders

Developmental disorders are a collection of disorders that manifest in early childhood, and severely impact the child's normal growth and development. In the UK, estimates of congenital abnormalities and/or developmental disorders ranges from ∼2-5% (Deciphering Developmental Disorders Study, 2017). Usually when no other environmental causes are identified, the disorder is thought to be genetic. Developmental disorders often include abnormalities affecting the central nervous system (neurodevelopmental disorders), resulting in cognitive and motor delay, and impairment of social functioning (Sontheimer, 2015). Examples of neurodevelopmental disorders include global developmental delay, intellectual disability and autism. However, developmental disorders can also affect other organ systems than the nervous system, and can include morphological anomalies (dysmorphology). The treatment and disease management opportunities largely depend on the specific disorder. For example, metabolic disorders if diagnosed early may be managed with dietary changes. Other forms of disease management include e.g. language and behavioural therapy (Myers et al., 2007).

Often children with developmental disorders show severe symptoms and phenotypes, whilst their parents appear normal. Research into the underlying genetic architecture of these diseases has shown that particularly in families with high autozygosity (parents are related), the disorder can often be the result of recessive inheritance of rare variants (homozygous for the deleterious variant) (Martin et al., 2017b). More often though, developmental disorders with genetic causes are due to *de novo* variants in dominant or recessive developmental disorder genes, particularly when there is no inbreeding in the family (Martin et al., 2017b). As these variants are on their own sufficient enough to cause severe phenotypes, they are likely to be

located in the protein-coding region of the genome, disturbing protein structure or function (MacArthur and Tyler-Smith, 2010; Garcia-Alonso et al., 2014). A high proportion of *de novo* variants observed in developmental disorders is also a reflection of the reduced reproductive capacity of the patients. Chromosomal abnormalities can also span monogenic disease genes, resulting in developmental disorders, though in these cases identifying the causal gene(s) has been difficult (Moreno-De-Luca et al., 2015; Bergbaum and Ogilvie, 2016).

There are currently 1,767 genes associated with developmental disorders (Firth et al., 2009) (downloaded 30th August 2018), of which 32% are monoallelic and 61% biallelic (including overlapping genes). In order to find causal rare variants, and to assess whether they were inherited or occurred *de novo*, it is particularly helpful to look at exome sequence data for trios (Wright et al., 2015). This is because exome sequencing is still relatively cheap compared to whole genome sequencing (Sazonovs and Barrett, 2018), and due to the severity of the phenotypes it is expected that the causal variant is within the protein-coding region of the genome (Wright et al., 2015). The benefit of the trio design is to help identify the pattern of inheritance, which greatly helps to inform families on the recurrence risk of the disorder.

Many neurodevelopmental disorders shared phenotypic symptoms, including reduced cognitive function, seizures, autism and schizophrenia. Studying chromosomal abnormalities has shed some light to which genes could be causal for such diverse symptoms. By assessing the symptoms of e.g. patients with differential length deletions spanning the same region, it is possible to refine likely causal genes within that region (Theisen and Shaffer, 2010). With the availability of WES and WGS, there is now evidence that variants in the protein-coding regions of genes are also associated with different neurodevelopmental and neuropsychiatric disorders (Zhu et al., 2014). For example, a study by Singh et al. (2017) found that rare variants in genes associated with neurodevelopmental disorders were enriched in schizophrenia patients with intellectual disability. These studies point towards overlap between genes associated with different disorders of the brain.

These previous studies have sensibly focused on identifying rare genetic variants, since such rare and severe disorders seem likely to be caused by highly deleterious

variants that would rapidly be removed from the population. However, (as detailed in section 1.3.4) the idea that common variant background could modify severe disorder outcomes or expressivity is not new in the field. In order to detect common variant effects contributing to these disorders, a genome-wide analysis would need to be done on a large number of patients, ideally in a large batch using the same genotyping chip in order to avoid biases from combining numerous small datasets. There now exists a dataset in which such a scan is possible, namely the Deciphering Developmental Disorders (DDD) Study (Wright et al., 2015).

This DDD study aims to find a genetic diagnosis for patients suffering from previously genetically undiagnosed developmental disorders. In 2011 to 2015, the study recruited ∼14,000 patients with neurodevelopmental disorders, congenital, growth or behavioral abnormalities, and dysmorphic features. Recruitment was done in the UK and Ireland through 24 genetics services. Each patient was assessed by a clinical geneticist, who deemed that the likely cause for the disorder was genetic (monogenic). Most patients had undergone previous genetic testing, but all had remained genetically undiagnosed at the time of recruitment. Phenotypic data were recorded for the majority of DDD patients, and some individuals had growth measurements and prenatal information as well.

In order to find genetic diagnoses for families, the DDD Study is exome sequencing all individuals recruited to the study, including parents whenever possible. This trio design has so far been successful in unravelling new developmental disorder associated genes (Deciphering Developmental Disorders Study, 2017), aiding new diagnoses to be made. Importantly, clinically relevant variants are curated by the clinical geneticists who recruited the families. When a diagnosis is found, the information is then fed back to the families who receive genetic counselling.

Exome sequencing of DDD trios has unveiled plenty of information about the genetic architecture of previously undiagnosed developmental disorders (Short et al., 2018; Lord et al., 2018; Martin et al., 2017b). The most recent published analysis of *de novo* mutations in the first ∼4,000 trios estimated that ∼40% of the cohort have causal protein-coding *de novo*s, but these have not all been identified yet due to lack of power (Deciphering Developmental Disorders Study, 2017). In

addition, the DDD Study has for the first time shown evidence that *de novo* mutations in non-coding elements of fetal brain-expressed genes also contribute to the disease burden in the cohort (Short et al., 2018). *De novo*s in canonical splice site regions have also been shown to contribute to disease in the cohort (Lord et al., 2018). Finally, it has also been estimated that from the European subset of DDD patients, 3.6% carry diagnostic autosomal recessive variants, whereas in patients with Pakistani ancestry this fraction is much higher at 31% due to elevated autozygosity. It is important to bear in mind though, that since the patients recruited to the Study have already undergone clinical assessment and usually genetic testing before recruitment, the cohort is depleted for clinically recognisable genetic disorders (Deciphering Developmental Disorders Study, 2017).

Importantly for the work presented in this thesis, the majority of DDD patients were also genotyped on a DNA chip. This means that DDD represents the largest cohort of genotyped patients with rare, undiagnosed developmental disorders. This dataset allows us to test whether common genetic variants contribute to heterogeneous, rare developmental disorders, or whether the genetic contribution to these disorders is solely attributable to rare variants. In addition, since we have data available on the rare variants for these individuals, we can ask whether there is a difference between the polygenic background of patients with and without rare diagnostic variants in the exome. In Chapters 2 and 3, I explore the common variant architecture of specifically neurodevelopmental disorders, using data from the DDD Study.

2.2.2 Polygenic scores in genetic studies

In this chapter, I describe two main analyses. The first is to conduct a GWAS using data from the DDD Study. The second analysis uses polygenic scores, which are an important tool for leveraging genome-wide data in the complex trait field. Polygenic scores can be used to quantify polygenic effects in a given cohort by utilising pre-existing information from other GWAS (International Schizophrenia Consortium et al., 2009). A polygenic score for an individual is the sum of tiny predicted effects from thousands (or millions) of variants discovered in an independent study of a given phenotype given the individual's genotype. The genetic effect estimates

for the variants are typically obtained from GWAS. The polygenic scores are constructed by taking the β (effect) for the effect allele at a single variant, and multiplying this by the individual's allele count at that locus (0, 1 or 2 for variants with two possible alleles). The same procedure is repeated at each variant of interest, and these are then summed to give one polygenic score for each study participant (Polygenic score = $\Sigma\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$), (Figure 2.1).

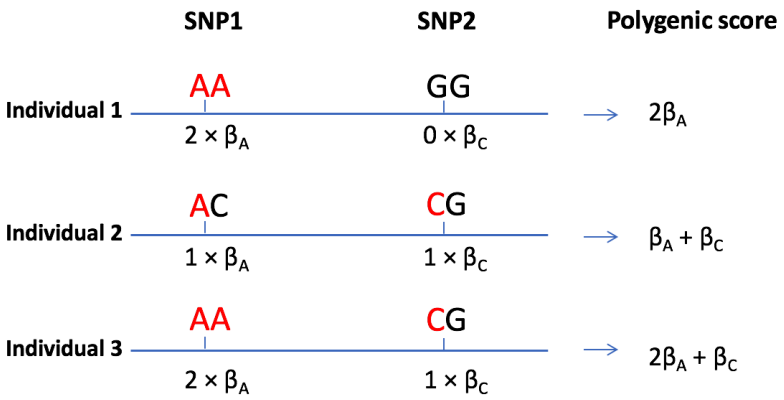


Figure 2.1: Illustrative figure of how polygenic scores are calculated in three study individuals, who have each been genotyped for two variants. Variant effect sizes or betas (β) from a GWAS are multiplied by the effect allele count in each individual (red = known effect allele). All effects are summed over for each individual to create a risk score for that individual.

Polygenic scores can enable us to compare the distribution of polygenic burden of a specified set of variants for a given trait, between two or more groups (Figure 2.2). As an example, a study by Talmud et al. (2013) used a polygenic score for higher low density lipoprotein C, to assess polygenic risk for high cholesterol in individuals diagnosed with familial hypercholesterolemia, with and without a confirmed diagnostic variant. They found that individuals with an unexplained case of the familial disease had elevated polygenic scores compared to those with a known variant. However, affected carriers of diagnostic variants also had significantly higher polygenic scores than controls with normal cholesterol levels. In this study, the authors used a polygenic score constructed from 12 SNPs which had previously been associated with higher cholesterol in blood. It is more common practice now to also include variants which are not all necessarily associated with the trait at genome-wide significance (International Schizophrenia Consortium et al., 2009).

Typically in this approach, variants are selected to be added to the score set if they pass a certain P-value threshold. The variants are also thinned so that only one variant from each independent locus is included in the score. Some studies then construct multiple polygenic scores using different thresholds, and report results from all of these. Others choose to use a cutoff that has previously been shown to explain the most variance for the phenotype in an independent cohort. Methods also exist to include all variants, but these will have to account for decreased accuracy when estimating the effect sizes for low frequence variants in a GWAS (Vilhjálmsson et al., 2015). By when adding more variants to a polygenic score, it is likely that more of the effects that truly contribute to trait h^2 are also included. However, the drawback to this is that it can add noise or bias to the scores.

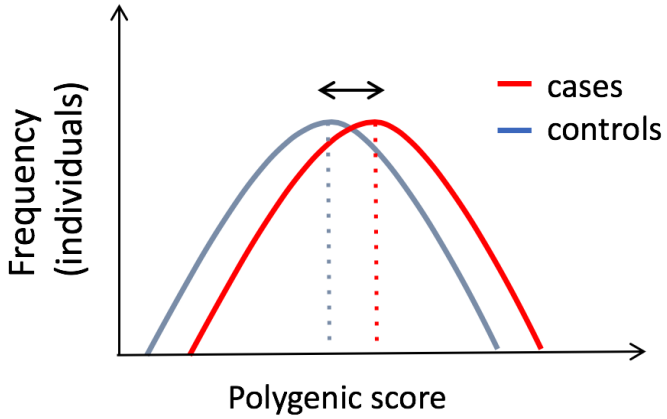


Figure 2.2: In this illustration, the distribution of polygenic scores in cases (red) is shifted to the right (higher risk) from the control distribution (in blue). The cases have on average a slightly elevated polygenic risk compared to the controls for the trait or disease in question. The tails of the distributions may also give important information about differences in polygenic burden in two groups.

2.3 Contributions and publication note

Genotyping of DDD samples was done by the Sanger Institute genotyping facility. Daniel Rice wrote custom scripts that allowed for efficient computation of polygenic scores. Hilary Martin contributed to the supervision of this work. The work

described in this chapter was completed under the supervision of Jeffrey Barrett, and was published in Niemi et al. (2018).

2.4 Methods

2.4.1 Datasets and quality control

In order to measure polygenic effects on neurodevelopmental disorders in the DDD Study, I performed a GWAS. As controls for this study, I used individuals recruited as part of the UK Household Longitudinal Study, who are not suffering from early childhood onset, severe neurodevelopmental disorders. In this section I will describe the cohorts I used, the genotype data, and the selection of samples and variants for GWAS.

Deciphering Developmental Disorders

Recruitment and phenotyping of DDD patients is described in detail in Wright et al. (2015) and Deciphering Developmental Disorders Study (2015). Families gave informed consent for participation. Briefly, the DDD study recruited patients with a previously undiagnosed developmental disorder, in the UK and Ireland. Recruitment was done by senior clinical geneticists who had assessed that each patient's disorder was sufficiently early-onset and severe that it was likely monogenic. Patient phenotypes were systematically recorded by clinical geneticists using Human Phenotype Ontology (HPO) terms in a central database, DECIPHER (Firth et al., 2009). Most patients are recruited at a young age, with the mean decimal age at assessment being 7.7 years, but, 6% of patients were recruited as adults over the age of 18 years. The DDD Study genotyped 11,304 patients on the Illumina HumanCoreExome. In addition, a cohort of 930 full trios, with the addition of some parent-proband duos and proband singletons were genotyped on the Illumina HumanOmniExpress chip. Genotyping was carried out by the Wellcome Trust Sanger Institute genotyping facility. All data were on GRCh37, and detailed information of genotyping chips is shown in Table 2.1.

Table 2.1: Quality control for UK cohorts.

		Quality control steps - DDD and UKHLS data	DDD trios	DDD probands (1)	DDD probands (2)	UKHLS controls
DNA chip		NA	HumanOmniExpress 12v1 BeadChip	Human CoreExome 24v1.0 BeadChip	InfiniumCoreExome 24v1.1_A Beadchip	Human CoreExome-12v1-0_B BeadChip
Pre QC	samples	NA	930	3,000	8,304	10,484
	variants	NA	811,844	547,644	551,839	538,403
Post sample and variant QC	samples	samples that passed QC	911	2,832	7,724	10,391
	variants	variants that passed QC and had MAF≥0.5%	587,655	246,506	246,506	246,506
Post imputation, neurodevelopmental GBR subset	samples	samples with non-GBR ancestry or without a neurodevelopmental phenotype excluded, one individual from related pairs removed (excl. trios)	728	1,966	5,021	9,270
	variants	imputed variants filtered for INFO≥0.9	4,934,465	4,134,438	4,134,438	4,134,438

UK Household Longitudinal Study

We obtained data from the UK Household Longitudinal Study (UKHLS) to use as controls for our discovery GWAS (University of Essex Institute for Social and Economic Research, 2018). The UKHLS cohort consists of a continuation of the British Household Panel Survey (BHPS) and additional ongoing recruitment of individuals living in the UK. Individuals were recruited to the study based on their postcode, with the aim to capture a representative population of the people living in the UK, and to collect extensive longitudinal data on these individuals. Study participation was incentivised with a monetary reward for every questionnaire completed. A registered-nurse visit was offered to UKHLS and BHPS participants during waves 2 and 3, spanning years 2010-2012 (University of Essex Institute for

Social and Economic Research, 2014). All those aged 16 and above were invited to take part. Upon consent, blood samples were taken during the nurse visit, and these were used for genotyping. Genotyping of 10,484 UKHLS samples was carried out by the Wellcome Sanger Institute on the Illumina HumanCoreExome chip. All data were on GRCh37, and detailed information of genotyping chips is shown in Table 2.1.

All participants in the nurse visit were asked about their general health, but they were not excluded from giving a blood sample based on disease status other than blood borne disease e.g. HIV, as long as they were healthy enough to undergo nurse interview and assessment (University of Essex Institute for Social and Economic Research, 2014). Since the study participants who were invited to give a blood sample were at least 16 years of age at the time of sampling, the UKHLS cohort mean age was higher than that of DDD. However, due to the fact that individuals were not deliberately excluded on the basis any diseases or traits, the expectation is that the distributions of alleles associated with these traits are relatively close to the population distributions. This also means that the UKHLS may include some individuals who have a diagnosis for complex diseases, e.g. nonpsychiatric diseases such as inflammatory bowel disease, diabetes type 1 and 2, etc. Therefore, since neither the UKHLS nor DDD ascertains participants based on the presence of any of these diseases (or who would develop them), we would expect the distribution of risk alleles for these non-neurodevelopmental complex diseases to be similar within the DDD and UKHLS cohorts. Therefore the difference in age or potential presence of individuals with complex diseases in the UKHLS is not a concern for our GWAS.

Quality control of datasets

Sample and variant quality control is essential to remove biases that may arise from e.g. ascertainment or the genotyping process, which may lead to spurious variant-phenotype associations in a GWAS. I performed variant and sample quality control for each dataset separately, adapting the protocol suggested by Anderson et al. (2010). I received all data in PLINK format as hard-called genotypes. Specific steps that I took are summarised in Table 2.2. Briefly, I removed variants that were

missing in ≥3% of samples, and samples that had ≥3% of their genotypes called as missing by the genotyping algorithm. I also assessed the proportion of heterozygous genotypes per individual, and removed samples that had high heterozygosity to control for admixture or low heterozygosity which implies consanguinity of the parents (±3 standard deviations from the mean). I then removed one of each pair of sample duplicates, which I defined on the basis of two samples sharing alleles identical by descent ≥0.98. The HumanCoreExome chip contains a high proportion of rare variants with minor allele frequency (MAF) ≤0.005 (45% of variants), which are likely to be enriched for genotyping errors. In order to minimise potential biases resulting from this, I removed rare variants before imputation for these dataset. As an additional quality control step in the DDD trios data, I removed families that had an elevated numbers of Mendelian errors.

Table 2.2: Sample and variant quality control parameters.

Sample quality control
Reported sex inconsistent with data
Sample genotype missingness $\geq$3%
$\pm$3 standard deviations from mean heterozygosity (control for inbreeding and admixture)
Sample duplicates (alleles identical by descent $\geq$98)%
High number of mendel errors in trio >2000
Variant quality control
Variant genotype missingness $\geq$3%
Chromosome and position duplicates
Position other than chromosomes 1-22
Hardy Weinberg Equilibrium test P<1 $\times 10^{-5}$
Strand information unavailable for SNP
Alleles discordant between case and control datasets
Alleles and frequency in Europeans discordant with HRC v1.1
Differential missingness between cases and controls P<1 $\times 10^{-20}$
Post-imputation
Variants INFO $\leq$0.90
Samples non-GBR ancestry
Relatives in discovery GWAS (proportion of alleles identical by descent >0.12), sample with higher missingness removed

Selecting samples with European ancestry

Variation between populations is typically correlated with geographical location (Campbell et al., 2005; Novembre et al., 2008). Therefore comparing the allele frequencies between two randomly selected cohorts would result in associations that have nothing to do with the intended trait measured but instead are different due to systematic differences in the population structure between these groups. Therefore case-control GWAS typically attempt to maximise homogeneity between individuals with respect to their genetic ancestry before analysis. Although genetic

ancestry tends to be more homogeneous among samples collected in the same geographical region or country, neither the DDD nor the UKHLS recruited their study participants using information about their genetic ancestry (more discussion below), and so the expectation was that these included individuals from different ancestral populations. I therefore checked the genetic ancestry of individuals before conducting the GWAS.

I defined sample genetic ancestry based on a projection principal component (PCA) analysis using PLINK with 1000 Genomes Phase 3 populations (1000G). For this analysis, I used only variants with a minor allele frequency (MAF) of $\geq$0.10 to reduce bias from rare alleles, and variants that were not in linkage disequilibrium with each other. A principal component analysis tries to fit a statistical model, where independent components explain variance in the data. The first principal components also typically correlate with ancestry and geographic location, and therefore selecting samples based on their clustering on PC1 and PC2 axes is often used for determining population ancestry groups (Novembre et al., 2008). Datasets such as the 1000 Genomes panel include individuals from a number of different geographical and ancestral groups, and therefore these are often used as reference panels for determining ancestry of new datasets. In my projection PCA, I used the 1000G samples to determine the genetic distance between individuals within 1000G (Figure 2.3a), and then projected my DDD and UKHLS samples on top of these, in order to assess where they lie in relation to the 1000G samples. The largest cluster of DDD and UKHLS samples that overlay with the 1000G samples had Great Britain ancestry. As illustrated by the PCA plots in Figure 2.3 b and c, the genetic ancestry of DDD patients and parents was quite diverse. Because the DDD Study recruited patients with the aim of exome sequencing trios, and within-trio analyses are immune to population structure, recruitment of individuals with heterogeneous ancestries was not a concern for the Study. The PCA may also reflect an enrichment of South Asian ancestry in the DDD, since consanguinity increases the risk of developmental disorders (Martin et al., 2017b).

When doing case-control GWAS, it is also customary to only include individuals who are not related more closely than second-degree relatives. Otherwise it is possible for genotypes found within these families to become over-represented relative to

the population allele frequency (Anderson et al., 2010). Increasing the sample size for GWAS can increase power to detect association, so some GWAS software such as BOLT-LMM (see section 2.4.3) build a relationship matrix from the genotypes of individuals. This should in theory make the analysis immune to population structure. In this Chapter, I use BOLT-LMM to conduct my discovery GWAS, however, in practice I saw that removing relatives from the analysis strengthened our downstream findings. I therefore identified pairs of related individuals equivalent to second-degree relatives or closer (alleles identical by descent >0.12, using PLINK) from the case-control cohorts, and removed the individual who had a higher variant missingness rate out of the two. I also checked that individuals in the discovery case-control DDD and UKHLS cohorts were not related to individuals who were included in the DDD trios (alleles identical by descent >0.12).

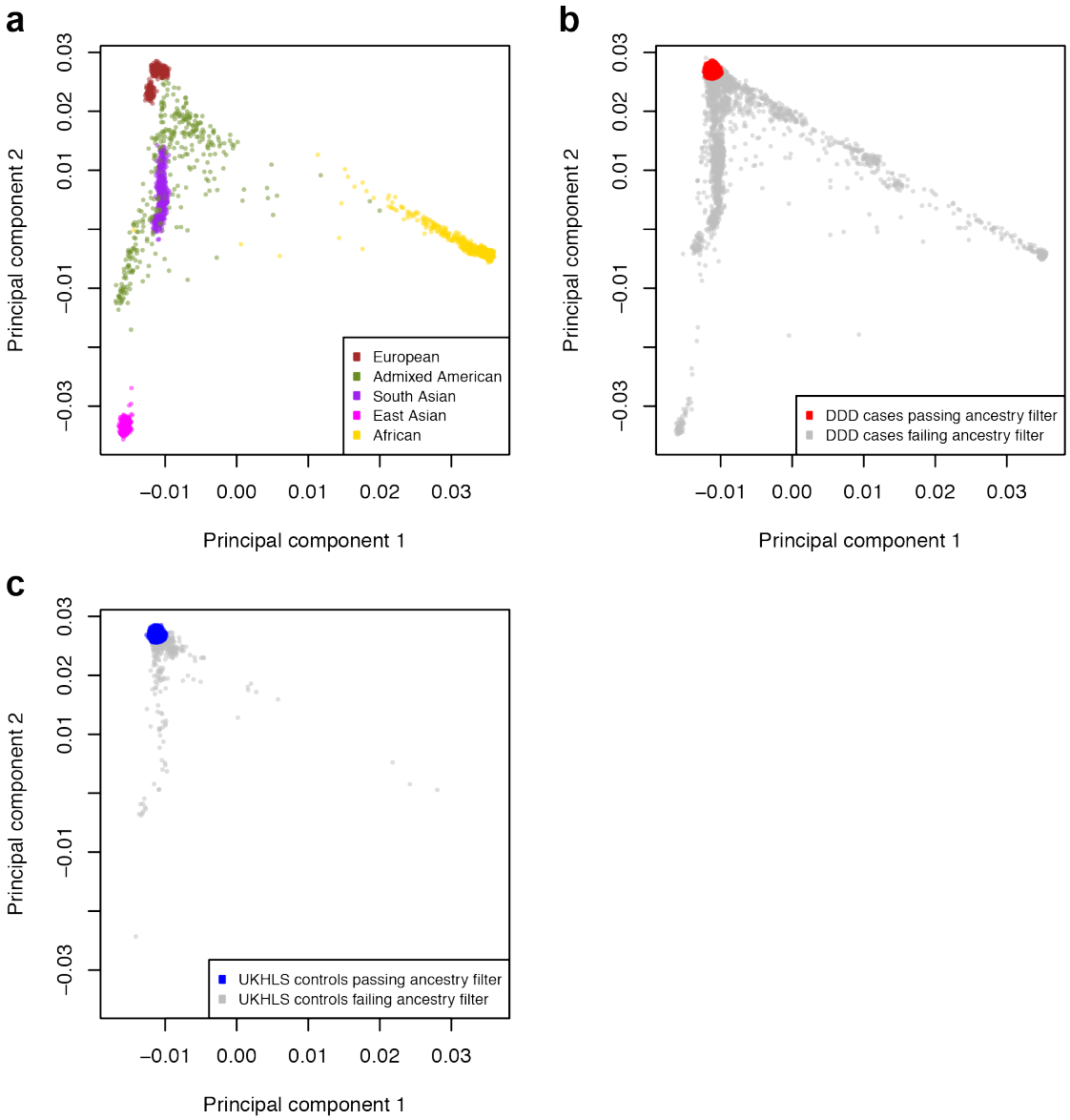


Figure 2.3: **a**, Reference samples (N=2,504) from 1000 Genomes Phase 3, colored by the five super populations, used for a projection PCA of UK cohorts (DDD and UKHLS). **b**, All DDD cases (discovery N=11,304 and trios N=930) from projection PCA with 1000 Genomes. Case samples with European ancestry are plotted in red and non-Europeans in grey. **c**, All UKHLS controls (N=10,396) from projection PCA with 1000 Genomes. Control samples with European ancestry are plotted in blue and non-Europeans in grey. All cases and controls coloured in grey (panels **b** and **c**) were excluded from analysis due to non-European ancestry.

Phasing and imputation

Genotyping chips capture only part of the known common variation in the genome, but it is possible to fill in the gaps, or to 'impute', the missing variants by inferring them from the surrounding markers (Marchini and Howie, 2010). The sample is compared to a panel of reference haplotypes, which allows for the best guess of the missing genotypes in the target sample. This process of inference from surrounding markers is made more efficient by first phasing, i.e. constructing haplotypes from the genotyped markers. Imputation is often done for GWAS samples to boost the coverage of the genome for association testing, and to increase overlap between datasets genotyped on different chips. In the context of this study, imputation allowed me to include more variants shared between the DDD cases and healthy controls, boosting coverage, as well as between the discovery and replication cohorts. Stringent quality control before imputation is key to avoiding amplification of biases that may arise from subtle differences between batches of data, arising from e.g. differences in missingness or genotyping error in the original genotype-calling.

After sample and variant quality control, I phased and imputed the discovery GWAS cohorts (DDD singletons and UKHLS), genotyped on the HumanCoreExome backbone, together using variants that intersected between the different versions of the chip (Table 2.1). I then phased and imputed trios that were genotyped on the HumanOmniExpress in a second batch, due to the small number of overlapping variants with HumanCoreExome chips. I used the Sanger Institute Imputation Service (McCarthy et al., 2016) to carry out phasing and imputation, using Eagle2 (v2.0.5) (Loh et al., 2016) and PBWT (Durbin, 2014) software respectively. For imputation, I selected the Haplotype Reference Consortium as the reference genotype panel (release 1.1, chr1-22, X) (McCarthy et al., 2016). After imputation I removed variants that had a missingness of >0.05 or an INFO score <0.9.

Phenotype data in DDD Study

In a case-control GWAS of a heritable trait, the phenotype for which cases are recruited to the study is usually well defined. However, the DDD cohort comprises

of patients with thousands of different phenotypes, likely with numerous different genetic contributors. I therefore first tried to increase power for association testing by refining the phenotype which we wanted to test. We decided to take the approach of selecting patients with at least one phenotype HPO term that indicated an abnormality of the central nervous system. The HPO tree begins with the root term phenotypic abnormality, and descends into organ system level, and further down to more specific phenotypic terms, as illustrated in Figure 2.4. I first manually studied the HPO term tree in order to define which groups of terms were associated with the central nervous system. I then ran a HPO text search for patients who had at least one of the following HPO terms or daughter terms of abnormality of the nervous system morphology (HP:0012639) or the following physiological sub-abnormalities: abnormal metabolic brain imaging by MRS (HP:0012705), abnormal brain positron emission tomography (HP:0012657), abnormal synaptic transmission (HP:0012535), abnormal nervous system electrophysiology (HP:0001311), behavioural abnormality (HP:0000708), seizures (HP:0001250), encephalopathy (HP:001298), abnormality of higher mental function (HP:0011446), neurodevelopmental abnormality (HP:0012759). The neurodevelopmental patient subset included both individuals who have, since recruitment to the DDD study, been found to carry diagnostic mutations in protein-coding genes (Wright et al., 2015; Deciphering Developmental Disorders Study, 2017; Martin et al., 2017b; Short et al., 2018), and individuals for whom no likely diagnostic rare variant has yet been found. The definition for the main phenotype in this study is therefore neurodevelopmental disorder risk, which is the risk of having a previously undiagnosed developmental disorder, being recruited to the DDD study, and having at least one neurodevelopmental HPO (Figure 2.5).

In addition to HPOs, some DDD patients had clinical growth measurements for height (78% of unrelated patients with European ancestry), birth weight (93%) and head circumference (87%). These measurements had been standardised by the DDD Study to reflect departure (standard deviations) from the age and sex-adjusted population means. I pulled these adjusted metrics from the DDD Study internal phenotype database, which is acquired from DECIPHER database.

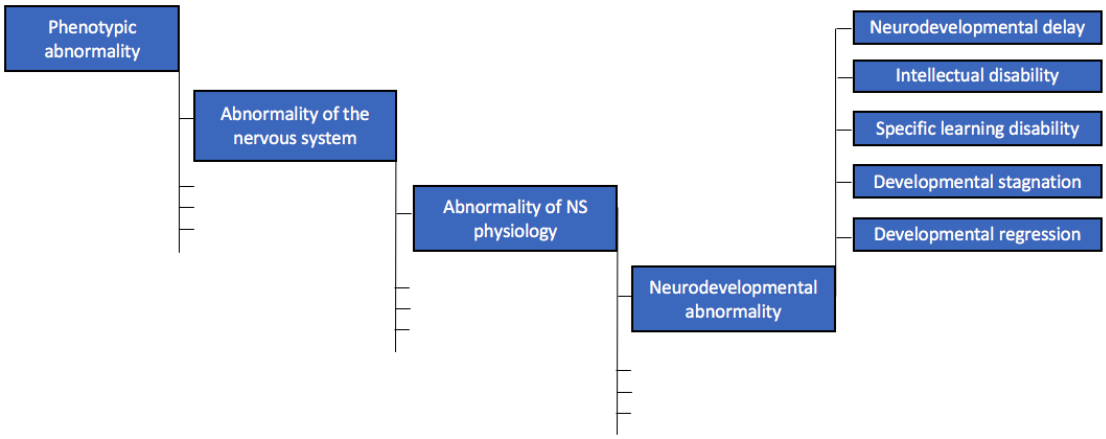


Figure 2.4: Illustration of the HPO tree. The term "phenotypic abnormality" descends down to abnormalities of different organ systems, and further down to more specific phenotypic terms.

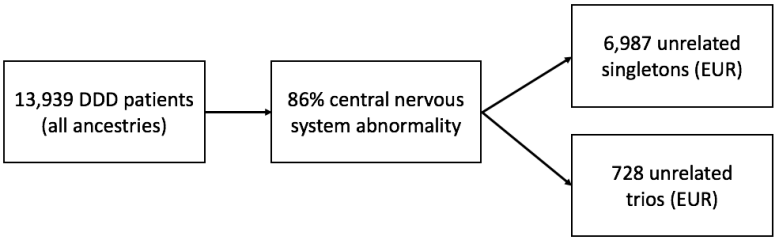


Figure 2.5: Summary of the DDD study samples.

2.4.2 Counting affected organ systems

We wanted to determine how many distinct organ systems were affected in each patient who was in included in the final sample (Figure 2.5). This was complicated by the fact that many HPOs fell under more than one organ system category. For example, microcephaly, which is a common term in the cohort, falls under "nervous system", "head or neck" and "skeletal system". In order to assign each HPO into only one organ system, I used a ranked organ system approach. To do this, I first ranked organ systems based on the number of raw counts of individuals with at least one term under that system (Table 2.3) in the full DDD cohort. After ranking organ systems, I then looked for individuals with at least one HPO under the system ranked most commonly affected (in this case the nervous system), and

assigned these individuals an organ system count of 1. I then removed these HPOs from the patients' lists, before continuing to identify individuals with at least one HPO in the organ system ranked second most prevalently affected (in this case head or neck). I continued to count organs and remove HPOs until we had assigned all individuals a count of organs systems affected out of 19 non-overlapping systems.

Table 2.3: Proportions of DD patients who have at least one HPO term belonging to a particular organ system category. The HPO tree descends down from "phenotypic abnormality", to different organ systems, down to specific terms describing particular phenotypes. Each HPO term used by clinicians to describe patients was traced up the tree to the organ system level. However, some HPOs may belong to more than one organ system category. For example, microcephaly will be counted under "nervous system", "head or neck" and "skeletal system" in the HPO tree, whilst global developmental delay will only appear under "nervous system".

Rank	Organ system	% All DDD patients (N=13,558)	% Neurodevelopmental subset of unrelated DDD patients, GBR ancestry (N=6,987)
1	Nervous system	87	100
2	Head or neck	68.9	71.2
3	Skeletal system	61.7	61.8
4	Limbs	35.1	35.3
5	Eye	34.9	35.3
6	Integument	31.2	31.9
7	Ear	20.1	19.7
8	Digestive system	20	19.1
9	Musculature	19.9	18.7
10	Cardiovascular system	15.1	13.5
11	Genitourinary system	12.4	11.4
12	Respiratory system	8.1	7.3
13	Connective tissue	7.4	6.3
14	Immune system	6.8	6.5
15	Endocrine system	4.1	4.1
16	Metabolism homeostasis	4.1	4
17	Breast	3.7	3.7
18	Blood and blood forming tissues	2.1	2.1
19	Voice	1.1	1.1

2.4.3 Genome-wide association study

To conduct the GWAS of neurodevelopmental disorder risk, I used BOLT-linear mixed models (BOLT-LMM) (Loh et al., 2015b). The method first builds a genetic relationship matrix (GRM) using a set of ∼500,000 thinned variants. Although our study phenotype is a dichotomous trait, the data is suitable for using BOLT-LMM, because it fulfils recommendations by the authors of the software: the discovery GWAS sample size is large, the MAF threshold we use is high (≥0.05), and cases and controls are well balanced (0.43 fraction of case) (*BOLT-LMM v2.3.2 User Manual* 2018; Loh et al., 2018). Using this method should control for cryptic relatedness and any remaining ancestry bias more accurately than e.g. adding ancestry PCs as covariates in a logistic regression for association testing. For the GWAS described in this thesis, I included sex as a covariate in the model.

I report a genomic inflation factor for the GWAS. Genomic control (λ_{GC}) quantifies the deviation of observed chi^2 test statistics from the expected null-distribution in a genome-wide association study. It is defined as a ratio median(observed chi^2)/median(expected chi^2) (Devlin and Roeder, 1999). Inflation of lambda (from λ_{GC}=1) indicates either true polygenic signal, or biases such as population structure in the data. Often the observed test statistics are visualised against the expected values in a quantile-quantile plot which can give an idea as to whether the data are behaving appropriately. Traditionally in GWA studies, λ_{GC} has been used to correct for confounding, as it is expected that SNP effects on all chromosomes are affected by the same bias (e.g. from population stratification), and therefore the test statistic is divided by λ_{GC}. In this thesis, however, I do not use genomic control on the test statistics, because BOLT-LMM should in theory handle potential non-polygenic biases.

2.4.4 SNP heritability using LD score regression

To estimate SNP heritability for discovery neurodevelopmental disorder GWAS, I used Linkage Disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015a), as the method is able to distinguish between confounding and polygenic effects.

One of the key benefits of using this method is that individuals' genotype-level data are not needed. Instead, the method uses summary statistics as input, which also greatly reduces processing and analysis time.

The principle of LDSC lies in the assumption that variants that tag (are in high LD with) more variants are also more likely tag a causal variant for a given phenotype. For this reason, those variants will have on average stronger association test statistics when there are real polygenic effects. However, confounding (e.g. population stratification) does not cause inflation in test statistics in proportion to LD. Specifically, the method involves regressing the association χ^2 test statistic at each SNP against the average linkage disequilibrium (LD) in that region, which reflects the extent to which that variant tags other variants (the LD score). The LD score for each SNP can be estimated from reference panels such as the 1000 Genomes European cohort. Best practice is to use scores derived from a population with matching ancestry to the cohort studied. The intercept from this regression can be transformed into an estimation of the proportion of phenotypic variation explained by effects other than polygenic, such as population substructure. The SNP heritability estimate is achieved by rescaling the slope of the LD score regression. If genomic control is required in a GWAS, the LD score intercept can be used as an effective alternative to the more conservative λ_{GC}, which does not distinguish between inflation from true polygenic signal and bias (Bulik-Sullivan et al., 2015a). However, in the absence of major inflation of test statistics and the fact that correction can downward bias LDSC estimates, I did not apply any genomic control to variant effects (betas) in work described in this thesis.

I used the LD score website LD Hub (Zheng et al., 2017), to estimate SNP heritability from the discovery GWAS neurodevelopmental disorder risk summary statistics (BOLT-LMM output). As recommended by the authors (Zheng et al., 2017), I removed the major histocompatibility complex (MHC) region (chromosome 6, 26-34MB) from the GWAS results before analysis due to its complex LD structure.

LDSC default output is SNP heritability (h^2) on the observed scale, however h^2 on the liability scale can be obtained by specifying the ratio of cases to controls in the study and the estimated population prevalence of the trait. In our scenario

though, estimating true population prevalence of neurodevelopmental disorders such as those included in my discovery GWAS is difficult due to several factors, including: (1) DDD participants included in the discovery GWAS were selected from a cohort of undiagnosed disorder patients specifically for neurodevelopmental abnormalities defined by HPO terminology; therefore any neurodevelopmental disorder patient whose condition was diagnosed through NHS clinical genetics clinics would not have been recruited to the DDD Study. (2) In addition, the DDD cohort consists of heterogeneous etiologies, including but not restricted to *de novo* coding mutations (Deciphering Developmental Disorders Study, 2017), bi-allelic inherited mutations (Martin et al., 2017b), *de novo* non-coding mutations (Short et al., 2018), and unexplained cases. I report SNP heritability assuming a prevalence of 1% in the population. By varying the prevalence between 0.2% and 2%, the SNP heritability estimate remained approximately within the 95% CI of the reported SNP heritability.

2.4.5 Polygenic scores

The formula I used for calculating polygenic scores took into consideration the β (effect size) of each known effect allele, and their allele counts in the target individual (Polygenic score = $\Sigma\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$). To select which variants to include in polygenic scores calculated from summary statistics for our developmental disorder risk discovery GWAS, I started by identifying which variants existed in both the discovery GWAS and the imputed data for DDD trios (replication set). This is because I could only use variants for which we know the effect (β) on neurodevelopmental disorder risk, to calculate the polygenic scores for this trait in the target trios. All variants that I selected had a MAF$\geq$0.05 in both the discovery GWAS and probands from the European trios, and had been directly genotyped or imputed with high confidence (INFO$\geq$0.9) in both datasets. To find independent variants to include in the scores, I pruned the remaining intersecting variants in the trios data using PLINK, which takes the top variant and removes variants within 500kb and that have $r^2 \geq 0.1$ with the top variant. PLINK then repeats the process until no SNP has a P-value below a pre-defined threshold. To

obtain this threshold, I did ten rounds of simulations where I first repeated the neurodevelopmental disorder risk GWAS having removed a random subset of 20% of cases and controls. I then calculated a neurodevelopmental disorder (NDD) risk polygenic score in the leave-out subset, and performed a logistic regression with 10 ancestry principal components to assess association of case-control status with the score. I tried different P-value thresholds which were $P<0.005$, $P<0.01$, $P<0.05$, $P<0.1$, $P<0.5$, $P<1$. I then chose a P-value threshold which resulted in a score that was most strongly associated with case/control status. The threshold $P<1$ performed best in ten independent permutations. As a note, there is currently no uniform protocol for how to define a P-value cutoff for polygenic scores. Some studies choose to use an a priori P-value cutoff that explains the most variance in a replication cohort as we have done, others report results for a range of thresholds, and some report only the analysis which post hoc resulted in the most significant results. After calculating the scores for each study individual using the predefined threshold of $P<1$, I normalised the proband scores and parental scores to have a mean of 0 and variance of 1.

2.4.6 Polygenic transmission disequilibrium test

I used the polygenic transmission disequilibrium test (pTDT) method (Weiner et al., 2017) to replicate neurodevelopmental disorder risk using trios data. The method compares the means of two polygenic score distributions: one comprising of scores of the probands, and the other of the average scores of parent-pairs. The test is equivalent to a paired, one-sample t-test, and assesses whether the mean of the score distribution in probands deviates from the mean of parent-pair score average, which is the expected score when there is random transmission. For pTDT analysis of neurodevelopmental disorder risk in trios, I report a one-sided P-value because our expectation was that the direction of transmission would be accumulation of risk alleles in affected children.

2.5 Results

2.5.1 Discovery GWAS for neurodevelopmental disorder risk

After removing relatives and non-European ancestry in the DDD Study, 86% of the remaining patients had at least one abnormality affecting the central nervous system. This left 6,987 unrelated DDD patients for our discovery GWAS. Some of the most common phenotypic abnormalities in this neurodevelopmental subset included global developmental delay, intellectual disability, cognitive impairment or learning disabilities (in 86% of the neurodevelopmental subset) and autism spectrum disorders in (16%), among others. Some of the more clinically relevant phenotypes (Wright et al., 2015) observed in the full DDD cohort and neurodevelopmental subset are shown in Figure 2.6 a. In addition to the neurodevelopmental phenotype, 88% of these patients' disorder included an abnormality affecting at least one other distinct organ system (Figure 2.6 b).

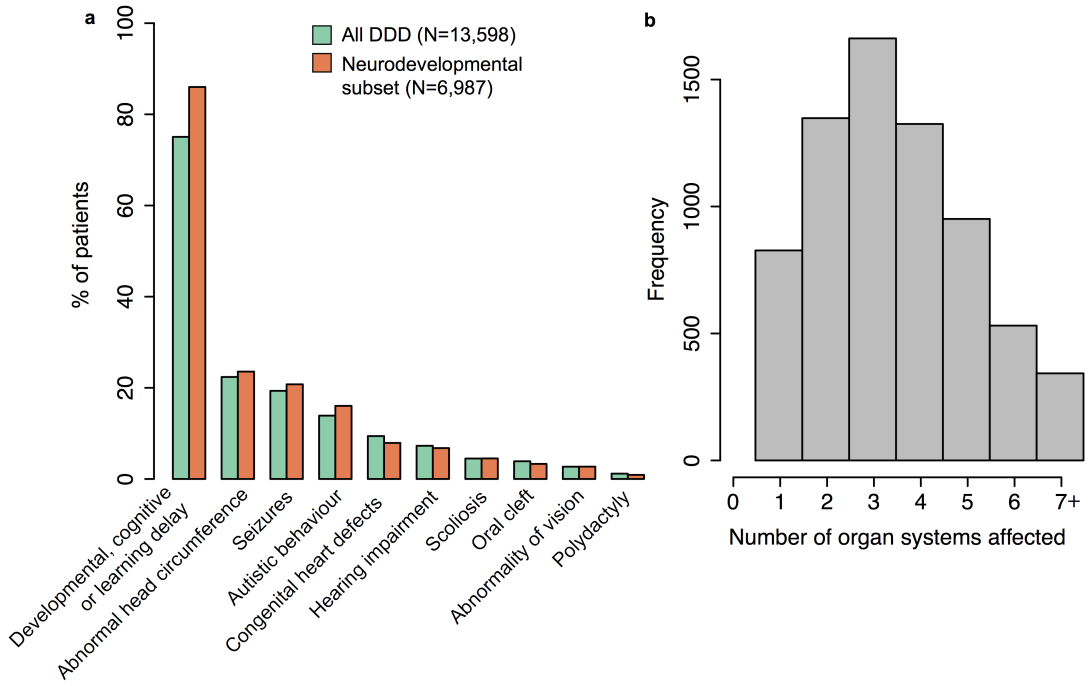


Figure 2.6: Patients recruited to the DDD study have diverse phenotypes. **a**. Examples of specific phenotypes affecting different organ systems, observed in the full DDD cohort and the neurodevelopmental subset of patients. These phenotypes were determined to be clinically relevant for developmental disorders in a previous publication (Wright et al. 2015). **b**. Distribution of the number of distinct organ systems affected in the set of 6,987 patients with neurodevelopmental abnormalities (Methods).

I carried out the GWAS in 6,987 DDD cases with neurodevelopmental disorders and 9,270 ancestry-matched controls in ∼4,1M genetic variants on chr 1-22, with a MAF≥0.05, genotyped on or imputed from the HumanCoreExome chip. No single variant reached genome-wide significance for association ($p< 5 \times 10^{-8}$) (Figure 2.7 a), which was unsurprising given then phenotypic heterogeneity between patients. In fact, the heterogeneity among cases would have led us to be suspicious had there been any significant hits, as these would likely have arisen due to genotyping error or bias rather than real signal. Despite no significant hits in the GWAS, the quantile-quantile plot of observed P-values versus expected (under assumption of no association), were modestly inflated (λ_{GC}=1.097) across the genome (Figure 2.7 b). This inflation could either originate from residual bias due to e.g. cryptic

relatedness or hidden ancestry, or be evidence of a real polygenic contribution from common variants to neurodevelopmental disorder risk.

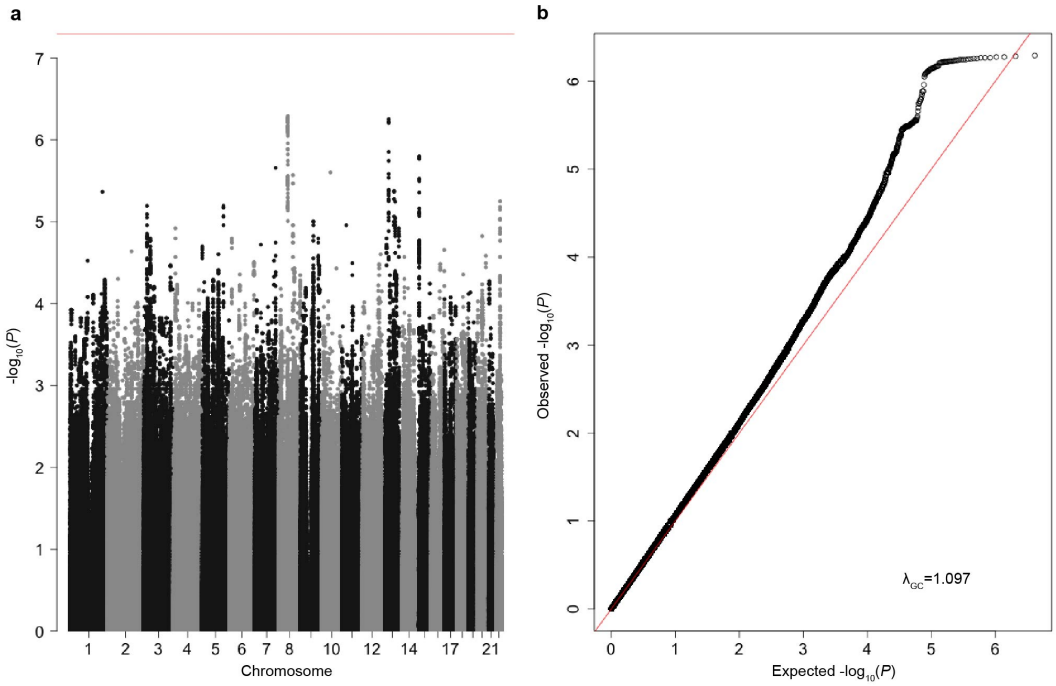


Figure 2.7: Discovery GWAS of neurodevelopmental disorder risk. **a**. Manhattan plot of neurodevelopmental disorder discovery GWAS, with 6,987 DDD cases and 9,270 ancestry-matched UKHLS controls (both European ancestry), using 4,134,438 variants MAF≥5% chr1-22. P-values were from a two-tailed chi squared distribution. Red line = threshold for genome-wide significance (P=5 $\times 10^{-8}$). **b**. Quantile-quantile plot of neurodevelopmental disorder discovery GWAS. Red line = expected values under the null.

2.5.2 Estimating SNP heritability

To investigate whether the inflation in test statistics in the discovery GWAS was due to confounding or real polygenic effects, a natural progression of analyses was to look for evidence of SNP heritability. When put into context with the traditional view that developmental disorders are monogenic conditions, and therefore the patients' phenotype is explained solely by rare variants or environmental factors, we would expect the SNP heritability for neurodevelopmental disorder risk not to

significantly depart from zero. However, the LDSC analysis showed that common variant (MAF≥0.05) heritability was 0.077 (SE=0.021, 95% confidence interval (CI): [0.036, 0.118]), when assuming an overall 1% population prevalence of severe neurodevelopmental disorders (observed scale h^2=0.138, SE=0.037, 95%CI: [0.066, 0.211]).

Strikingly, this is similar to what has been reported for major depressive disorder (MDD) (h^2=0.089, SE=0.004, assuming lifetime population risk of 15%) (Wray et al., 2018) and autism spectrum disorders (h^2=0.118, SE=0.01, population prevalence 1.2%) (Grove et al., 2017). Both these studies were carried out with much larger numbers of cases than our study with 130,664 MDD cases and 18,381 autism cases. However, our h^2 estimate was substantially lower than what has been recently reported for some other neuropsychiatric traits such as schizophrenia (h^2=0.244, SE=0.007, population prevalence 1%) (Pardiñas et al., 2018) and attention deficit hyperactivity disorder (ADHD) (h^2=0.216, SE=0.014, population prevalence 5%) (Demontis et al., 2017). The significant SNP h^2 finding for neurodevelopmental disorders (NDDs) directly contradict the monogenic view for these disorders, and warrants further analysis into understanding how common variants are playing a modifying role in disease liability and how they affect the presentation of clinical symptoms. I return to this in Chapter 3. Additionally, LDSC determined that 66% (SE=11.5%) of the variance observed in the GWAS were due to true polygenic effects. This is lower than what has been reported for other traits by studies that used the same software, but those traits are known to be polygenic and therefore likely also have a cleaner phenotype that they are measuring.

2.5.3 Replication in DDD trios

Having shown that a significant contribution to neurodevelopmental disorder risk in our discovery case-control cohort comes from common genetic variants, I then sought to replicate the findings in an independent dataset. For this, I used data for trios who were also recruited as part of the DDD Study, but who not related to the individuals in the discovery GWAS. This cohort of around one thousand trios had been genotyped on a significantly denser chip (Illumina HumanOmniExpress)

than the larger DDD singletons cohort used for the discovery GWAS (Illumina HumanCoreExome). These two chips had an overlap of only ∼100,000 common variants. Imputation from such a small number of shared variants would have likely have resulted in poor quality data, and I therefore treated the trios data separately. I initially attempted to increase power for the GWAS discovery phase by meta-analysing association results from the trios with the discovery GWAS (I describe these analyses in more detail in section 2.5.4). However, due to issues with the data quality and sample size of the trios dataset, I eventually employed an independent replication approach, instead of adding them to the GWAS.

In order to replicate my finding of a contribution of common variation to neurodevelopmental disorder risk, I wanted to assess whether the effect alleles (and thus their conferred risk) from the discovery neurodevelopmental disorder GWAS were over-transmitted from parents to the affected DDD children. For this approach, I used the polygenic transmission disequilibrium test (pTDT), developed by Weiner et al. (2017). Specifically for our purpose, the benefits of this method included the fact that if any effects in the discovery GWAS were driven by residual bias instead of real differences between cases and controls with respect to NDD risk, we would not expect to see these same effects over-transmitting in a family-based design. This residual bias was not a concern for h^2 estimation using LDSC in the discovery GWAS, but polygenic scores, even when controlling for population structure, are more susceptible to this type of error. In addition, by using a test only within the trios, this eliminates genotyping chip biases.

Having constructed the polygenic risk scores using 71,356 variants, I performed the pTDT for neurodevelopmental disorder risk in 728 European ancestry trios from the DDD Study. I found that parents were over-transmitting neurodevelopmental disorder risk-increasing alleles to the affected child (P=0.0035, one-tailed t-test), replicating the finding of significant polygenic contribution to severe neurodevelopmental disorders.

2.5.4 Pitfalls and lessons learnt from the DDD GWAS

Discovery GWAS

Initially when carrying out the discovery GWAS, I had included in the analyses another control dataset, a Dupuytren's contracture cohort (EGAS00001001206) (Ng et al., 2017). These 4,201 individuals suffered from Dupuytren's disease, which is a common, heritable, late-onset connective tissue disease that causes contracture of digits. The cohort was collected as part of the British Society for Surgery of the Hand Genetics of Dupuytren's Disease consortium, and samples were genotyped on the Illumina HumanCoreExome chip. We initially reasoned that these individuals would represent a random sample of the population with respect to their distribution of risk alleles for complex traits and diseases other than Dupuytren's hand contracture phenotype, so this cohort could be used as extra controls for our neurodevelopmental disorder risk study.

I therefore initially carried out a developmental disorder GWAS (before the decision to refine the phenotype to neurodevelopmental disorders) combining the Dupuytren's cohort with controls from the UKHLS. Quite surprisingly, this analysis showed some near genome-wide significant loci. Upon further inspection, I realised these loci were among the significant loci from the Dupuytren's phenotype GWAS (Ng et al., 2017). This served as a lesson to us that including a cohort, even if as controls, that had been ascertained for a specific phenotype not related to the one we were interested in, could introduce biases into our results. Even though in the event that there had been significant loci associated with neurodevelopment in our GWAS, the downstream analyses paths that we wanted to take would have been affected: e.g. when looking for genetic overlap between neurodevelopmental disorder risk and other published traits (Chapter 3), we would have run into problems trying to decipher whether the correlations would were driven by genetic architecture of neurodevelopmental disorders or an unrelated complex trait for which the controls had been ascertained. In order to avoid any more less obvious biases that could have arisen from the ascertainment of the Dupuytren's cohort, we decided to exclude the cohort from the neurodevelopmental disorder GWAS. This

analysis served as a cautionary tale about the how sample ascertainment can induce spurious genetic correlations, an important lesson that I return to in Chapter 3.5.6.

Trios data

As mentioned above, I initially attempted to utilize the trios data for boosting the discovery GWAS instead of using them for independent replication, before eventually opting for the pTDT method and independent replication. Here, I describe that analysis and discuss some lessons learnt from this.

I attempted to use 911 probands from the trios set, combined with healthy controls, to perform a GWAS and meta-analyse with the larger discovery GWAS. Because I was not able to find suitable controls genotyped on the HumanOmniExpress chip, I instead used 4,612 controls from the Wellcome Trust Case Control Consortium 2 (WTCCC2) project. These individuals had been genotyped on a combination of Illumina 1.2M and Affymetrix500 chips.

In this smaller GWAS using DDD probands (from the trios dataset) and WTCCC2 controls, I found that, despite extensive data quality control, multiple variants were associated with NDD risk at genome-wide significance. It seemed highly likely that these were spurious due to the fact that the better-powered larger GWAS had not detected any significant associations. The cause for some of these associations turned out to be genotyping error, and I subsequently removed these variants. However, many associations were not obviously due to error. The most likely explanation was that these were a result of chip biases. Due to the unreliability of the results, we decided to explore other options for using the extra DDD samples. Eventually, we opted for using the pTDT method and polygenic scores constructed from neurodevelopmental disorder discovery GWAS summary statistics to test for over-transmission of these effect alleles in the independent trios.

2.6 Discussion

In this chapter I have shown that there is a significant contribution from common genetic variants to severe, rare neurodevelopmental disorders. The SNP heritability of these disorders is 7.7% on the liability scale, when assuming a population prevalence of 1%. I also show that alleles increasing risk for these NDDs are over-transmitted from parents to affected children in an independent cohort of DDD Study trios. Patients were ascertained to participate in the DDD Study to be exome sequenced because the clinical geneticists who assessed them believed that their disorder was likely monogenic. Therefore this study represents one of the first GWAS of a large scale heterogeneous cohort of disorders that match the phenotypic profile of monogenic disease. The SNP heritability estimate for NDDs is similar to what has been estimated e.g. for major depressive disorder (Wray et al., 2018) and autism spectrum disorder (Grove et al., 2017).

One of the limitations of my work on neurodevelopmental disorder risk in this thesis, is that I have not included variants with MAF<0.05. The decision for a MAF cutoff at 0.05 was taken for two main reasons. Firstly, the purpose of the study was to assess whether truly common variants contribute to developmental disorders, as has been shown for other related brain disorders such as autism (Grove et al., 2017) and schizophrenia (Loh et al., 2015a). In addition, including only variants with a higher MAF, particularly when cases and controls are genotyped on the same chip, reduces the number of false positives. Ultimately, the finding of significant SNP heritability in the common variant range is important for shaping our understanding of the genetic architecture of neurodevelopmental disorders, since specifically rare variants have previously thought to be the sole genetic contributors to these disorders.

Although assessing the lower frequency MAF ranges, e.g. MAF=0.005-0.05, was out of scope of this project, we can expect there to be at least some SNP heritability to be discovered there. Martin Kelemen, a PhD student in our group, performed some investigative analyses into this lower frequency variant space. His analyses using methods other than LDSC suggested that potentially much more SNP heritability can be found, particularly in the MAF=0.0001-0.005 range. Though at these very

low minor allele frequencies, analyses are easily prone to bias. However, a good example of a report of a large contribution to a trait from low frequency variants comes from a recently published paper on intelligence. Intelligence is a trait that is likely under purifying selection, and therefore variants with large effect sizes are removed from the population. Most of the SNP heritability for intelligence had previously been explained by common variants with smaller effect sizes. Hill et al. (2018), however, showed that including imputed variants down to MAF=0.001-0.01 almost doubled the variance explained. This brought their h^2 estimate to around 0.50, which is in the range of heritability estimates from family studies of intelligence (though the confidence intervals for this estimate were wide). Analyses looking more deeply into including low frequency variants for neurodevelopmental disorders would be something that can be considered in the future. However, this would require careful consideration of potential caveats relating to estimating SNP heritability attributable to low frequency variants. These include reports that even subtle population stratification between cases and controls can lead to biased estimates when dealing with lower frequency variants (Bhatia et al., 2016). Delving into even lower MAF ranges, other members of the DDD Study team are currently exploring an oligogenic model using inherited low frequency variants, that may individually have moderate effects on the phenotypes we observe in the cohort

A critical caveat of our study, which could downward bias our discovery GWAS h^2 estimate, is that the analysis was performed on a very sparse genotyping chip. This can lead to incomplete tagging of common variants, therefore affecting h^2. Additionally, I applied a stringent imputation quality cutoff before analysis. Another general consideration for SNP heritability analyses in case-control studies is that SNP heritability can be underestimated if the controls are not screened for the disease (Peyrot et al., 2016). This is because affected individuals may be included as controls. However, in the case of NDD risk, this is unlikely, since individuals who participated in the UK Household Longitudinal Study (i.e. my GWAS controls) were likely not suffering from neurodevelopmental disorders of the severity that the DDD patients have. However, since I do not have data on rare variants from the controls it is possible that some individuals may be carriers of deleterious variants in neurodevelopmental disorder genes, but the individual

is on the milder spectrum of cognitive or neurodevelopmental phenotypes. These individuals could potentially end up in a population survey study, however this would likely be a small proportion of the cohort.

In this chapter I have also discussed how sample size and selecting for a cleaner phenotype among all cases can increase power for association, and consequently for estimating h^2. Finding additional cohorts similar to the DDD Study to boost our sample size would be challenging. There are several datasets that have been collected for traits such as intellectual disability, e.g. the Northern Finland Intellectual Disability study (Kurki et al., 2018) and a cohort in Nijmegen, Netherlands. However, sample ascertainment is a needs to be considered when combining datasets, as the particular phenotypes and genetic architectures between DDD Study and other cohorts may be somewhat different. Although a large proportion of DDD patients (∼70%) suffer from intellectual disability or developmental delay, the majority are also affected in organ systems other than the nervous system. Therefore the genetic architecture of this cohort may be different to e.g. a cohort ascertained for non-syndromic intellectual disability. Intellectual disability as a trait also has a phenotypic range from mild to profound intellectual disability, and it is thought the extremes of phenotype have somewhat different genetic underpinnings (Reichenberg et al., 2016). Despite these notions, combining the DDD Study with other intellectual disability cohorts would likely boost power for association testing.

Another limitation of the work presented in this thesis is the exclusion of chromosome X from the GWAS. There is a known enrichment of developmental disorder associated genes on chromosome X, with almost over 20% of known monoallelic developmental disorder genes being found on this chromosome (Firth et al., 2009). The decision to drop chromosome X from the analyses came in two parts. Firstly, including it in a GWAS would require additional quality control steps and using a different model for association, to account for the the fact that females have two copies and males only one. The pseudoautosomal regions would need to be removed or treated separately. At the time, we BOLT-LMM did allow for including chromosome X in the data, however it did not (from my understanding) treat it in a different way to autosomes. The BOLT-LMM team have recently released a software update, which now allows for more specialist treatment of the chromosome X.

Therefore including chromosome X using BOLT-LMM, or applying other software e.g. PLINK logistic regression models, would be worth looking into in the future. Secondly, downstream analyses from the GWAS mainly require autosome data only, as LDSC estimates h^2 from only autosomes. Polygenic scores for example also rely on published GWAS, which typically exclude the X. In Chapter 3, I also describe analyses which utilise data from other published GWAS, and these typically once again use data from autosomes only.

In this Chapter, I have also discussed different caveats of GWAS data in trios. We had data for a small cohort of DDD families, which we wanted to utilise to our best ability. Having attempted replication GWAS in a small cohort through both case-control GWAS and family-based GWAS (TDT), I proceeded with the polygenic transmission disequilibrium test, which uses polygenic scores instead of genotypes for the test of transmission. Whilst the more conventional way to utilise the data would have been to meta-analyse with the larger discovery GWAS, the challenges that arose during this process were good examples of the types of considerations that need to be made when planning a GWAS study. Had the trios been genotyped on the same HumanCoreExome chip, which is also cheaper than the HumanOmniExpress chip, it would have been easier to meta-analyse TDT results with the NDD discovery GWAS. Had this been the case, we would not necessarily have considered using the pTDT method. In hindsight, pTDT was perhaps an even more useful tool for us, as it provided the opportunity to perform replication, whilst a meta-analysis would not have resulted in a large increase in power for association, with an addition of only $\sim$700 cases to the GWAS. These exploratory analyses show a good example of how GWAS data can be used in multitude of ways.

Finally, an obvious limitation of the analyses presented here, and also in the following chapters, is that they focus on individuals of European ancestry. Populations not only differ in LD structure, but also the causal variants may be different. Therefore, we cannot make generalisations about the genetic architecture of neurodevelopmental disorders in populations with other ancestries from our results. Although the DDD cohort includes other ancestries, particularly South Asian ancestry, we were not able to find suitable controls for these cohorts, and we did

not have genotype data for many trios of non-European ancestry. This a general issue in the field, and since most published GWAS to date are in Europeans only, the downstream analyses in Chapter 3 would not have been possible to carry out in non-European DDD patients. Hopefully in the future many more studies on non-Europeans will be carried out.

Chapter 3

Investigating shared genetic architecture and polygenic substructure in severe neurodevelopmental disorders

3.1 Chapter overview

In this chapter, I aim to further understand the observed common variant effects contributing to risk of rare severe neurodevelopmental disorders. I first attempt to partition the overall SNP heritability for these disorders into categories of variants that have specific functional roles, or which are within regions of the genome that are expressed in different tissues. By doing this I hope to learn about the biology underlying the signal I found. After this, I compare the neurodevelopmental disorder GWAS results to other published GWAS for a variety of traits. I do this to look for genetic overlap, to learn more about shared underlying biology between our GWAS and other traits. Finally, I employ polygenic scores to look for differences between patient groups within the DDD cohort. These analyses

may tell us whether the polygenic burden is concentrated in patients with specific phenotypes or molecular aetiologies to their disease.

3.2 Background

In Chapter 2, I showed that there is a significant contribution from common genetic variants to severe rare neurodevelopmental disorders. A typical GWAS study would focus downstream analyses on deciphering which variants in the discovered trait-associated regions are more likely to be causal for the association, and honing in on which genes and biological pathways are affecting the trait. The aim of this process is usually to understand the basic biology of the trait, and to hopefully find potential new candidate drug targets for treatment of diseases. With our neurodevelopmental disorder (NDD) risk GWAS, however, I was not able to go down the path of finding candidate causal variants. This is because we had no genome-wide significant variant associations with the phenotype. Instead, in this chapter I explore different avenues to learn more about polygenic effects contributing to NDD risk. Much of this involves utilising information from other already published work in the field.

One of these possibilities is to partition the SNP heritability for NDD risk. This analysis is to find out whether variants expressed in specific tissues are enriched for common variant effects in our GWAS, or whether particular functional element classes are disproportionately responsible for any of this polygenic burden. In addition, by comparing NDD risk to common variant architectures of other GWAS'd traits, we can potentially gain insight into whether the SNP heritability for NDD risk is capturing effects previously associated with brain or neurodevelopment. Finally, utilising the phenotypic data available for the DDD cohort, we can ask the question whether the polygenic effects we are observing contribute more to particular patient groups, or whether these are distributed equally among all patients.

One way to investigate these questions is to use polygenic scores. DDD patients and UKHLS controls have been ascertained for whether or not they have a severe neurodevelopmental disorder phenotype. We can then ask whether they are

significantly different from each other with respect to their allele frequencies for variants associated with previously published GWAS looking at other traits. If they are different, this would be an indication that NDD risk shares effect alleles with the trait in question. This type of approach to look for genetic overlap between traits using polygenic scores was first used by International Schizophrenia Consortium et al. (2009). In this study, the authors showed that patients with bipolar disorder had elevated schizophrenia polygenic scores compared to healthy controls, which indicated shared biology between the two diseases. Importantly, the study also showed that polygenic scores for schizophrenia predicted the disease in an independent cohort, but the prediction was better when applying a higher P-value cutoff and including more variants in the scores.

Nonetheless, polygenic scores still have some caveats. In the field there are no set rules for how to construct polygenic scores, and there are multiple parameters that can be tweaked when deciding which variants to include. This can cause issues when attempting to replicate results and when looking for genetic overlap between diseases and traits, as different studies will use different approaches to defining which variants to use. Often there are multiple published GWAS for the same trait, so the decision on which data to use to construct scores is also an important one. One may choose to use the GWAS with the largest sample size, or potentially a GWAS with a more homogeneous measured phenotype. Another caveat is that polygenic scores can only be reliably constructed in a target population with the same genetic ancestry as the original GWAS. This is because the allele frequencies between populations are different, and subsequently polygenic scores do not necessarily follow an expected normal distribution in a target population with different ancestry (Weiner et al., 2017). Additionally, a polygenic score derived in one population may not capture risk in a second population because linkage disequilibrium patterns differ between them. This means the causal variants may not be tagged in the second population (and indeed, there may be different causal variants). Polygenic scores also typically explain a small proportion of variance in the phenotype. The predictive power of a polygenic score relies heavily on the sample size of the discovery GWAS, and the SNP heritability of both the discovery and target trait. Even if the discovery GWAS is well powered and has high h^2, an

analysis may be underpowered if the h^2 of the target cohort is small, the sample size is small, or the individuals are heterogeneous with respect to their common phenotype.

A more traditional approach to looking for genetic overlap between traits was to utilise family studies, particularly twin studies (Plomin et al., 2008). These genetic correlation approaches involve estimating genetic correlation by comparing cross-twin cross-trait correlations between monozygotic and dizygotic twins. If the traits were more correlated in MZ twins than DZ twins, this was an indication of shared genetic influences between the traits. For example, a study comparing correlations of brain volume and intelligence (IQ) using ∼100 twin pairs, estimated the genetic correlation of these traits to be 0.23-0.30 (Leeuwen et al., 2009).

More recent approaches utilising molecular genetic data have also been developed to look for overall genome-wide shared patterns of genetic effects between pairs of traits. These methods typically use information on hundreds of thousands of variants obtained from GWAS results. One of these methods, termed GCTA, has been quite widely used in studies of genetic correlation. However the downside of this method is that it requires genotype-level data from both studies assessed, which may not always be available. In addition, because the model first builds a relationship matrix for each pair of individuals using the genotype data, the runtime can become substantial when sample sizes increase. A more recent method, which has become very popular for genetic correlation analysis, is bivariate LDSC (Bulik-Sullivan et al., 2015b). This method only requires summary-level data from GWAS, removing issues to do with data sharing, since an individual's genotype cannot be determined from this summary format. This greatly reduces the processing and analysis time, making it easy to perform numerous pairwise analyses. Bivariate LDSC also accounts for sample overlap between studies. To facilitate genetic correlation analyses in the field, the authors have built a web tool named LD Hub (Zheng et al., 2017). Here, a researcher can easily perform heritability analysis on their data and genetic correlation with other traits of interest. The downside of LDSC is that it requires sample sizes in the thousands, and therefore smaller studies tend to use alternative methods such as GCTA.

Bivariate LD score regression method (Bulik-Sullivan et al., 2015b) derives from the univariate LD score regression which is used for SNP heritability analysis (introduced in Chapter 2.4.4). Bivariate LDSC relies on the assumption that for a single SNP, the product of z scores from the two GWAS will, on average, be higher if the traits are genetically correlated than if they are not. This product is then regressed against the LD score (amount of genetic variation tagged) for that SNP. The genetic correlation (r_g) between the two GWAS can then be estimated as a function of the slope of this regression. The advantage of this method over polygenic scores is that as long as each GWAS used has individually controlled for population stratification bias, the resulting r_g analysis should be unaffected by confounders; the method accounts for genetic distance between individuals, whereas polygenic scores may be more biased by e.g. cryptic relatedness. Polygenic score analyses usually include covariates in attempt to correct for stratification.

Bivariate LDSC has greatly advanced our understanding of shared genetic architecture between traits. A landmark paper published in 2015 by authors of the bivariate LDSC method (Bulik-Sullivan et al., 2015b), described results from 276 genetic correlations between 24 different traits using summary statistics from published GWAS. At the time, the studies included in their analyses represented the largest available datasets for the traits, including Rietveld et al. paper on educational attainment which with ∼126k samples had only found three genome-wide significant loci (Rietveld et al., 2013). In comparison, the currently largest study on the trait found 1,271 independent SNPs (Lee et al., 2018). Some other GWAS included in the Bulik-Sullivan et. al paper had no significant loci associated with them. However, using LDSC bivariate analysis, the authors showed that significant overlap of common variant effects could be detected between traits even in the absence of significant loci in the GWAS, illustrating the power of leveraging genome-wide data in these analyses. This paper found shared genetic architecture between traits that had been suspected to have shared causes based on previous epidemiological studies, but also highlighted some unexpected correlations such as the positive genetic correlation between schizophrenia and anorexia. In addition, the study found examples where trait pairs that had been expected to be genetically linked turned out not to be. Since then, more studies have analysed the genetic overlap

between common cognitive, neuropsychiatric (e.g. schizophrenia, bipolar disorder), neurodevelopmental (e.g. autism) and neurological traits (e.g. Alzheimer's and Parkinson's) (Anttila et al., 2017; Okbay et al., 2016).

Whilst genetic correlation analysis is a powerful tool for investigating the overall shared genetic effects between traits, it does not provide information on the internal genetic architecture of a study. In the case of our neurodevelopmental disorder GWAS, we cannot make conclusions about whether the DDD cohort as a whole are all contributing to SNP heritability and shared genetic effects with other traits. To do this, we need to employ other methods such as polygenic risk scores, to assess the burden of genetic variation between subsets of patients determined utilising the phenotypic data we have of the patients. One huge advantage of the DDD Study over many other cohorts for severe neurodevelopmental defects is that the patients have detailed phenotypic data that have been systematically recorded using HPO terms. This essentially reduces heterogeneity between different clinicians' use of terminology, and increases our chances of capturing all individuals with a specific abnormality when text mining the data. In addition, many of patients phenotype record also includes different measurements of growth, and information on when they reached developmental milestones such as age they first walked or talked. Additionally, some records include a free text note written by the clinician, which sometimes includes more detail about e.g. severity of the phenotypic abnormality which may not have been logged with the HPO terms. In the context of this thesis, I can leverage these data to understand more about whether effects from common variants are more important in particular patient groups who are phenotypically more similar to each other.

3.3 Contributions and publication note

Elizabeth Radford provided data on developmental milestones in the general population. Wendy Jones provided useful conversations about the clinical genetics assessment of developmental delay. The Australian data were collected by Sui Yu, Jozef Gecz Nicholas Martin, and the raw data were prepared by Kerrie McAloney

and Scott Gordon. Hilary Martin performed the quality control and imputation of Australian datasets, helped generate PCA plots for these, performed the AVENGEME power calculation in Australians, and contributed to supervising this work. The work described in this chapter were completed under the supervision of Jeffrey Barrett, and the key findings were published in Niemi et al. (2018).

3.4 Methods

3.4.1 Partitioned heritability

I used partitioned LDSC (Finucane et al., 2015) software to look for neurodevelopmental disorder (NDD) SNP heritability enrichment using the baseline model LD scores and regression weights available online. The method captures the proportion of variation in the phenotype that is explained by a pre-defined subset of variants in the genome. If a category of variants are enriched for heritability for that trait, then SNPs that are in high LD with variants in that category will have increased χ^2 test statistics compared to if the SNPs were in high LD with a category that is not enriched for heritability (Finucane et al., 2015). For cell type groups and functional categories I set the significance threshold to P<0.005 (0.05/10 tests) and P< 9.2×10^{-4} (0.05/54 tests), respectively.

The method also allows for partitioned heritability analysis of custom regions of interest in the genome. For NDD risk, I was particularly interested in two custom sets of variants. The first set of variants were those within the boundaries of genes that are known to cause developmental disorders, namely the Developmental Disorders Genotype-Phenotype Database (DDG2P) genes. These are a set of 2,044 genes that have been curated by clinicians in the DDD Study, and confirmed or presumed probably causal for developmental disorders (Firth et al., 2009). The second group of genes of interest were the highly constrained genes (Lek et al., 2016). These are genes where loss-of-function mutations are depleted from the expected numbers in a large exome sequence database of relatively healthy individuals in the ExAC consortium. To more accurately estimate the expected observations of rare variants in genes, this model incorporates information about the gene's length and the sequence-context based mutation rates under minimal selection. The authors (Lek et al., 2016) describe this metric as a probability for being loss-of-function intolerant (pLI). I analysed the partitioned heritability in the set of 3,230 genes with high evidence (pLI≥0.9) for selective constraint. Selecting for high pLI scores captures most of the severe haploinsufficiency genes that are known to cause human disease (Lek et al., 2016), out of those that are sufficiently large to estimate

constraint for. For partitioned heritability in variants within DDG2P genes, I added 35kb upstream and 10kb downstream of the gene boundaries, following the example of Grove et al. (2017). I then annotated each variant used in the LDSC baseline model according to whether it was within any of the DDG2P boundaries or not. I then calculated new LD scores for variants using LDSC, and ran the partitioned heritability analysis following instructions on github (Finucane et al., 2015). I repeated the same steps for a list of highly constrained genes (pLI≥0.9).

3.4.2 Genetic correlation

For the traits that were available through LD Hub, I used the online server, and for others I downloaded the LDSC software from github and ran the analyses on the command line. The 19 traits tested included cognitive performance, education, psychiatric traits and diseases, anthropometric traits and non-brain related traits and diseases. I set the significance threshold to $p<0.0026$ (0.05/19 tests).

3.4.3 Australian replication cohort

Datasets and patient phenotypes

To replicate findings from the genetic correlation analyses, we collaborated with a group from Australia, who provided us data for Australian cases with neurodevelopmental disorders and ancestry-matched population controls. The majority of the patients (>95%) were under 18 years old when recruited. They were originally genotyped as part of routine clinical care to ascertain pathogenic copy number variants; 50-60% were recruited through clinical genetics units, and the rest through neurologists, neonatologists, paediatricians and cardiologists. Our Australian collaborators reviewed information on the request forms, and found that the majority of patients had developmental delay/intellectual disability and malformations involving at least one organ (e.g. brain, heart, and kidney). 15-20% were recruited as neonates with multiple malformations involving brain, heart and/or other organs, and were too young to be diagnosed with developmental delay/intellectual disability. The

population-matched controls came from the Brisbane Longitudinal Twin Study (Queensland Institute of Medical Research, (Wright and Martin, 2004; Mina-Vargas et al., 2017)).

Quality control and imputation

Both case and control data were on GRCh37, and detailed information of genotyping chips is shown in Table 3.1. Sample and variant quality control was performed following steps described in (Chapter 2.4.1). Patients without a neurodevelopmental phenotype were removed prior to data quality control, therefore all cases quoted in the table have the relevant phenotype (compared to DDD cohort quality control steps in Table 2.1). Rare variants MAF$\leq$0.005 were removed before phasing and imputation. The samples were then phased and imputed in a single batch, using SNPs that intersected between the CytoSNP-850K chip (cases) and the Illumina 610K chip (controls). The Sanger Institute Imputation Service (McCarthy et al., 2016) was used to carry out phasing and imputation, using the same software Eagle2 (v2.0.5)(Loh et al., 2016) and PBWT (Durbin, 2014), and the Haplotype Reference Consortium as the reference panel (release 1.1, chr1-22, X)(McCarthy et al., 2016) as I had used for DDD and UKHLS (Chapter 2). Samples of European ancestry were then selected by defining a cluster around the 1000 Genomes Great British (GBR) Phase 3 samples in a projection PCA (Figure 3.1).

Table 3.1: Quality control for Australian datasets.

		Quality control steps - Australian data	Australian ID cases	Australian (BLTS) controls
DNA chip		NA	CytoSNP-850K	Illumina 610K
Pre QC	Samples	NA	2,283	4,274
	Variants	NA	854,413	526,217
Post sample and variant QC	Samples	samples that passed QC; one individual from related pairs and non-GBR samples removed	1,270	1,688
	Variants	variants that passed QC and had MAF≥0.5%; intersection of CytoSNP-850K and Illumina 610K SNPs	282,595	282,595
Post imputation, neurodevelopmental GBR subset	Samples	NA	1,270	1,688
	Variants	imputed variants filtered for INFO≥0.9	4,636,561	4,636,561

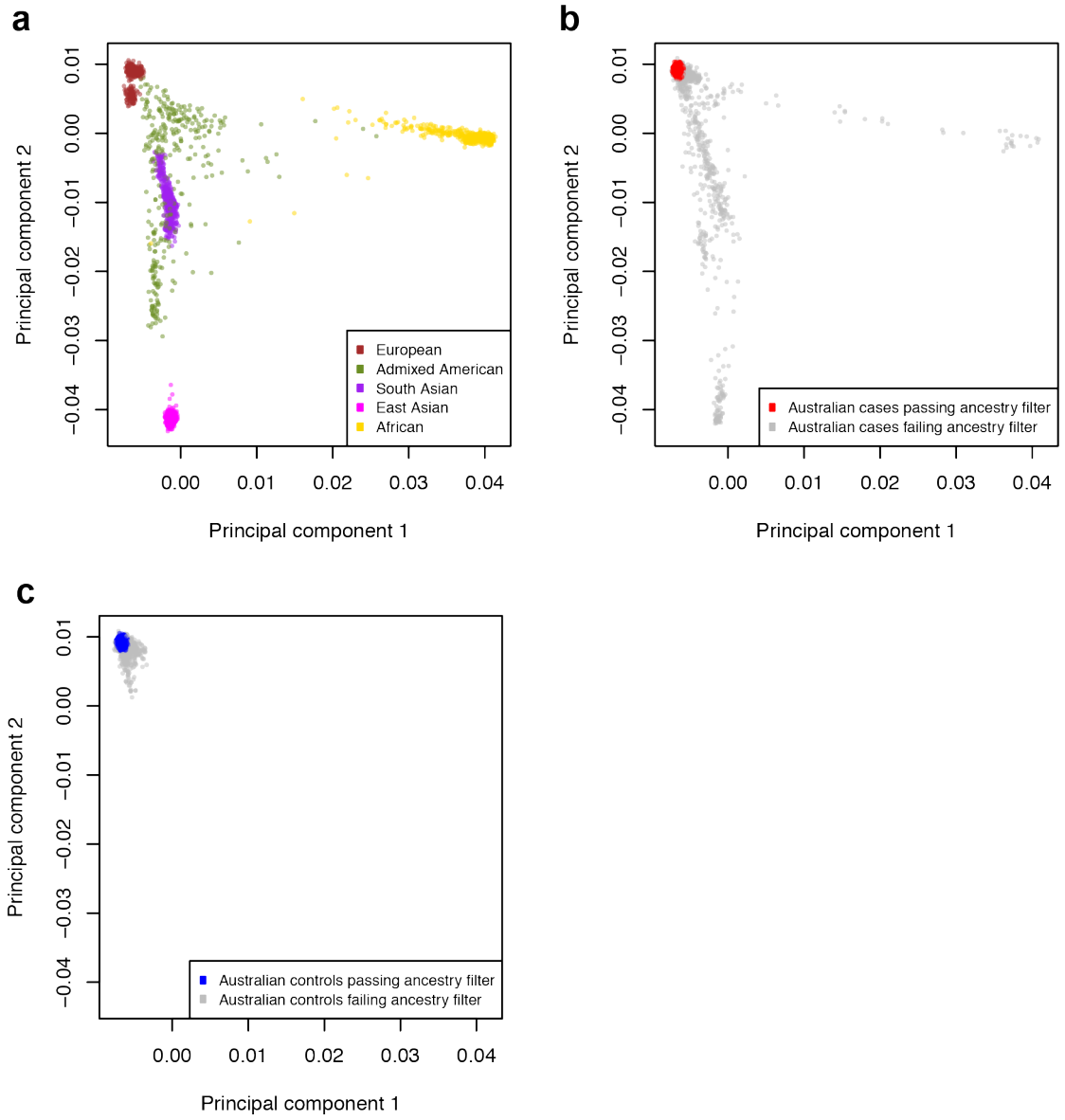


Figure 3.1: Ancestry principal components analysis of Australian cohorts. **a**. Reference samples from 1000 Genomes Phase 3, colored by the five super populations, used for a projection PCA of Australian cohorts (cases and controls). **b**. All Australian cases (N=2,283) from projection PCA with 1000 Genomes. Case samples with European ancestry are plotted in red and non-Europeans in grey. **c**. All Australian controls (N=4,274) from projection PCA with 1000 Genomes. Control samples with European ancestry are plotted in blue and non-Europeans in grey. All cases and controls coloured in grey (panels b and c) were excluded from analysis due to non-European ancestry.

Polygenic scores in Australians

The principle behind constructing polygenic scores is described in chapter 2.4.5. I used P<1 threshold for choosing SNPs for NDD risk scores in Australian analyses. I also constructed scores for seven published GWAS (educational attainment (Lee et al., 2018), intelligence (Sniekers et al., 2017), schizophrenia (Pardiñas et al., 2018), autism (Grove et al., 2017), intracranial volume (Adams et al., 2016), height (Wood et al., 2014) and birth weight (Horikoshi et al., 2016)). Some of these traits were correlated with NDD risk (educational attainment, intelligence, schizophrenia), whilst e.g. autism is prevalent in NDD patients, and intracranial volume is correlated with head circumference that is often abnormally small or large in DDD patients (micro- or macrocephaly, see Figure 2.6). Again, for all traits, I included only variants that had a MAF≥0.05 and that were directly genotyped or imputed with high confidence (INFO≥0.9) in the Australian sample. As P-value thresholds for published GWAS, I used the threshold that had been found to explain the most variation in the most recent available published studies for the trait (educational attainment P<1 (Okbay et al., 2016), intelligence(Sniekers et al., 2017), schizophrenia P<0.05(Pardiñas et al., 2018) and autism P<0.1 (Weiner et al., 2017)). Note that for educational attainment and autism, the paper cited for the P-value threshold is different than that of the summary statistics used for the trait because at the time of analysis we had obtained the summary statistics through personal communication without access to the manuscript associated with the data as these were yet unpublished. For traits which we had phenotype data for in the DDD, I used thresholds that explained the most variation in DDD cases using linear regression and R-squared: P<1 for intracranial volume, P<0.01 for birth weight and P<0.005 for height. Thresholds and the number of SNPs used for each score are shown in Tables 3.2. All scores were normalised to a mean of 0 and variance of 1.

The schizophrenia PGC-CLOZUK study (Pardiñas et al., 2018) included some controls from the Brisbane Longitudinal Twin Study that I would be using as controls. If I constructed polygenic scores from these summary statistics this would result biased score differences between the Australian cases and controls. Through

Table 3.2: Summary of polygenic score parameters in Australian cohorts.

Polygenic score	Polygenic score parameters		
	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score
Educational attainment	0.1	1	92,092
Height	0.1	0.005	9,809
Intelligence	0.1	0.05	21,551
Schizophrenia (QIMR removed)	0.1	0.05	23,878
Intracranial volume	0.1	1	90,928
Autism	0.1	0.1	26,846
Birth weight	0.1	0.01	6,828
Developmental disorder risk (discovery GWAS)	0.1	1	67,001

personal communication, Antonio Pardias reran the schizophrenia GWAS having removed the Brisbane Longitudinal Twin Study from PGC-CLOZUK data, and I used these summary statistics instead of the published ones.

To test for differences in scores between cases and controls, I used R (version 1.90b3) to perform logistic a regression, including the first ten principal components from the ancestry PCA as covariates to control for potential population stratification. We used AVENGEME (Palla and Dudbridge, 2015) to calculate power to find significant association, assuming that the SNP heritability was the same (h^2=0.077) in both the Australian and British cohorts, and that the genetic correlation between them was 1.

3.4.4 Subsetting the DDD Study patients

By diagnostic variant

We wanted to investigate whether DDD patients with diagnostic rare variants were different from individuals with no diagnostic variants, with respect to their polygenic burden. Identification of clinically relevant rare variants from the exome data was performed by the DDD exome analysis team. This process was based on the clinical filtering procedure described in (Wright et al., 2015), which focuses on identifying

rare, damaging variants in a set of genes known to cause developmental disorders (https://www.ebi.ac.uk/gene2phenotype/), that fit an appropriate inheritance mode. Briefly, variants that pass clinical filtering are uploaded to DECIPHER (Firth et al., 2009), where the patients' clinicians classify them as definitely pathogenic, likely pathogenic, uncertain, likely benign or benign. This process of clinical classification is necessarily dynamic as new disorders are identified and patients manifest new phenotypes.

Our diagnosed set of 1,127 patients fulfilled one of these criteria: a) they were amongst the diagnosed set in a recent reanalysis of the first 1,133 trios (Wright et al., 2018b), or b) had at least one variant (or pair of compound heterozygous variants) rated as definitely pathogenic or likely pathogenic by a clinician, or c) had at least one variant (or pair of compound heterozygous variants) in a class with a high positive predictive value that passed clinical filtering but had not yet been rated by clinicians. *De novo* or compound heterozygous loss-of-function (LoF) variants were considered to have high positive predictive value, since of the ones that had been rated by clinicians, 100% of compound heterozygous LoFs and 94.% of *de novo* LoFs had been classed as definitely or likely pathogenic. My undiagnosed set consists of 2,479 patients who had no variants that passed the clinical filtering, or in whom the variants that had passed clinical filtering had all been rated as likely benign or benign by clinicians, or who were amongst the undiagnosed set in the first 1,133 trios that have previously been extensively clinically reviewed (Wright et al., 2015). Note here, that my diagnosed versus undiagnosed analysis shown excludes 3,375 patients who had one or more variants that passed clinical filtering in a class with a relatively low positive predictive value, but who have not yet been rated by clinicians.

By severity of intellectual disability or developmental delay

I defined patients as having mild intellectual disability or delay if their HPO phenotypes included borderline, mild or moderate intellectual disability (HP:0006889, HP:0001256, HP:0002342) and/or mild or moderate global developmental delay (HP:0011342, HP:0011343). Patients were included in the severe ID or delay set

if they had severe or profound intellectual disability (HP:0010864, HP:0002187) and/or severe or profound global developmental delay (HP:0011344, HP:0012736). I excluded patients with ID or global developmental delay of undefined severity. A comparison of polygenic scores between all three categories of severity (mild, moderate, severe) would have been possible, but my concern was that the power for this analysis was too low due to small sample numbers. Instead, we decided to group mild and moderate DD/ID patients together as mild.

3.4.5 Polygenic scores in DDD patients

I constructed polygenic scores for educational attainment (Lee et al., 2018), intelligence (Sniekers et al., 2017), schizophrenia (Pardiñas et al., 2018), autism (Grove et al., 2017), intracranial volume (Adams et al., 2016), height (Wood et al., 2014) and birth weight (Horikoshi et al., 2016) in the 6,987 DDD patients (Table 3.3), the same way as described for the Australian cases and controls. I then performed a linear or logistic regression in R of the phenotype against each polygenic score, including 10 PCs from the ancestry PCA as covariates, with threshold $P<0.007$ for significance ($P<0.05/7$ correcting for seven polygenic scores).

Table 3.3: Summary of parameters used to construct polygenic scores for DDD patients cohort (European ancestry, N=6,987).

Polygenic score	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score
Educational attainment	0.1	1	79,296
Intelligence	0.1	0.05	19,387
Schizophrenia	0.1	0.05	21,321
Autism	0.1	0.1	23,648
Intracranial volume	0.1	1	76,788
Birth weight	0.1	0.01	6,212
Height	0.1	0.005	9,019

3.4.6 Power for detecting differences in polygenic scores

I assessed power to detect differences in scores between diagnosed and undiagnosed patients, by testing the hypothesis that diagnosed patients were effectively a random sample of population controls with respect to their polygenic profiles. To test this, I randomly sampled 1,127 controls (i.e. the same number as we had diagnosed patients) and compared the polygenic scores between them and the undiagnosed patients (N=2,479) using logistic regression. I repeated this 10,000 times and determined the proportion of iterations where there was a significant difference $P<0.007$ ($P<0.05/7$ correcting for seven polygenic scores) as proxy for power. For educational attainment, this was 99.1% of iterations, 93.6% for schizophrenia, 61.2% for intelligence, 34.8% for height, 2.2% for autism, 0.75% for birth weight and 0.08% for intracranial volume.

3.5 Results

3.5.1 Partitioning neurodevelopmental disorder SNP heritability

In Chapter 2, I described an overall 0.077 (95% CI : [0.036, 0.118]) contribution to NDD risk coming from common genetic variants in the discovery case-control GWAS, calculated using LDSC. As a first approach to tease apart this common variant burden, I used an an extension of the LDSC method by Finucane et al. (2015), termed stratified LDSC. This method can be used to further break down trait SNP heritability into functional genomic categories (e.g. conserved regions, enhancers or histone marks, etc.) or cell type groups (e.g. central nervous system, liver or cardiac, etc.) that are enriched for the heritability observed.

The partitioned LD score regression results showed that SNP heritability for neurodevelopmental disorders was nominally significantly enriched in cells of the central nervous system (P=0.025) (Table Appendix A), and in mammalian constrained regions (Lindblad-Toh et al., 2011) (P=0.009) (Appendix A), consistent with sim-

ilar analyses for other neuropsychiatric and cognitive traits. Neither the highly constrained genes (pLI$\geq$ 0.9) nor the DDG2P genes showed significant enrichment or depletion of h^2.

Although significantly greater than zero, the h^2 for NDD risk is still relatively low and the study has a small sample size compared to other neuropsychiatric disease studies where specific enrichment has been detected (Grove et al., 2017; Pardiñas et al., 2018). The nominally significant finding of CNS variant enrichment in the partitioned heritability analysis supports CNS involvement, but it does not give us much more information about the potential mechanisms behind NDD risk. Similarly, enrichment in mammalian constrained regions would be plausible if real, as we know that genes that are highly conserved are more likely to also be important for normal development. It would be interesting to see whether the heritability enrichment results became stronger if we could increase our power by obtaining more samples, or if further refinement of the NDD phenotype could increase the h^2 estimate and subsequently power for partitioned heritability analysis.

3.5.2 Shared genetic architecture with other traits

To further investigate the genetic architecture of NDD risk, I next looked for overall genetic overlap of common variant effects on this trait with other published GWAS. Due to the increasing amount of evidence for polygenic effect sharing between neuropsychiatric and cognitive phenotypes, we were particularly interested in knowing whether severe neurodevelopmental disorder risk shared effects with common cognitive and neuropsychiatric traits. We decided to investigate genetic correlation (r_g) with autism, which is a neurodevelopmental disorder, educational attainment, which is a proxy phenotype for cognitive performance, and schizophrenia, which, at the start of this project, had one of the largest neuropsychiatric GWAS available. In addition, we were interested in intelligence, which became available later on as this project progressed. In addition, we also decided on a list of good-quality available non-neurodevelopmental GWAS for different types of traits to check for genetic correlation against. Due to the fact that both the DDD patients and UKHLS controls had not been ascertained for the presence or absence

of any complex diseases or traits, our expectation was that the allele frequencies for traits unrelated to neurodevelopment would likely not differ between DDD patients and UKHLS controls (as described in section 2.4.1). Therefore, I would not expect to find that polygenic effects in the NDD discovery GWAS overlapping with those found in GWAS on these traits. These kind of traits can therefore be regarded as negative control GWAS, shown in green in Figure 3.2. An example of such a trait is Crohn's disease which has a later onset in life and does not involve the brain. If we found no evidence for genetic overlap of such traits with NDD risk, this would be an indicator that there are no subtle ascertainment differences between the cases and controls that are affecting the genetic architecture of NDD risk we detect. Other traits that we were interested in included anthropometric traits such as height and birth weight, because developmental disorders often include growth, skeletal system and muscular abnormalities, as illustrated in Figure 2.6.

I carried out genetic correlation of the neurodevelopmental disorder risk discovery GWAS against nineteen published traits, using bivariate LD score regression (Bulik-Sullivan et al., 2015b). NDD risk was significantly negatively correlated with genetic predisposition to higher educational attainment (r_g=-0.49, SE=0.08, P=5.3×10^{-10}) and intelligence (as measured by Spearman's g; see Chapter 4) (r_g=-0.44, SE=0.10, P=2.2×10^{-5}), and positively correlated with genetic risk of schizophrenia (r_g=0.28, SE=0.07, P=2.7×10^{-5}) (Figure 3.2 and Table 3.4). Interestingly, educational attainment and schizophrenia have both been are linked to neurodevelopment (Owen et al., 2011; Noble et al., 2015), however they do not share effects with each other (see discussion below). Although genetic correlation analysis is a powerful way to find genetically related traits, we cannot extrapolate more about which particular effects may be the ones shared between NDD and these traits. None of the anthropometric traits, nor the negative control traits, were significantly genetically correlated with our NDD GWAS after accounting for multiple testing. These results, together with the findings from the partitioned heritability analysis in section 3.5.1, suggest that thousands of common variants have individually small effects on brain development or function, which in turn influences neuropsychiatric disease risk, cognitive traits, and risk for severe neurodevelopmental disorders.

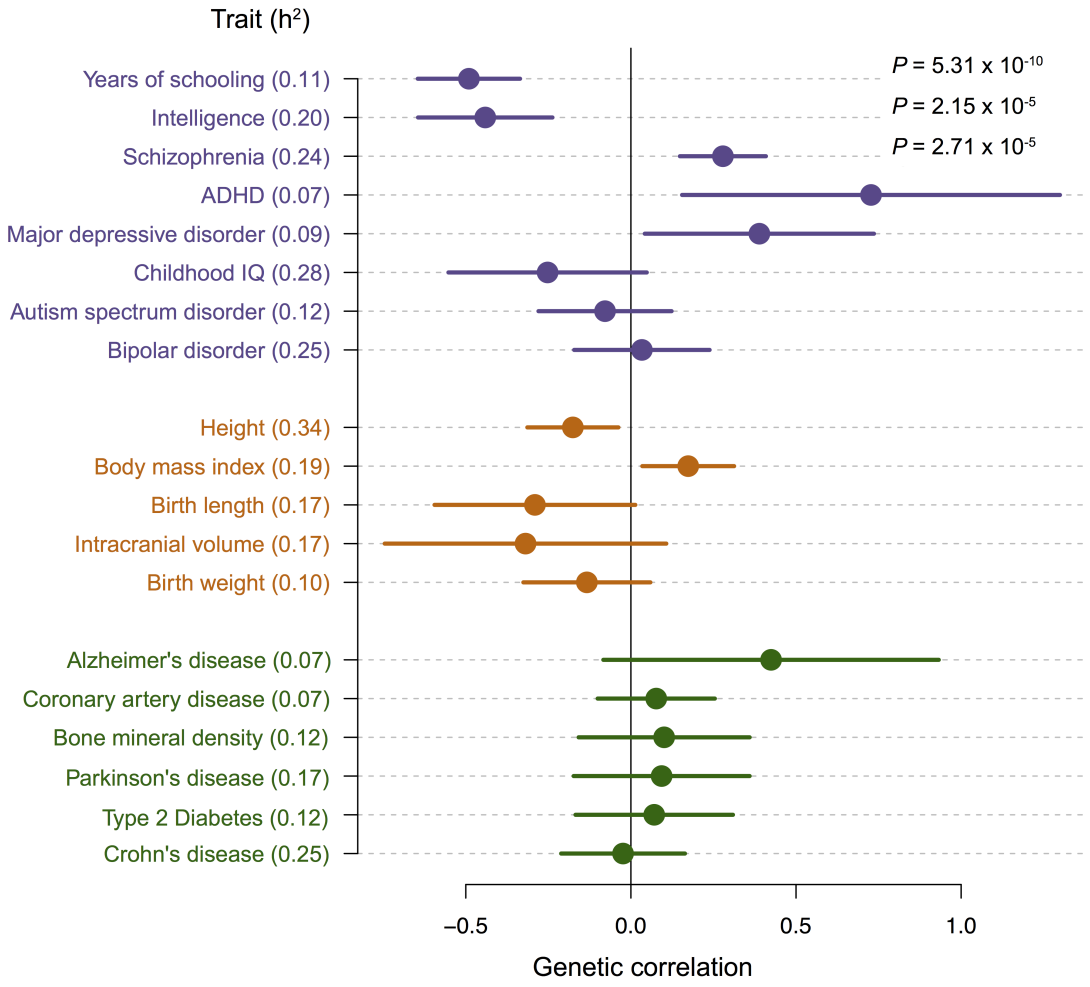


Figure 3.2: Genetic correlations between neurodevelopmental disorder risk (6,987 cases and 9,270 controls) against nineteen other traits. Cognitive or psychiatric (purple), anthropometric (orange) and negative control traits (green) with SNP heritability (h^2) displayed for the trait. SNP heritability for dichotomous traits is displayed on the liability scale. Genetic correlation was calculated using bivariate LD score correlation, with the bars representing 95% confidence intervals (using standard error) before correction for multiple testing. Uncorrected P-values are only shown if they pass Bonferroni correction for 19 traits. Sample sizes for 19 other GWAS are shown in Table 3.4.

Educational attainment is used as a proxy trait for cognitive performance (Rietveld et al., 2014), but it is also a trait that can easily be influenced by other factors than cognitive performance. Therefore, general intelligence, which is measured through cognitive tests, would arguably be a more robust measure of cognitive performance,

Table 3.4: Results from genetic correlation analyses with discovery neurodevelopmental disorder GWAS. Population prevalence for categorical traits was used to calculate trait 2 SNP heritability on the liability scale.

Trait 2	r_g between developmental disorder risk and trait 2	Standard error (SE)	95% confidence interval (SE, lower bound)	95% confidence interval (SE, upper bound)	P-value	h^2 for trait 2 (liability scale)	SE for h^2	Population prevalence
Years of schooling	-0.491	0.079	-0.336	-0.645	5.31×10^{-10}	0.112	0.004	
Intelligence (Spearman's g)	-0.441	0.104	-0.237	-0.645	2.15×10^{-5}	0.203	0.013	
Schizophrenia	0.279	0.066	0.148	0.409	2.71×10^{-5}	0.242	0.008	0.010
ADHD	0.727	0.292	0.155	1.299	0.013	0.071	0.031	
Major depressive disorder	0.389	0.177	0.042	0.736	0.028	0.087	0.017	0.150
Childhood IQ	-0.252	0.153	0.048	-0.553	0.100	0.279	0.051	
Autism spectrum disorder	-0.078	0.103	0.123	-0.28	0.445	0.118	0.010	0.012
Bipolar disorder	0.033	0.105	-0.172	0.238	0.751	0.250	0.023	0.010
Height	-0.176	0.07	-0.038	-0.314	0.012	0.336	0.021	
Body mass index	0.174	0.071	0.035	0.312	0.015	0.189	0.010	
Child birth length	-0.291	0.155	0.013	-0.595	0.061	0.165	0.027	
Intracranial volume	-0.319	0.218	0.107	-0.746	0.142	0.167	0.053	
Birth weight	-0.133	0.098	0.059	-0.326	0.174	0.095	0.008	
Alzheimer's disease	0.424	0.259	-0.083	0.932	0.101	0.068	0.013	0.050
Coronary artery disease	0.077	0.091	-0.101	0.254	0.396	0.070	0.005	0.050
Lumbar Spine bone mineral density	0.101	0.132	-0.158	0.36	0.447	0.116	0.018	
Parkinson's disease	0.093	0.136	-0.173	0.359	0.494	0.167	0.050	0.002
Type 2 Diabetes	0.071	0.122	-0.168	0.309	0.562	0.120	0.012	0.080
Crohn's disease	-0.024	0.096	0.164	-0.211	0.804	0.252	0.027	0.003

and therefore a preferable trait for using in polygenic analyses also in this thesis. However, it is difficult to obtain both intelligence test results and genotype data for large (ideally ancestrally homogeneous) cohorts, whereas it is easier to obtain data on how many years of schooling study individuals obtained. Therefore the sample sizes for educational attainment GWAS are vastly larger (now 1.1M (Lee et al., 2018)) than cognitive GWAS, and as long as this trait correlates well genetically with intelligence, it can be used as a proxy. It is therefore not surprising that NDD risk r_g analysis results with these traits are very similar. However, for downstream

analyses, it may be that educational attainment polygenic scores are more powerful for detecting significant differences between groups of individuals, as the GWAS is much larger than the GWAS for intelligence.

3.5.3 Replication of genetic overlap findings in Australians

Having shown that there is significant polygenic contribution to neurodevelopmental disorder risk which has shared genetic effects with other brain-related phenotypes, the next question was whether these findings were specific to the DDD Study, or would they replicate in other similar cohorts. If the results replicated in a completely independent cohort which had been ascertained for similarly severe neurodevelopmental disorders, this would strengthen our findings of polygenic contribution to developmental disorders.

For this attempt at replication, we obtained data for 1,270 South Australian patients who with neurodevelopmental disorders, and 1,688 population-matched controls. However, the small cohort size for the Australians meant that I was not able to do direct genetic discovery or subsequently genetic correlation analysis, as this requires >5,000 samples (as stated in LDSC github). Instead, I tested whether there was a difference in common variant polygenic scores between cases and controls for a number of traits that I had found to be significantly correlated with NDD risk (Chapter 3.4.3). A significant difference in scores would signify replication of the r_g findings in this smaller cohort. To do this, I calculated polygenic scores using summary statistics from our discovery NDD GWAS, and the publicly available GWAS, including educational attainment (Lee et al., 2018) and intelligence (Sniekers et al., 2017).

The results showed that Australian neurodevelopmental disorder patients had lower polygenic scores for educational attainment and intelligence compared to controls ($P=1.4 \times 10^{-8}$ and $P=7.6 \times 10^{-4}$ respectively) (Table 3.5). I initially observed suspiciously significantly increased in polygenic scores for schizophrenia ($P=1.2 \times 10^{-36}$) in the Australian cases. However, this turned out to be due to the fact that some of the Australian controls were included in the schizophrenia GWAS (Pardiñas et al., 2018). We obtained new summary statistics from the authors, in

Table 3.5: Summary of polygenic score results in Australian cohorts.

Polygenic score	Results*		
	Beta	Standard error	P-value
Educational attainment	-0.216	0.038	1.4×10^{-8}
Height	-0.155	0.040	8.8×10^{-5}
Intelligence	-0.126	0.038	7.6×10^{-4}
Schizophrenia (QIMR removed)	0.092	0.038	0.014
Intracranial volume	-0.078	0.038	0.041
Autism	0.070	0.038	0.063
Birth weight	-0.062	0.038	0.098
Developmental disorder risk (discovery GWAS)	-0.047	0.038	0.212

which the Australian controls had been removed from the GWAS, and I repeated the analysis. Here, Australian cases had nominally significantly higher scores for schizophrenia (P=0.014).

For neurodevelopmental disorder risk, I did not see a significant difference between cases and controls for the scores constructed from our discovery GWAS. We therefore wondered if we had enough power to detect a significant association (at P<0.05) between our polygenic score for neurodevelopmental disorders and case/control status in the Australian dataset. This analysis showed that we should have had 95% power to detect a difference if the two cohorts had identical phenotypes. This suggests that differential phenotypic ascertainment between the British and Australian cohorts diluted our ability to quantify their shared genetics.

The fact that the NDD-risk polygenic score were not significantly different between cases and controls, despite the fact that we have shown that NDD and intelligence are negatively genetically correlated, likely reflects low power to estimate variant effects in our NDD GWAS. Out-of-sample polygenic score prediction is affected both by the discovery sample size, and the total amount of heritability that can be predicted. Our neurodevelopmental disorder risk GWAS found significant but still relatively low heritability (0.077). In comparison, the educational attainment GWAS had a huge sample size of 1.1M, so polygenic scores using these SNP effects will be much better powered despite the trait also having a relatively low SNP

heritability (0.11). If we compare this to an early GWAS study of educational attainment in 9,538 Australian individuals, a similar size to our study, they also failed to explain any significant variation (r^2<0.0023, P≥0.14) in an independent target cohort of 968 individuals (Martin et al., 2011a)

Interestingly, the Australian patients also had lower scores for height than controls (P=8.8×10^{-5}) (Table 3.5). Even though NDD risk GWAS was not significantly genetically correlated with height after multiple testing correction, the direction of effect was in the same direction as in this analysis. A possible explanation for this finding is residual population structure differences between Australian cases and controls: height is well known to correlate with latitude within Europe (Novembre et al., 2008), and the cases are recruited from Adelaide, where there is more Mediterranean (Greek, Italian) ancestry, and controls were recruited in Brisbane, which has more Irish ancestry. It is possible that height scores could differ for this reason if the PC covariates were not sufficient to control for this. This result is, however, potentially interesting, since developmental disorder patients often have growth abnormalities associated with their condition.

3.5.4 Polygenic substructure in DDD patients

Having replicated the NDD risk discovery GWAS results in an independent Australian cohort, I returned to the DDD Study data to answer more specific questions about the distribution of polygenic risk among patients with heterogeneous phenotypes. The DDD cohort is one of the largest severe neurodevelopmental disorder cohorts in the world, and furthermore, the deep phenotyping of patients by clinical geneticists adds a whole new layer of valuable information for studying the genetic architecture of these diseases. Specifically, this phenotypic information, coupled with the data from exome sequencing of DDD trios, allows us to explore whether polygenic burden is more enriched in certain patient subgroups than others. Some of the key questions we were interested in answering included: (1) Are patients who had a diagnostic variant through the exome sequencing project any different from patients for whom we have not identified a likely severe pathogenic mutation? (2) Do common variant effects correlate with severity of the developmental disorder?

(3) Are specific phenotypes correlated with differences in polygenic risk? Here, I investigate these questions using polygenic scores.

Comparing patients with and without diagnostic rare variants

We hypothesized that the so far genetically undiagnosed DDD patients would be contributing to NDD polygenic risk, because in other complex neuropsychiatric traits individuals without causal mutations have overall higher polygenic risk for the trait than healthy controls (International Schizophrenia Consortium et al., 2009). In contrast, for the DDD patients who had a rare diagnostic variant, we thought the polygenic contribution in these individuals could be slightly different. As an example from another trait, a study on hypercholesterolemia (Talmud et al., 2013) found that carriers of familial mutations causing the disease also had an elevated polygenic risk for the disease compared to healthy controls. However, their polygenic risk was lower than in patients with disease but no familial mutation. It would therefore seem plausible a hypothesis, that some of the polygenic burden in the DDD cohort was also carried by individuals with diagnostic variants. However, since the neurodevelopmental disorders in DDD patients are so severe, we thought it could also be that the these individuals did not carry elevated polygenic liability to neurodevelopmental disorders, since the single diagnostic variant in a developmental disorder gene could be enough to cause disease. In order to find out whether the polygenic burden discovered in Chapter 2 was more different in patients without diagnostic variants, I compared polygenic scores between patients with and without diagnostic variants.

From the cohort of 6,987 European ancestry and unrelated DDD patients who I used for NDD discovery GWAS, all had undergone exome sequencing as part of the DDD Study. From these patients, 1,127 have so far been found to carry *de novo* or inherited candidate diagnostic variants. In addition, 2,479 patients had no variants that passed the clinical filtering. From this analysis, I excluded 3,375 patients who had variants that were not likely to be diagnostic, but which had not been rated by clinicians.

Table 3.6: Polygenic score analysis comapring DDD patients who have a genetic diagnosis (N=1,127) to those who are genetically undiagnosed (N=2,479). Diagnosed cases were labelled as 1 in the logistic regression.

Polygenic score	Estimate	Std.Error	P
Educational attainment	0.080	0.037	0.028
Intelligence	0.063	0.036	0.080
Schizophrenia	0.017	0.036	0.644
Autism	-0.077	0.036	0.032
Intracranial volume	0.005	0.036	0.891
Birth weight	0.002	0.036	0.966
Height	0.001	0.036	0.971

The analysis comparing polygenic scores showed that diagnosed patients were not significantly different from undiagnosed patients with respect to any of the polygenic scores tested, after correcting for multiple testing (Table 3.6). Since the sample sizes for this analysis were quite low, this analysis was potentially underpowered. I tested our power to detect a significant difference in polygenic scores between these groups. These power analyses showed that diagnosed patients were not as different from undiagnosed patients as population controls were, at least for educational attainment and schizophrenia (Methods). This suggests that both common and rare variants are contributing in many neurodevelopmental disorder patients. As the DDD project continues to identify new diagnoses, we anticipate that the increase in power by adding more patients to the diagnosed or undiagnosed group may show that monogenic and polygenic contributions are not purely additive.

In the meantime, in attempt to increase the power for detecting differences between diagnosed and undiagnosed patients, I tried using different criteria to add samples or to refine the set of individuals included. I then performed the logistic regression between diagnosed and undiagnosed patients based on new criteria:

– Including uncertain cases in the undiagnosed set (low predictive value for the variant(s) and not rated by clinicians)

- Restricting the undiagnosed set to only definitely undiagnosed (note here the difference to the main analysis is that this excluded the likely benign variants who are probably undiagnosed)
- Restricting diagnosed set to those who have a *de novo* loss-of-function variant or large deletion in a gene with pLI>0.99 or pLI>0.999

None of these new criteria showed significant differences in any of the polygenic scores tested between the two patient groups. This indicated that the finding of common variants contributing to both diagnosed and undiagnosed patients with severe neurodevelopmental disorders is quite robust (at these sample sizes) to changes in how we define the diagnosed and undiagnosed patient groups.

Comparing patients with mild or severe developmental delay/intellectual disability

The detailed phenotype information annotated for DDD patients also allowed me to look into the impact of common genetic variation to the severity of global developmental delay and intellectual disability. Intellectual disability (HP:0001249) is a neurodevelopmental disorder where individuals suffer from deficits in intellectual and adaptive functioning, that begin during the developmental period (American Psychiatric Association, 2013). Intellectual functioning is typically measured with psychometric testing, where scores more than two standard deviations below the mean are regarded as intellectual disability. The severity of the condition is usually determined by the level of adaptive functioning. Global developmental delay is used to describe delay in reaching a number of intellectual performance developmental milestones when children are typically under the age of five, and therefore they are too young to be assessed using tests for intellectual disability (American Psychiatric Association, 2013). Within the cohort of 6,987 European ancestry (unrelated) DDD patients, 69% had global developmental delay and/or intellectual disability (DD/ID). Of these, 13.3% had mild DD/ID, 26.2% had moderate and 18.9% had severe DD/ID; the remaining 41.6% had DD/ID of unspecified severity.

Table 3.7: Polygenic score analyses comparing DDD patients with mild/moderate (N=1,902) or severe (N=911) developmental delay or intellectual disability. Severe cases were labelled as 1 in the logistic regression.

Polygenic score	Estimate	Std.Error	P
Educatinal attainment	0.116	0.040	0.004
Intelligence	0.089	0.040	0.028
Schizophrenia	0.078	0.041	0.054
Autism	0.001	0.041	0.974
Intracranial volume	0.010	0.040	0.800
Birth weight	-0.045	0.040	0.260
Height	-0.057	0.041	0.161

I wanted to assess whether the severity of developmental delay or intellectual disability was associated with any polygenic scores I had constructed for these patients. The results showed that severe DD/ID patients (N=911) were significantly enriched for educational attainment increasing alleles compared to mild or moderate cases (N=1,902) (P=0.004, variance explained Nagelkerke's R^2=0.008), after correcting for multiple testing (Table 3.7). Whilst this finding might seem initially counter-intuitive, it is consistent with epidemiological studies (Reichenberg et al., 2016) which found that the siblings of patients with severe intellectual disability showed a normal distribution of intelligence quotient (IQ), whereas siblings of patients with milder intellectual disability had lower IQ than average. This implied that mild intellectual disability represents the tail-end of the distribution of polygenic effects on intelligence and severe intellectual disability has a different etiology. At the time of writing this thesis, another study was published on bioRxiv (Kurki et al., 2018) which found individuals with intellectual disability in a Northern Finnish cohort had lower polygenic scores for educational attainment and intelligence, and higher scores for schizophrenia than matched controls. In addition, the authors found no significant difference between patients with and without diagnostic exome mutations in genes known to cause developmental disorders. These findings are in line with our observations, but the authors did not see a significant difference between mild and more severe forms of ID (though their sample sizes were smaller than ours).

Adding patients to developmental delay categories based on developmental milestones

In an attempt to increase power for the severity of DD/ID analysis, I tried different ways of moving patients from the DD/ID of unspecified severity category, who were excluded from the analysis above, to either mild or severe categories. In order to do this, I used phenotype data on the age when the child developed their first five words and the age when they first walked (Figure 3.3). If the child reached the milestone at the age of less than +4SD from the mean, I assigned them to have mild delay. Children who reached the milestone by +8SD from the mean were assigned to have severe delay. For the analysis, I also removed patients who at the time of assessment were younger than +4SD the population mean age of reaching the milestone, as we cannot distinguish whether these patients are severely delayed or not.

Re-categorising patients with unspecified severity DD/ID based on their milestones added a further 168 patients to mild category and 657 to severe category (total 2,070 mild and 1,568 severe). But the result was the same as in the original analysis despite the increase in sample size, as only educational attainment polygenic score was significantly associated with severity of delay (higher in the severe group, P=0.008).

I also attempted an analysis, whereby I re-categorised all patients from the previous analyses based solely on their recorded milestones. Here, anyone with DD/ID HPO regardless of the specific category who reached walking or talking milestone at age <4SD was labelled as mild, and >8SD as severe delay. For walking alone, the mild category contained 1,902 patients and the severe category 1,798. For talking alone, mild category had 1,023 patients and severe 2,966. I found no significant association between the delay of speech or walking and the polygenic scores. Finally, I asked whether individuals who were either mildly delayed or severely delayed in both talking and walking had different polygenic scores (mild N=586, severe N=1,424), but again there was no association. From these analyses, we can conclude that face-to-face clinical assessment is more effective in distinguishing patients with

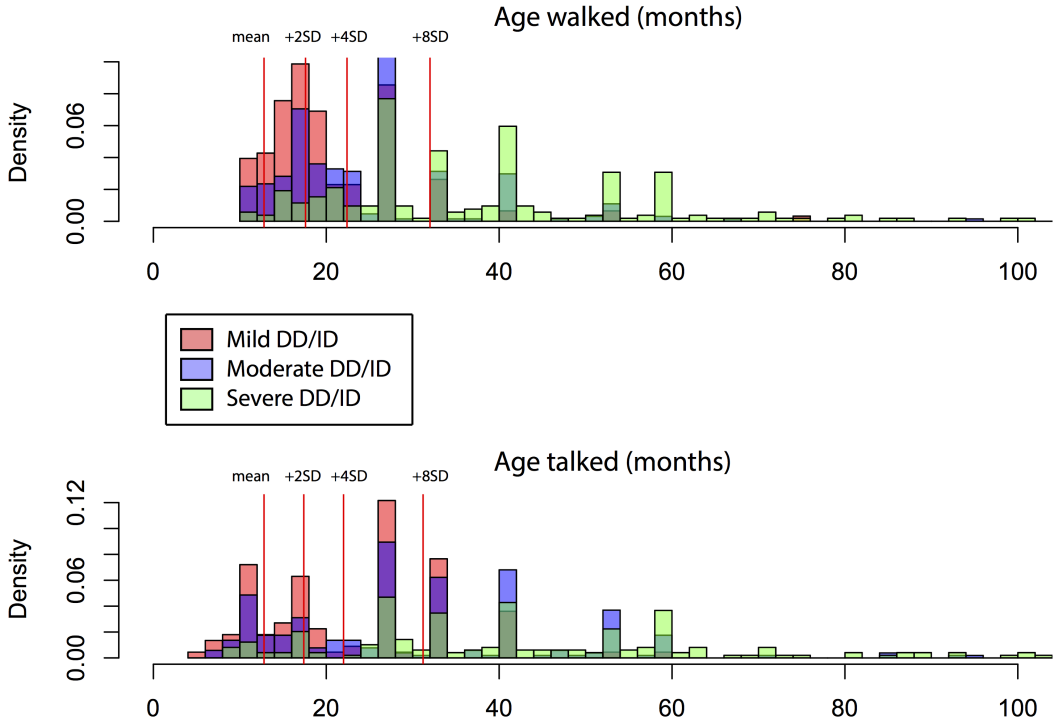


Figure 3.3: Histogram of the age that DDD patients with developmental delay or intellectual disability reached developmental milestones. The patients are coloured by severity of DD/ID. There is a long right hand tail of individuals from each of the categories, however, since mild DD/ID was plotted first and severe last, the tail appears green as the numbers in the severe category were highest. The population mean, +2SD, +4SD and +8SD from the mean are displayed as red lines.

differential genetic aetiologies with respect to their developmental delay, than is categorising patients by their developmental milestones.

3.5.5 Investigating phenotypic expressivity in DDD patients

A final question that I was able to investigate with the DDD cohort genotype and phenotype data was whether individual presentation of symptoms within the cohort was affected by common genetic variants for that trait. I identified four

phenotypes measured in our neurodevelopmental disorder cohort, which are often described as part of the symptoms or syndrome of patients in the DDD, and for which published GWAS were available. These were autistic behaviour (16% of cohort, HP:0000729), birth weight, height, and head circumference. On average, the 6,987 neurodevelopmental DDD patients I studied had a head circumference 1.20 SD smaller, they were 0.72 SD shorter than, and weighed 0.15 SD less than the age and sex-adjusted population average.

Using common variant polygenic scores for the four phenotypes described above, I tested for association between the phenotype and relevant score in our cohort, including 10 ancestry PCs as covariates. In all four traits, there was significant association with the score (Table 3.8), demonstrating that common variation contributes to the phenotypic expression of these traits in our study. Although this type of trait $\sim$ polygenic score association seems unsurprising to those in the complex trait field, in clinical genetics where these traits are typically considered to form part of the patient's profile of symptoms, the finding may have significance in understanding variable phenotypes among patients. Whilst these results do not directly answer whether common variants play a role in variable penetrance of specific severe neurodevelopmental disorders, the indication is that this could be the case. In order to investigate this we would need larger cohorts of individuals with rare variants in the same genes with deep phenotype data.

Table 3.8: Association between measured traits and the relevant polygenic score in 6,987 DDD patients (European ancestry). Linear or logistic regression of measured traits in the DDD Study against the respective polygenic score, including ten ancestry principal components as covariates. P-values are two-sided, from t-distribution (linear) and z-score distribution (logistic). Autistic cases were labelled as 1 in the logistic regression (Naegelkerke's R^2 reported).

Measured trait	Polygenic score	Estimate	Std.Error	P	R^2
Height	Height	0.408	0.033	1.2×10^{-35}	0.033
Birth weight	Birth weight	0.187	0.017	2.5×10^{-28}	0.020
Head circumference	Intracranial volume	0.132	0.031	1.8×10^{-5}	0.004
Autistic behaviour	Autism	0.120	0.033	2.5×10^{-4}	0.006

To better understand how well our polygenic scores were explaining variance in these phenotypes, I compared some of these to phenotype predictions performed in the original studies. The autism GWAS (Grove et al., 2017) showed that with

five target-training samples within their cohorts, the mean variance in the trait explained was 2.5% (Naegelkerke's R^2). As comparison, the autism polygenic score in DDD patients explains 0.6% of variance in whether the patient shows autistic behaviour or not. For height, although the polygenic scores explained 3% of the variance in height in DDD patients, this is substantially lower than what the original GWAS study reported for a test set of individuals, which was close to 30%. What this perhaps indicates is that whilst polygenic effects are still influencing the phenotypic expressivity of height in severe neurodevelopmental disorders, other genetic and non-genetic factors potentially have a relatively larger contribution to height in this patient group than in the general population.

3.5.6 Challenges in interpreting genetic correlation

Control ascertainment in NDD risk GWAS and effects on r_g

Whilst the analyses of genetic overlap between NDD risk and other common traits and diseases yielded very interesting results, I also came across some results that required careful consideration over potential differences in sample ascertainment between studies. In this section, I will discuss some of these observations and the implications that sample ascertainment might have on studies, particularly those looking for shared genetic effects between cohorts ascertained for the same phenotype and between different traits.

The strongest findings from bivariate LDSC of NDD risk with 19 other traits was a negative genetic correlation with educational attainment and intelligence. Although intuitively this finding makes sense in the context of NDD patients being severely cognitively affected, we need to consider the possibility of sample ascertainment affecting the r_g results. It is, for example, possible that individuals who are more highly educated and cognitively better functioning are more likely to consent to participate in studies, particularly as controls. In context with our findings from the r_g analysis, the question then was whether the effect was driven by real depletion of educational attainment and intelligence increasing alleles in DDD patients, or

by enrichment of educational attainment- and intelligence-increasing alleles in the controls.

The concern over sample ascertainment initially arose during the NDD discovery GWAS phase. I had at the time included another control cohort in the analysis, from Born in Bradford study (BiB). This longitudinal study recruited expectant mothers at the Bradford Royal Infirmary, between 2007 and 2010 (Raynor and Born in Bradford Collaborative Group, 2008), with the aim to study both genetic and environmental factors affecting the wellbeing and health of families. The genetic data collected for the study included mothers who were genotyped on a version of the Illumina HumanCoreExome chip. Out of these 3,033 had European ancestry and I included them as controls in my study, along with UKHLS.

After I had performed an initial NDD GWAS, I also checked the genetic concordance between the two control cohorts by performing a GWAS of BiB mothers (as cases) against UKHLS (controls). The findings, summarised in Table 3.9, were surprising. First of all, the SNP heritability estimate for BiB vs. UKHLS was greater at h^2=0.141 (SE=0.043) than our NDD GWAS (h^2=0.043, SE=0.020) even with a smaller sample size. This was alarming, so I carried out a r_g analysis with educational attainment (Lee et al., 2018), which showed that there was a more significant negative genetic correlation between BiB and UKHLS than between DDD and UKHLS with respect to genetic factors influencing educational attainment (Table 3.9). I then carried out a GWAS comparing DDD (as cases) to BiB (controls), and found that there was no significant SNP heritability or genetic correlation, although the direction of effect was that BiB were even more depleted for educational attainment increasing alleles than DDD were. Other education and cognitive traits showed the same trend for these r_g analyses. To test whether the lack of significant differences between DDD and BiB was due to decreased sample size, I performed three simulation GWAS in which I randomly sampled 3,033 non-overlapping sets of UKHLS and used these as controls against DDD. These analyses demonstrated that the DDD vs. BiB was likely underpowered, but also that there is potentially less of a difference between DDD and BiB than DDD and UKHLS with respect to common variants.

Table 3.9: Results from LDSC SNP heritability analysis for GWAS on different cohort pairs. Genetic correlation of the GWAS with educational attainment. Liability scale conversion was done assuming a population prevalence of 1% for developmental disorders. Note, compared to our final discovery NDD GWAS, the GWAS used in these analyses included relatives, which I later removed to reduce noise in the analysis.

Cases	Controls	SNP heritability			r_g with Years of Education (2018)		
		h^2 (liability scale)	SE_h^2	Prop. polygenic effects	r_g	Standard error (r_g)	P
DDD (N=7,274)	UKHLS + BiB (N=13,087)	0.043	0.020	0.296	-0.489	0.124	8.3×10^{-5}
DDD (N=7,274)	UKHLS (N=10,054)	0.068	0.023	0.395	-0.537	0.100	8.8×10^{-8}
DDD (N=7,274)	BiB (N=3,033)	0.026	0.039	0.069	0.358	0.296	0.226
BiB (N=3,033)	UKHLS (N=10,054)	0.141	0.043	0.275	-0.515	0.079	7.9×10^{-11}

The conclusion that could be drawn from these analyses was that mothers recruited to BiB were more depleted for education-increasing alleles than DDD children with rare severe neurodevelopmental disorders. This could perhaps be explained by the fact that Bradford is one of the most deprived areas in the UK. Since educational attainment correlates with socio-economic status (White, 1982), there may be differences in the geographical distribution of these effect alleles even within a country, that may result from a migration of more highly functioning individuals from less affluent to more affluent places; a finding that is supported by a recent preprint paper on UK Biobank data (Haworth et al., 2018). These results highlight the importance of the choice of controls in GWAS, particularly those on traits related to cognition, as their ascertainment can greatly affect the resulting estimations of genetic effects. In the light of these findings, I dropped the BiB mothers from the analysis, because I did not believe they were a good representative population for the UK based on the information we had about the socio-economic differences between Bradford and the rest of the UK.

UKHLS are a good (enough) representative sample of the UK population with respect to educational attainment

Since one of our key results from the genetic correlation of NDD risk analysis was the depletion of education and cognition involved alleles in the DDD, the question

then remained, whether the UKHLS were representative of the UK population in terms of the distribution of these alleles. In other words, if the UKHLS was not a suitable comparison group, this could greatly impact the interpretation my findings.

The UKHLS collects longitudinal data on its participants. These phenotypic data included information about the highest level of educational qualification obtained. Epidemiological studies looking at these data have argued that individuals in the UKHLS who consented to their health data being recorded may be slightly biased towards those who achieved secondary education (Cruise et al., 2015; Knies and Burton, 2014). However, since the individuals who consented to giving blood for DNA analysis were a subset of this group, we did not have information on whether the subset was also biased with respect to their education. We therefore obtained the educational attainment phenotype data on these UKHLS individuals, to see whether they were skewed with regards to their educational attainment compared to census and labour market data on the UK population (Table 3.10).

For all genotype participants in the UKHLS phenotype data, I extracted the highest educational qualification they had achieved by 2012 (the year the nurse visits were completed). The variable also took into consideration what the participant had answered during previous data collections during the longitudinal study, so this variable would always represent the highest qualification recorded at any point. I then compared this data to the UK census 2011 data. Both datasets only included responses from individuals who at the time were 16 years or older, though it was likely the UKHLS age distribution would have been skewed to higher ages compared to the census. The proportions of individuals who achieved a certain level of qualification in all three datasets is summarised in table 3.10. It would seem that the UKHLS is not particularly enriched for individuals who achieved a degree, but there are larger differences in the lower categories. This may be partly due to the fact that the census data categorises qualifications in a different way, and partly due to real differences. When comparing the genotyped UKHLS cohort to official labour market statistics from 2012, these match for the higher categories of education, however the major caveat is that these are statistics for 16 to 64 year olds, whereas a substantial proportion of UKHLS will probably be over

this maximum age. Data from census and labour market survey are displayed in Appendix B.

Table 3.10: Comparing UKHLS to census and labour market data. UKHLS variable "hiqual_b" from 2012 with available categories, UK census 2011 data for equivalent categories, and official labour market statistics from 2012. Missing data for UKHLS was removed prior to calculation of percentages.

	UKHLS 2012 (%)	UK census 2011 (%)	Nomis statistics 2012 (ages 16-64) (%)
Degree/other higher degree	33.8	27.2	34.0
A-level	19.6	12.3	19.0
GCSE	21.2	15.3	18.7
Other qualification	11.2	22.5	18.4
No qualification	14.2	22.7	10.0

In attempt to convince ourselves that the NDD risk discovery GWAS had found true polygenic effects that were driven by the patients, and not UKHLS controls, I tested the unrealistic scenario whereby I removed from the controls all UKHLS individuals who had achieved a degree qualification. I repeated the GWAS, and the SNP heritability analysis still showed significant h^2=0.059 (SE=0.025) common variant heritability (assuming population prevalence 1%). The negative genetic correlation with educational attainment (Lee et al., 2018) also remained significant r_g=-0.22 (SE=0.075, P=0.0037), although it was attenuated. The positive genetic correlation with schizophrenia r_g=0.23 (SE=0.081, P=0.0048) also remained. Together, these results imply that even if the UKHLS genotyped samples were slightly biased towards individuals who had achieved secondary school education, after removing a the top third of the whole cohort with respect to their educational attainment (and reduced power to detect polygenic burden and genetic correlation due to the reduced sample size) there is still significant polygenic burden associated with neurodevelopmental disorder risk. It is worth noting here that the over-transmission of NDD-risk alleles from parents to probands (Chapter 2.5.3) already provided strong support for true polygenic contribution driven by the cases. The distribution of educational attainment associated alleles in the UKHLS controls when having removed a third of the cohort is highly likely not a realistic representation of the population. But the point of this analysis was to show that the results from our NDD GWAS are quite robust to this extreme subsetting.

3.6 Discussion

In this chapter, I have shown that common variant effects that contribute to the risk of severe, rare neurodevelopmental disorders in the DDD Study are shared with other traits that involve brain function or development. Similar findings in an independent cohort of Australian neurodevelopmental patients provides evidence for replicability of the results, with a cautionary note that differential ancestry and ascertainment of neurodevelopmental disorders may cause heterogeneity between cohorts. However, when interpreting analyses assessing genetic overlap, an important consideration is that any given results will be affected by sample size and ascertainment of the original studies. As GWAS continue to grow in sample size, they gain more power for association and heritability analyses. New reports on genetic correlations thus continue to emerge, expanding our understanding of the genetic architecture of various traits and diseases. But this also means that GWAS cohorts, and thus the underlying genetic architectures, used for these analyses undergo changes over time. The consecutive addition of more samples to existing datasets, and the analyses of completely newly ascertained cohorts, can sometimes change the estimates of genetic correlation between traits, or remove them completely. This brings attention to some of the drawbacks of looking for genetic overlap between studies where samples may have been ascertained in very different ways, and calls for more careful consideration when interpreting genetic correlation results. In this chapter, I have particularly discussed examples of how recruitment of study participants might affect our understanding of subtle sharing of genetic effects between traits.

These results from our neurodevelopmental disorder GWAS and Australian cohorts indicated that NDD patients are depleted for alleles that increase educational attainment and intelligence, and enriched for those contributing to the risk of schizophrenia, a neuropsychiatric disease. This is interesting in the light of published literature on genetic overlap between different neuropsychiatric and cognitive traits (Okbay et al., 2016; Pardiñas et al., 2018; Brainstorm Consortium et al., 2018; Grove et al., 2017), and particularly that of intellectual disability which found essentially the same results (Kurki et al., 2018). In our study, the strong correlation of NDD discovery GWAS with both educational attainment and schizophrenia,

which do not share common variant effects with each other (Pardiñas et al., 2018), shows that NDD risk arises from a more complex combination of variant effects. We also do not see a significant association between NDD discovery GWAS and bipolar disorder, which is known to share polygenic effects with schizophrenia (International Schizophrenia Consortium et al., 2009). Potential reasons for these include the fact that the schizophrenia GWAS is just much better powered with a larger sample size, or that NDD risk shares with schizophrenia specifically more of the effects that are not shared with bipolar disorder. A recent study by Bansal et al. (2018) found evidence that the polygenic relationship between educational attainment and schizophrenia is not homogeneous across patients, indicating that both traits are genetically heterogeneous. They suggest that some patients' polygenic background is more concordant with bipolar disorder and higher cognitive performance, and others are more independent of these. The findings could explain part of the genetic correlations of NDD risk with other traits that we observe, and some of the more or less unexpected results.

The reported genetic correlation between educational attainment and schizophrenia on its own is interesting to us in the context of GWAS power and sample ascertainment. Epidemiological studies have shown that individuals suffering from schizophrenia have poorer educational attainment (Swanson et al., 1998) and cognitive performance (Bowie and Harvey, 2006) even before the onset of disease. But contrarily, an educational attainment GWAS from 2016 (Okbay et al., 2016) described a small, but significant positive genetic correlation with the schizophrenia GWAS from 2014. Both traits now have a newer, larger GWAS, but neither of these studies report on genetic correlation between the traits (Lee et al., 2018; Pardiñas et al., 2018). Out of interest, I ran the r_g analysis between these two studies from 2018. The results showed that there was no significant genetic correlation between these studies (r_g=0.009, SE=0.018, P=0.62). On the other hand, intelligence (Sniekers et al., 2017) and schizophrenia (2018) show significant negative genetic correlation r_g=-0.226 (SE=0.0298, P=3.6×10^{-5}), which is in line with findings from epidemiological studies. In this particular example, one of the many reasons for why educational attainment is first correlated with schizophrenia but then not in the newer studies could be differences in sample ascertainment between the studies

on the same trait. Educational attainment is a quantitative trait and therefore probably easier to measure in populations, whereas recruitment of patients with a neuropsychiatric disease such as schizophrenia requires clinical recruitment. One possibility therefore could be that the schizophrenia patients recruited to the 2014 study (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) could have represented a cohort of higher functioning patients, who were able to consent to taking part in a research study. A bias such as this could in turn result in the positive genetic correlation with higher educational attainment. The 2018 schizophrenia study included the CLOZUK patient cohort, who were recruited under the requirement that they were taking the oral antipsychotic drug clozapine. There might thus be an argument to say that this sample may be less biased towards sampling higher functioning patients. Although in this particular example these notions are just speculation, the topic of recruitment bias in GWAS cohorts is an important one which perhaps does not get as much attention in the field as it should.

The interpretability of genetic correlation results is also affected by the SNP heritability of the traits. As Wray et al. (2018) note, high r_g with a trait is more reliable when the SNP heritability for both traits is high. In the case of our discovery NDD risk GWAS, h^2 is still relatively low. Resulting from this, the magnitude of r_g with other traits that have a h^2 below $\sim$0.10 have very wide confidence intervals, as with the example r_g between NDD risk and ADHD, although this overlap was not significant after multiple testing correction (Figure 3.2). It is possible though that if we had more power for detecting shared effects, these other neuropsychiatric or neurodevelopmental traits like ADHD and major depressive disorder may pass the threshold for significant association. Both these traits have been shown to be correlated with schizophrenia (Wray et al., 2018) and bipolar disorder (Hulzen et al., 2017), and therefore a positive direction of correlation with NDD risk was perhaps expected. It would potentially have been interesting to include other neuropsychiatric and developmental polygenic scores in our analyses (e.g. for bipolar disorder) but we wanted to select either well powered GWAS with sample sizes in the tens of thousands, or GWAS that we had a measured phenotype data for in the DDD cohort.

The genetic correlation results did not show significant overlap of overall NDD risk with autism spectrum disorder, and again there was no significant difference in autism polygenic scores between Australian cases and controls. This is despite 16% of the neurodevelopmental DDD patient subset showed autistic behaviour. Autism is known to be associated with rare variants (Koch, 2014; O'Roak et al., 2014; Iossifov et al., 2014), but also has a substantial contribution from common variants with a common-SNP heritability of 0.09 (Grove et al., 2017). As the DDD cohort is a mixture of patients with a range of phenotypes and severities, and with ∼30% of the cohort genetically diagnosed, it would have seemed slightly implausible that autism in each of these patients was explained solely by rare variants. Indeed, when separating the DDD cohort into those who showed autistic behaviour and those who did not, there was significant association between this behaviour and increased autism polygenic scores. The same pattern was seen for birth weight, height, and intracranial volume, all which are traits for which DDD children are on average below the population mean. Together these results illustrate how despite no overall genetic correlation with NDD risk, common variants that affect these traits in the general population are also affecting phenotypic expressivity of the trait in our neurodevelopmental cohort.

The negative result of no association between case/control status and NDD risk polygenic score in the Australians is interesting, particularly since out-of-sample prediction for e.g. schizophrenia has been successful at similar sample sizes (International Schizophrenia Consortium et al., 2009). However, schizophrenia has a much higher SNP heritability (h^2=0.24 in the study I use for polygenic scores (Pardiñas et al., 2018)) than DDD (h^2=0.077), and we know that predictive power depends on the amount of heritability to be found. As a comparison, we can look at how prediction using educational attainment scores with similar sample sizes has performed in other studies. The current estimate for educational attainment SNP heritability from the largest GWAS to date is h^2=0.11 (Lee et al., 2018). This GWAS was conducted using 1.1M samples, and we find significant association in our Australian cohort. However, an earlier study by Martin et al. (2011b) found that an educational attainment polygenic score constructed from a discovery GWAS of N=9,538 Australian individuals failed to explain any significant variation

(r^2<0.0023, p≥0.14) in an independent cohort of 968 individuals. This illustrates how the discovery GWAS sample size and SNP heritability of the trait can affect the ability to detect association with polygenic scores. It is also likely that there is more heterogeneity between our UK and Australian cohorts due to international ascertainment differences, than there would be between two schizophrenia cohorts. Schizophrenia is a clinically well characterised trait, although the specific combinations of symptoms may be more heterogeneous than diseases such as Crohn's disease. Overall, our power analysis supports phenotypic heterogeneity due to ascertainment differences between the NDD discovery cohort from the DDD Study and the Australian cohorts. A final important notion from the polygenic score analyses in Australians demonstrated, was that scores constructed from even large meta-analyses of dichotomous traits are vulnerable to bias from sample overlap with the target population, as we saw from the schizophrenia analysis.

An important finding from our study was that DDD patients with diagnostic rare variants were not significantly different from patients without a genetic diagnosis. This suggests that rare and common variants are both contributing to disease risk in the DDD cohort. The study by Kurki et al. similarly found that individuals with intellectual disability and likely causative rare variants were no different from patients without a likely causal variant with respect to their polygenic scores for educational attainment, intelligence and schizophrenia. Another study consistent with these findings, by Weiner et al. (2017) similarly found no evidence for a difference in polygenic risk scores between autism cases with a *de novo* diagnostic mutation compared to those without. This suggests that both common and rare variants are contributing in many neurodevelopmental disorder patients. If common and rare variants are acting together in these patients with severe disorders, a question that then arises is whether common variants could be affecting the penetrance of disease in patients but also in the general population. This lead us to wonder whether a polygenic profile skewed towards the opposite end of the spectrum, i.e. enriched for cognitive performance increasing alleles and depleted for neuropsychiatric disease alleles, could have a protective modifying effect on an individual in the presence of rare, damaging variants. In Chapter 4, I explore these questions in a cohort of seemingly healthy individuals from the general population.

Chapter 4

Do common variants protect against rare, deleterious variants in the general population?

4.1 Chapter overview

In this Chapter, I wanted to understand how rare and common variants affect the cognitive scores (general intelligence) in a cohort of healthy individuals. I test whether carriers of apparently deleterious rare variants were protected by common variant polygenic scores, and whether there was an interaction between rare variants and polygenic scores, as has been previously reported for a related trait educational attainment (Ganna et al., 2016). These two questions are related but not the same. I therefore first ask whether there is a difference in the overall distribution of polygenic scores between rare variant carriers and those without, and then look deeper into whether there is an interaction between the variant types.

4.2 Background

In Chapters 2 and 3, I showed that neurodevelopmental disorder risk has a polygenic component that overlaps with liability for complex neuropsychiatric phenotypes in the population, and that common variants affect expressivity of specific phenotypes in the DDD cohort. An interesting question which we did not directly touch upon in the previous chapters was whether common variants also affect the penetrance of rare variants in genes associated with developmental disorders.

Unpublished work in the DDD cohort by Kaitlin Samocha suggests that DDD patients are enriched for rare LoFs and missense variants in known developmental disorder genes and in genes depleted of such variation in the general population. These are generally inherited from unaffected parents, implying that such variants can be incompletely penetrant. We also know that patients with variants in the same genes show different levels of severity of the disorder. Given the results from Chapters 2 and 3, assessing whether common variants affect penetrance of rare variants in genes associated with neurodevelopmental disorders would therefore be interesting to us. However, there are several complicating factors to doing this in the DDD cohort. First of all, as our probands are all affected, it is difficult to assess penetrance of disease associated variants observed in the cohort since it is likely that they are causal for the symptoms and therefore highly penetrant on those individuals' genetic background. Instead, assessing DDD patients would probably tell us more about the expressivity of symptoms rather than penetrance. In addition, patients may have multiple genetic aetiologies responsible for different symptoms, which may or may not be masking the effects of other variants on phenotypes of interest. For inherited candidate variants, we could utilise genetic and phenotypic data from the parents or siblings where available, but often the phenotypic information recorded for family members is not detailed and therefore it may not be clear whether they are in fact somewhat affected or completely healthy. At present, we have genotype data for only ∼1,000 pairs of parents, and therefore assessing polygenic scores for parents is not feasible.

We have seen from previous studies of large sequenced cohorts, that even presumably healthy individuals from the population carry damaging rare variants that would

be expected to cause disease (MacArthur et al., 2012; Narasimhan et al., 2016). We therefore wondered whether we could find rare deleterious variants in genes associated with neurodevelopmental disorders or in other brain-expressed genes that have been shown to be intolerant to deleterious mutations, in a healthy population cohort. If we were to find individuals carrying these variants, we would then want to investigate further why these individuals were not suffering from severe phenotypes. Specifically, we would be interested in investigating whether common variants are modifying these variants' penetrance, by acting in a protective manner in healthy individuals.

We therefore sought out a cohort of healthy individuals with genetic data available on both common and rare variants, and with relevant phenotype data. The UK-based INTERVAL cohort of ∼50,000 healthy blood donors had genotyped all participants on a DNA chip, and also exome sequenced a smaller subset of participants. Using this data meant that we could assess both common variants and rare variants not captured by chip data. Conveniently for our purpose, the participants in INTERVAL were also asked to complete tests that give an overall measure of cognitive performance, or general intelligence. This general intelligence score is a continuous measure, and would allow us to assess the impact of common and rare variants on cognitive performance.

General intelligence is measure of overall intelligence or cognitive performance, described first by Robert Spearman (1904). Spearman believed that there is a single factor underlying different types of cognitive abilities. The single measure of cognitive performance, general intelligence (or cognitive score g) has been shown to explain a large proportion of variance in a variety of tests that can be used to measure different cognitive abilities. The genetic factors contributing to general intelligence have been extensively studied over decades. Intelligence is highly heritable, and increases with age to H^2 of ∼0.50-0.80 (Plomin and Deary, 2015). The largest GWAS (using data from 78,000 individuals) so far have found that common variants explain 20% of the variance in the trait (Sniekers et al., 2017), although a recent study by Hill et al. (2018) showed that more heritability may be found in lower frequency variants.

Directly measuring cognitive scores requires participants to fully complete one or more extensive questionnaires, which may not be possible in large cohorts recruited for unrelated research purposes. For this reason, many studies indirectly measure cognitive performance of participants by asking them about their educational attainment. Educational attainment (or the number of years of schooling) in GWAS has been used as a proxy phenotype for cognitive performance because the two traits share common variant heritability (Okbay et al., 2016; Rietveld et al., 2014) (genetic correlation between educational attainment (2016) and intelligence (2017) r_g=0.71, SE=0.02). Due to the high genetic correlation between the traits, it may be expected that phenotypes associated with polygenic scores for intelligence will also be associated with polygenic scores for educational attainment. The first study showing that ultra rare variants in the general population also significantly affect educational attainment was published in 2016 by Ganna et al. (2016). Therefore one of the first questions we want to address is whether rare variants are in fact associated with cognitive performance in INTERVAL, and then proceed to investigate this further by looking for protective effects of polygenic scores and testing whether common and rare variants ware intereacting with each other in INTERVAL.

4.3 Contributions

Quality control of INTERVAL GWAS data was performed by Heather Elding and Tao Jiang. Quality control of INTERVAL exome data was performed by Fernando Riveros McKay Aguilera and Tarjinder Singh, and further filtering was performed by Hilary Martin. Cognitive scores in INTERVAL were calculated by Hilary Martin, following instructions and guidance provided by Steven Bell and Ian Deary. The work described in this chapter was completed under the supervision of Hilary Martin and Matthew Hurles.

4.4 Methods

4.4.1 INTERVAL cohort

The INTERVAL study is a cohort of blood donors, recruited in 2012-2014 in the UK. The study was run through collaboration between the Universities of Cambridge and Oxford and the NHS Blood and Transplant Unit. The aim of the study was to assess the impact of different blood donation intervals on the wellbeing of 25,000 men and 25,000 women. This dataset is uniquely useful for our purpose, because the study collected three types of data on participants: GWAS chip data, sequence data and phenotype data including cognitive tests. For us, this meant that we could utilise data from individuals, for whom all three data were available, to assess the contribution of genetic variants in a joint model. In our study we use exome sequence data to find rare variants. INTERVAL has also performed whole-genome sequencing on a larger cohort of INTERVAL participants, but at the time of this thesis, these data were not ready for use.

4.4.2 Quality control of genotype data

Sample and variant quality control and imputation

Quality control (QC) and imputation of INTERVAL genotype chip data was performed by Heather Elding and Tao Jiang, with details described in (Astle et al., 2016). A total of 48,813 interval study participants were genotyped on the UK Biobank Affymetrix Axiom chip which assays 820,967 variants. Duplicate and non-European samples (from ancestry PCA with 1000G) were excluded before we received the data. The QC'd genotyped dataset included 43,059 European ancestry samples. The data were imputed using the Sanger Imputation Server (Loh et al., 2016), using UK10K-1000G Phase III imputation as the reference panel and SHAPEIT3 for imputation (Delaneau et al., 2011). All data were on GRCh37.

Ancestry check of European samples

The INTERVAL European sample selection had been performed as part of the study by Astle et al. (2016) before I received the data. I therefore wanted to check how tightly the European INTERVAL samples clustered together on an ancestry PCA, so I performed a new ancestry PCA of all the INTERVAL European samples using the same 1000G Phase 3 samples that I had used previously in Chapters 2 and 3.

For the PCA, I used the same protocol as previously in Chapter (Section 2.4.1). I ran the ancestry PCA in PLINK on directly genotyped variants, on 43,059 INTERVAL samples against 2,504 reference population samples from 1000G, using 73,444 variants (MAF>0.10) that overlapped between the two datasets. The PCA plot (Figure 4.1) showed that, compared to our European sample selection in Chapters 2 and 3, the Europeans selected by Elding and Jian were less tightly clustered around the 1000G European samples. A zoomed-in plot of the the INTERVAL samples (Figure 4.2) shows some substructure within the European INTERVAL subset that they had selected.

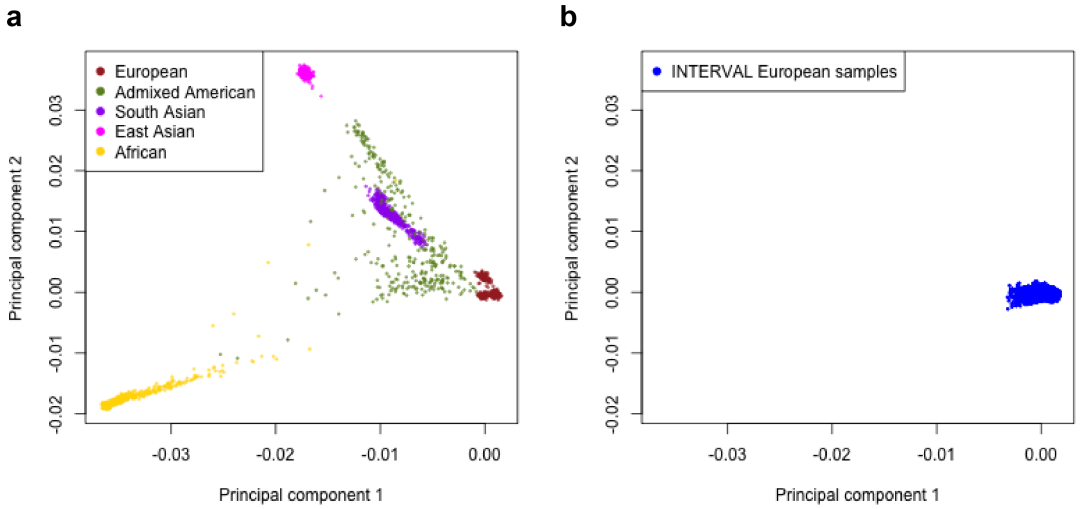


Figure 4.1: Ancestry principal components analysis of INTERVAL samples. **a**. Reference samples (N=2,504) from 1000 Genomes Phase 3, coloured by the five super populations, used for a projection PCA. **b**. European INTERVAL samples (N=43,059) that passed quality control (Astle et al. 2016).

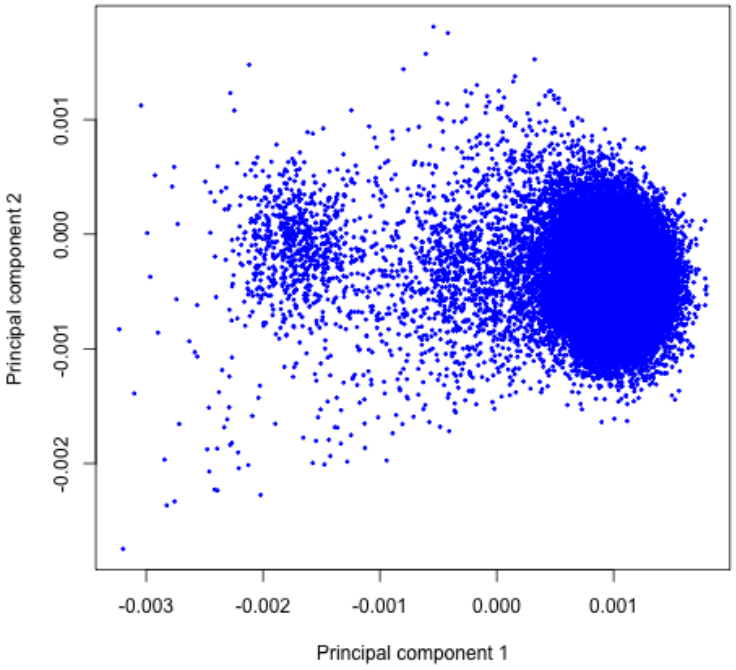


Figure 4.2: A zoomed-in plot showing European INTERVAL samples (N=43,059) from the ancestry PCA with 1000 Genomes.

Despite the apparent substructure, we decided to carry forward for further analysis all the INTERVAL samples presumed European by Elding and Jian. This is because we wanted to maximise the number of samples with all three types of data (GWAS, exome and cognitive scores) for our first-pass analysis of effects of rare and common variation on cognition. We reasoned that using ten ancestry principal components as covariates in our analyses should correct for any effects resulting from this substructure.

Removing relatives

To avoid bias in the planned analyses, I checked for relatedness in the INTERVAL cohort. I performed a relatedness check with 83,434 directly genotyped variants with MAF>0.10, using PLINK. I removed one individual from each pair of samples

equivalent to second-degree relatives or closer (alleles identical by descent >0.12), selecting the one who had the higher variant missingness rate. This resulted in a genotyped dataset of 41,580 unrelated individuals.

4.4.3 Polygenic scores

After removing relatives from the data, I filtered the imputed genotypes by removing variants with INFO<0.9, variants with MAF<0.05 and duplicate variants. I used these filtered, imputed data to construct polygenic risk scores, using the method described in Chapter 2.4.5 and Chapter 3.4.3. I constructed normalised polygenic scores for all 41,580 unrelated INTERVAL samples, using variant effects from our neurodevelopmental discovery GWAS, and GWAS on educational attainment, intelligence, schizophrenia, autism, intracranial volume, birth weight and height (parameters shown in Table 4.1). Our expectation was that that polygenic scores for intelligence and educational attainment would be most relevant to cognitive scores measuring general intelligence. Scores constructed from our neurodevelopmental disorder discovery GWAS could potentially be relevant as well, however as seen in Chapter 3 results in Australians, the GWAS is likely underpowered for these types of analyses. Although we did not expect polygenic scores for anthropometric traits to be associated with cognitive scores, these would still act as an additional check that the data were not bringing up unexpected associations indicative of potential biases.

Table 4.1: Parameters used for generating polygenic scores in INTERVAL cohort.

Polygenic score trait	r^2 for SNP pruning	P-value threshold for SNP pruning	Number of SNPs in score
Neurodevelopmental disorder (discovery GWAS)	0.1	1	71,978
Educational attainment	0.1	1	122,400
Intelligence	0.1	0.05	24,955
Schizophrenia	0.1	0.05	28,084
Autism	0.1	0.1	30,949
Intracranial volume	0.1	1	110,859
Birth weight	0.1	0.01	8,793
Height	0.1	0.005	10,660

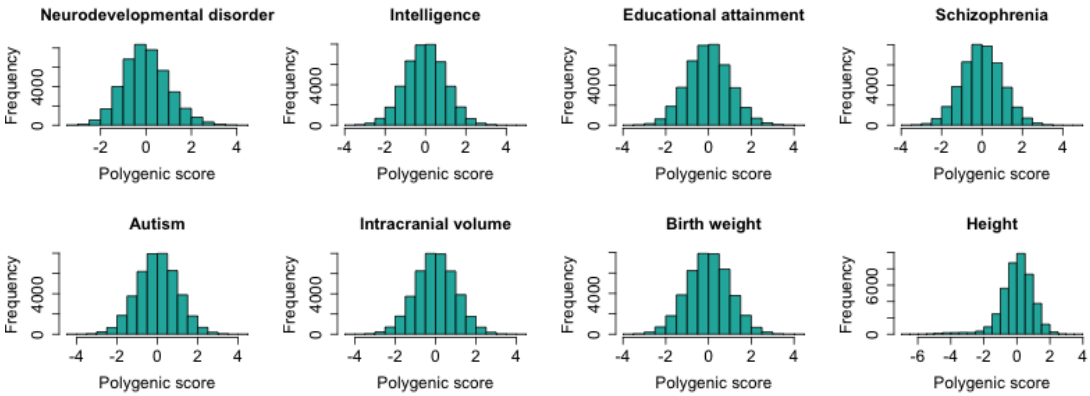


Figure 4.3: Distribution of polygenic scores for eight traits in the INTERVAL unrelated European cohort (N=41,580). Scores are normalised to a mean of 0 and variance of 1.

As shown in Figure 4.3, the polygenic scores were generally normally distributed within the full INTERVAL European dataset (N=41,580). The outlier plot is the one showing polygenic scores for height, which showed a heavy lower tail (bottom right panel Figure 4.3). As it is known that the distribution of height increasing alleles within Europe varies, particularly on the North-South axis (Novembre et al., 2008), I decided to investigate whether the observed heavy tail was due to population substructure within the INTERVAL cohort.

I first separated the individuals whose height scores were in the lowest 2% of the cohort (Figure 4.4). I then re-plotted the ancestry PCA of INTERVAL samples, but coloured those who fell into the lowest 2% in a different colour to the remaining 98%. From this plot, shown in Figure 4.5, it is evident that the 2% cluster together separately from the majority of the cohort. This illustrates that the heavy lower tail in height polygenic scores is accounted for by population substructure within the European subset of INTERVAL samples. However, using ancestry principal components as covariates in downstream analyses should in theory correct for bias resulting from this stratification.

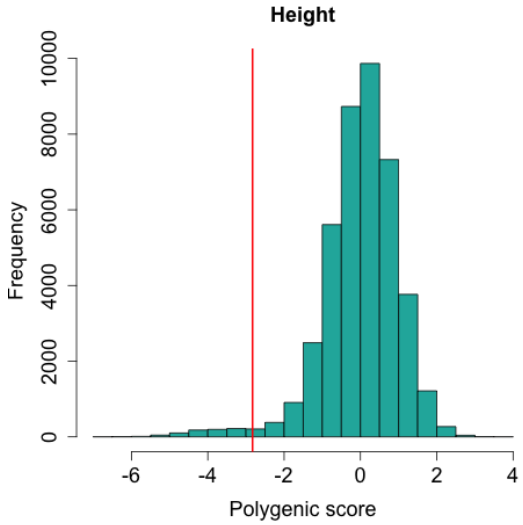


Figure 4.4: Distribution of height polygenic scores in INTERVAL (N=41,580). The histogram shows a heavy lower tail of height polygenic scores. The red vertical line is at the 2nd percentile, which is equivalent to -2.8SD from the mean.

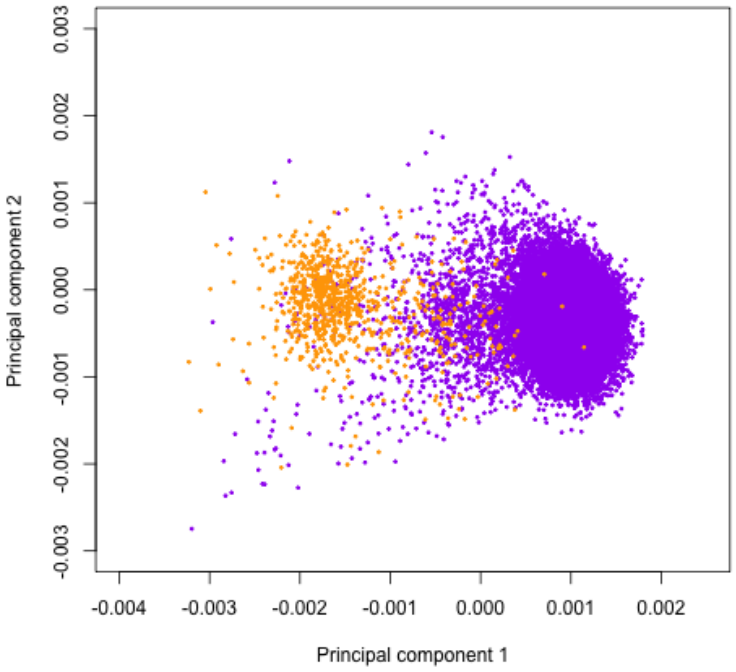


Figure 4.5: Ancestry PCA plot of INTERVAL Europeans (N=41,580), coloured by polygenic score for height. Interval samples with lowest 2% of polygenic score for height were coloured in orange, and the remaining 98% in purple to investigate whether the long tail for height scores was due to population substructure.

4.4.4 Quality control of exome data

We wanted to assess the contribution of rare loss-of-function (LoF) variants to cognitive functioning in INTERVAL, because these variants are more likely to be damaging and under purifying selection in the population. We also include in our analyses missense variants that are predicted to have a damaging consequence, and which lie within regions of a gene that are depleted of missense variants. High constraint refers to the gene being depleted of deleterious variants in large cohorts such as ExAC. These genes are likely haploinsufficient and deleterious variants in these genes will be under purifying selection. The pLI metric (Lek et al., 2016) is used to score genes for probability of loss-of-function intolerance. Variants in

genes with a pLI score of >0.9 are considered highly constrained. Previous studies have shown that deleterious rare variants in highly constrained genes are enriched in individuals with neurodevelopmental disorders such as autism (Kosmicki et al., 2017), schizophrenia (Genovese et al., 2016b), intellectual disability (Gilissen et al., 2014; Singh et al., 2017) and severe childhood developmental disorders (Deciphering Developmental Disorders Study, 2017; Singh et al., 2017). Many neurodevelopmental or neuropsychiatric (Pardiñas et al., 2018) associated genes fall within the category of high pLI genes, but the majority of pLI>0.9 genes do not yet have a disorder associated with them (Lek et al., 2016). We are therefore interested in assessing variants in pLI>0.9 genes, and particularly variants that have loss-of-function consequence on the the protein.

In total 4,502 individuals from the INTERVAL cohort were exome sequenced. Illumina paired-end sequencing was performed at the Wellcome Sanger Institute sequencing facility. Data were aligned and called by the Human Genetics Informatics team at the Sanger Institute. All data were aligned to GRCh37. We were looking for rare variants that are depleted in the population due to their damaging consequences. This means that true damaging variants are very rare, and the a proportion of apparent deleterious variants will be false positives. We therefore had to perform quality control for the exome data. For the purposes of this project, we applied further quality filtering on the data to find rare, loss-of-function (including small insertion/deletions of up to 10 basepairs) or deleterious (see below) missense variants in fetal brain-expressed genes. Genotypes were set to missing if they had GQ<20 (genotype quality), DP<7 (depth) to decrease the probability of missing a heterozygous call, and for heterozygous calls, a P-value from a binomial test of allele balance < 0.001 to remove false heterozygous genotypes. We then restricted to variants with MAF<0.001 to enrich for rare deleterious variants that are under negative selection, and that reside in genes expressed in fetal brain based on data from The Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013). Finally, variants were restricted to those that had (1) loss-of-function (LoF) consequence in all transcripts of a gene with pLI>0.9 (constraint score) to increase the probability that the variants are deleterious and were annotated as high confidence by LOFTEE (*loftee*), and are not in the last exon or intron; OR (2) were

missense variants with CADD>30 (in top 0.1% of deleterious variants, Combined Annotation Dependent Depletion) (Kircher et al., 2014) in (a) missense-constrained regions or (b) in genes with overall missense constraint that did not split into separate regions, and in a gene with pLI>0.9. These constrained genes and regions were defined in Samocha et al. (2017). For both the gene-wide and region-based analysis, we restricted to gene/regions with a ratio of observed to expected variation <0.4, and and chi-squared p<0.001. These came to a total of 1,029 LoFs and 711 missense variants in the cohort of 1,906 individuals, breakdown of samples with rare variants is shown in Table 4.2.

Table 4.2: Count of rare variants in INTERVAL individuals (total individuals N=1,906).

	Number of variants per person				
Variant class	0	1	2	3	4
Lof+missense pLI>0.9	1,491	365	43	6	1
LoF pLI>0.9	1,666	226	14	0	0
Missense pLI>0.9	1,702	192	9	3	0

4.4.5 Cognitive scores

As part of the INTERVAL study, participants were asked to complete sets of cognitive tests, that could be used to calculate a general intelligence or cognitive score (g). This testing was introduced part way through the study, meaning that many individuals were never asked to complete the study. Therefore, the missingness for test results is higher than simply from individuals opting out of responding or not completing the full questionnaire.

The four tests included in the score were: a pairs test, which is a summary/total score for a memory test completed by participants which was transformed due to skewed positive tail, the fluid IQ test which is a problem solving test, the stroop mrt test in which participants have to report the colour that the word is written in, and the trails B which tests ability to join letters and numbers alternately. The four cognitive measures were all taken at 24 months after the individual was enrolled in the INTERVAL study. In total, there were 21,503 individuals who had complete data for all four cognitive measures. The general cognitive score itself is calculated

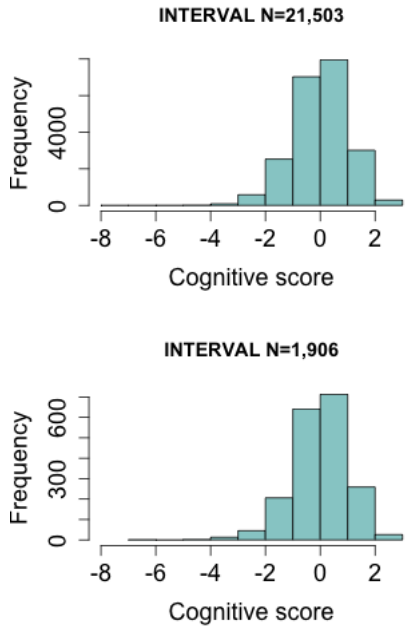


Figure 4.6: Distribution of cognitive scores in INTERVAL. **a**. Distribution of scores in all individuals who completed all four tests (N=21,503). **b**. Distribution in the subset of individuals who had data for polygenic scores, rare exome variants and cognitive scores (N=1,906).

by first performing a principal component analysis on the four test results. The cognitive score is then taken as the first unrotated principal component from this analysis, and it is positively correlated with fluid IQ and negatively correlated with inverse-normalised pairs test, stroop MRT test and trails B duration test results. Scores were normalised to a mean of zero and standard deviation of 1.

Figure 4.6 shows the distribution of these scores in all individuals who completed the four tests, and in the subset of individuals who had genotype and exome sequence data available. The distributions appear normal in the cohort and subset with some outliers.

Cognitive functioning is known to decrease with age (Deary and Batty, 2007). I therefore plotted the scores against the participants' age in the genotyped subset of 1,906 INTERVAL participants, and confirmed cognitive scores were negatively correlated with age (Pearson correlation -0.49, $P=6.7 \times 10^{-116}$) (Figure 4.7). I also

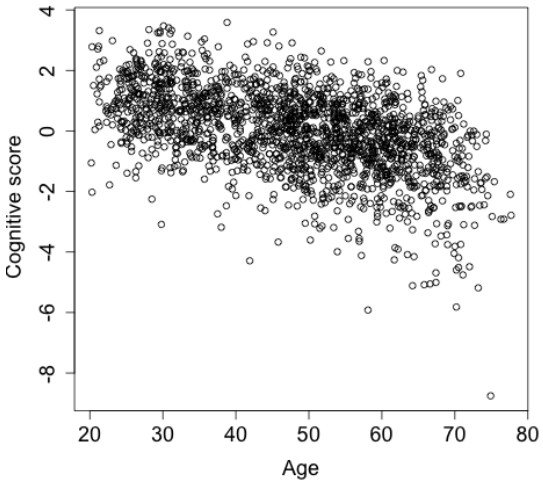


Figure 4.7: Cognitive scores are negatively correlated with age in INTERVAL (N=1,906). Pearson correlation -0.49, P=6.7×10^{-116}.

include age^2 as a covariate in the downstream analyses, to account for a possible non-linear relationship between age and cognition.

4.4.6 Power calculations

Power for detecting a difference in means

I carried out a power calculation to test our power to detect a significant difference in polygenic scores between individuals with (N=415) and without (N=1,491) LoF or missense variants. For the calculation I used software G*Power. Using this, I estimated power as a function of sample size, given three different values for the difference in polygenic score means. Since the polygenic scores were normalised, I assumed a variance of 1 in both groups, and calculated power at a range of sample sizes, assuming the same ratio of individuals with/without rare variants as we observe in our cohort (ratio=0.28).

Power for detecting an interaction

I carried out a power calculation to test what power we had to detect a significant interaction term between common and rare variants in the regression on cognitive scores in INTERVAL. To get an estimated effect size for an interaction term in a model that includes all the other independent variables, I used the partial r^2 for the interaction. Partial r^2 estimates the proportion of residual variation in the dependent variable explained by an independent variable, after the dependent variable has been regressed on all other variables. Partial r^2 essentially measures the additional explanatory power of the remaining independent variable. Partial r^2 can take any value between 0 and 1. I first estimated partial r^2 for the polygenic score, using the R function etasq(). I then used the software G*Power (v3.1) to obtain the effect size (β) corresponding to the partial r^2 and multiples of this.

4.5 Results

4.5.1 Assessing protective effect from common variants

To first confirm whether rare variants that we found were affecting the cognitive performance of individuals in INTERVAL (N=1,906), I performed a logistic regression of cognitive scores on rare variant status, controlling for age, age^2, sex and ancestry principal components. I found that having a LoF or missense variant in a brain-expressed pLI>0.9 gene (N=415 individuals with, N=1,491 without) was nominally significantly associated with a -0.10 SD change in cognitive scores (95% CI : [-0.007, -0.20], P=0.035). Similarly, having a LoF (N=240 individuals with, N=1,666 without) was associated with a -0.17 SD change in cognitive scores (95% CI : [-0.055, -0.29], P=0.004). Having confirmed that rare variants were affecting cognitive scores in INTERVAL, I proceeded with investigating the interplay between these rare variants and polygenic scores in the cohort.

Our hypothesis was that in a cohort such as INTERVAL, which consists of relatively healthy and cognitively functioning adults, individuals with rare deleterious variants

may be enriched for common variants that increase cognitive performance. In other words, we suspected that common variants could be having a protective effect against the effects of rare variants in this cohort. If the rare variants in individuals are deleterious with respect to their cognitive performance, the negative shift due to the rare variant is seemingly not sufficient to decrease the individual's chance of participating in the study. In other studies (e.g. UK Biobank) it has been noted that individuals who participate in studies tend to be more highly educated than average, implying that also their cognitive performance overall will be higher than average. We therefore reasoned that it was possible common variant polygenic scores were acting in a protective manner against the deleterious effects of rare variants within INTERVAL.

To assess whether polygenic scores were protective in INTERVAL, I compared the means of the distributions of polygenic scores for intelligence between individuals with and without rare variants. These distributions are shown in Figure 4.8. I performed a two-sample t-test, but neither the analysis comparing individuals with LoF+missense (N=415 individuals) or LoF-only (N=240) to those with no rare variants (N=1,491 and N=1,666 individuals, respectively) showed significant difference between the two groups (P=0.06 and P=0.11 respectively, one-tailed test). The trend in both analyses was that the mean of polygenic scores for individuals with rare variants was higher than in the group with no variants.

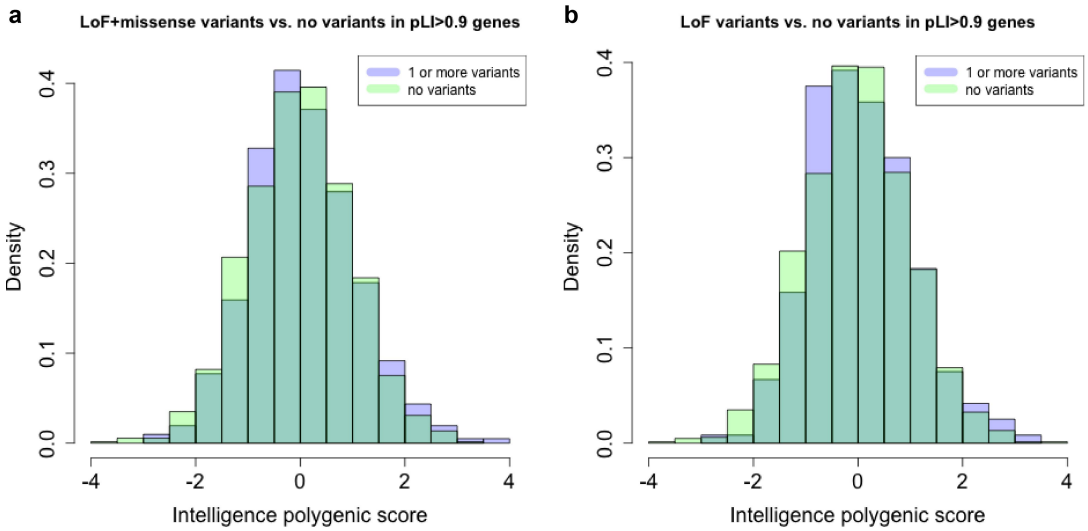


Figure 4.8: Distribution of intelligence polygenic scores in individuals with and without rare variants. **a**. Density distribution of intelligence polygenic scores in individuals with LoF or missense variants in pLI>0.9 genes (N=415) in purple, or without(N=1,491) in green, **b**, in individuals with LoF variants in pLI>0.9 genes (N=240) in purple, or without(N=1,666) in green.

We therefore wondered whether our small sample size meant we did not have enough power to detect a significant difference in the polygenic scores between these groups. As an estimate (beta) of the difference between the group means, I used the observed difference (beta=0.085). This showed the we had only 35% power to detect a significant difference between the means if the true difference was the observed mean. In this case, with the additional ∼6,000 samples from the WGS dataset, we would have good power (almost 90%) to detect a significant difference. I also tested our power to detect a difference in means, if the true difference was larger or smaller than we observe with our current samples. Figure 4.9 shows the power curves for these estimates. The larger difference estimate is the upper bound of the 95% confidence interval (beta=0.19) where the true difference lies. At our sample size, we had 80% power to detect a difference this large, so it is unlikely that the true difference is of this magnitude. Because the lower bound of the 95% confidence interval for the difference in means overlapped with zero, I took an arbitrary lower value of beta=0.01, which constitutes a rather small difference in standardised polygenic scores. It also appears that if the true difference in means

was this small, we would have almost no power to detect a significant difference, even at sample sizes in the hundreds of thousands.

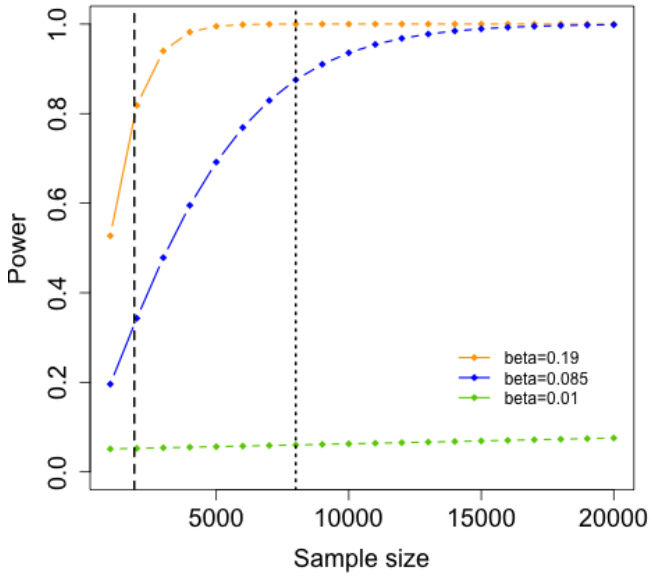


Figure 4.9: Power to detect a significant difference in mean polygenic score between individuals with and without a rare variant. The curves show our power to detect a significant difference ($P<0.05$) between the mean intelligence polygenic score in individuals with and without LoF or missense mutations in pLI>0.9 genes. The power is plotted as a function of sample size. In our analysis, we observe a difference in means (beta) of 0.085 (blue line). The orange line depicts our power to detect a significant difference if beta=0.19 (upper bound of 95% confidence interval for difference in means). The green line depicts our power to detect a significant difference that is no greater than beta=0.01. Long dashed line shows power at our current sample size 1,906 INTERVAL participants (with WES and cognitive data). Short dashed line shows power that we could obtain after adding approximately 6000 more samples with WGS and cognitive data from INTERVAL.

4.5.2 Joint contribution of rare and common variants to cognitive functioning

We next wanted to assess what the actual measured effects of polygenic scores and rare variants were on the cognitive performance of INTERVAL participants,

and whether there was an interaction between these two types of genetic variation. We were particularly interested in the latter question, which is whether the effect of polygenic risk scores on cognitive ability are the same in people with a rare deleterious variant as in people without. This type of an interaction effect between rare and common variants has previously been described for rare and common variant polygenic scores on educational attainment (Ganna et al., 2016).

We expected to find a significant positive association between a measure of general intelligence (cognitive scores) in INTERVAL and polygenic scores for intelligence and educational attainment as the two are genetically correlated. We also expected that deleterious variants may have a negative effect on the cognitive scores, although our sample size may be too small to detect this effect. Our main interest was whether we would also find a significant interaction between polygenic scores and rare variants on the cognitive scores. We were interested in seeing whether polygenic scores explained less (or more) variance in cognitive scores in the presence of a rare deleterious variant.

For our final dataset, we had 1,906 samples with polygenic scores, exome data and cognitive scores. To test the effect of common and rare variants, and their interaction on the cognitive scores in INTERVAL individuals, I performed a linear regression using R. I regressed the cognitive scores against each polygenic score (intelligence, educational attainment, schizophrenia, autism, intracranial volume, birth weight and height), their exome variant status (at least one variant/no variant, or a numerical count of the variants), the interaction of polygenic scores and rare variants, age, age^2, sex and ten ancestry principal components as covariates.

$$\text{Cognitive score} \sim \beta_{\text{prs}}\text{prs} + \beta_{\text{var}}\text{var} + \beta_{\text{prs}}\text{prs} * \beta_{\text{var}}\text{var} + \beta_{\text{age}}\text{age} + \beta_{\text{age}^2}\text{age}^2 + \beta_{\text{sex}}\text{sex} + \beta_{\text{PC1}}\text{PC1} + ... + \beta_{\text{PC10}}\text{PC10} + \varepsilon$$

Where prs = polygenic score, var = rare variant, ε = error. For rare variants, we had done filtering on the exome data to include only variants expressed in fetal brain and that had a pLI>0.9. As a first pass analysis, I chose variants that passed similar variant filters as what they authors of Ganna et al. used in their study. This first analysis included rare LoF and missense variants in pLI>0.9 genes, but we had restricted to fetal brain-expressed genes earlier in our variant filtering pipeline.

Out of 1,906 participants, 415 had at least one such variant, and 1,491 had none. I fitted the regression model:

As perhaps expected, in each regression the variable age^2 was negatively associated with cognitive scores and explained the most variance in overall. When age^2 is included in the regression, the association between age and cognitive score (as shown in Figure 4.7) becomes non-significant (and is in a positive direction). This implies that age has a non-linear relationship with the cognitive score, and the effect of age on the cognitive score is lesser in older participants.

As the polygenic score in this regression, I tested all eight polygenic scores to find which ones were relevant for explaining variance in cognitive scores. Table 4.3 summarises the effect of each of these polygenic scores on the cognitive scores in INTERVAL (N=1,906) in the joint regression model. Only intelligence and educational attainment polygenic scores were significantly associated with cognitive scores in the combined regression model, with individuals with a higher polygenic score having higher cognitive scores. The effect sizes of these two polygenic scores were very similar to each other, with 0.14 SD change in cognitive score for each 1 SD unit change in intelligence polygenic scores ($P=7.8 \times 10^{-10}$), and 0.14 SD change in cognitive scores for each 1 SD change in educational attainment polygenic scores ($P=9.4 \times 10^{-11}$). I will therefore focus on analyses where I use these two polygenic scores. To compare the effect of polygenic scores and other variables such as rare variant status on the cognitive scores, I have plotted the effect sizes of these in Figure 4.10 for the analyses using intelligence polygenic scores, and in Figure 4.11 for analyses using the educational attainment polygenic scores.

Table 4.3: Association of eight polygenic scores with cognitive scores in the combined regression model. These results show association of each polygenic score with the cognitive score, in a combined analysis with rare variant status (LoF or missense), an interaction term and other covariates. The estimate describes a change in standardised cognitive scores for a 1 SD change in polygenic score.

Polygenic score	Estimate	Std. error	P
Neurodevelopmental disorder	-0.047	0.023	0.042
Intelligence	0.139	0.022	7.8×10^{-10}
Educational attainment	0.143	0.022	9.4×10^{-11}
Schizophrenia	-0.033	0.023	0.146
Autism	0.018	0.023	0.428
Intracranial volume	0.008	0.023	0.718
Birth weight	-0.001	0.023	0.956
Height	0.011	0.023	0.669

In the analysis incorporating intelligence polygenic scores, in addition to the significant association of the polygenic score with cognitive score, we also see a significant effect of rare LoF or missense variants (Figure 4.10 a). Having one or more LoF or missense variant decreased cognitive scores by -0.11 SD (P=0.018), although the 95% confidence intervals for this estimate are wide. I also performed a version of the regression analyses using the actual count of rare variants, and the results of this analysis were very similar to the binary model using $\geq$1 or none. There was no significant interaction between rare variants and polygenic scores in this analysis (P=0.76).

I then tested whether refining criteria for rare variants would have an impact on our results. I reran the regression analyses using only LoF variants (N=240 samples with $\geq$1 variant, N=1,666 no variant) in pLI>0.9 genes. As shown in Figure 4.10 b, the effect of rare LoF variants appeared to be stronger than in the LoF+missense analysis, with an effect size equivalent to -0.19 SD change in the cognitive score for individuals with at least one LoF (P=0.0013), but again with large confidence intervals. As before, we did not detect a significant interaction between common and rare variants on the cognitive scores (P=0.30).

For comparison, the results from the regression analyses using polygenic scores for educational attainment are shown in Figure 4.11. The results are very similar to the regression using intelligence polygenic scores: the rare variants analysis shows a similar trend to the previous analysis in that in the regression model, LoFs only have larger effect on the cognitive score than LoFs+missense variants, and that polygenic scores explain.

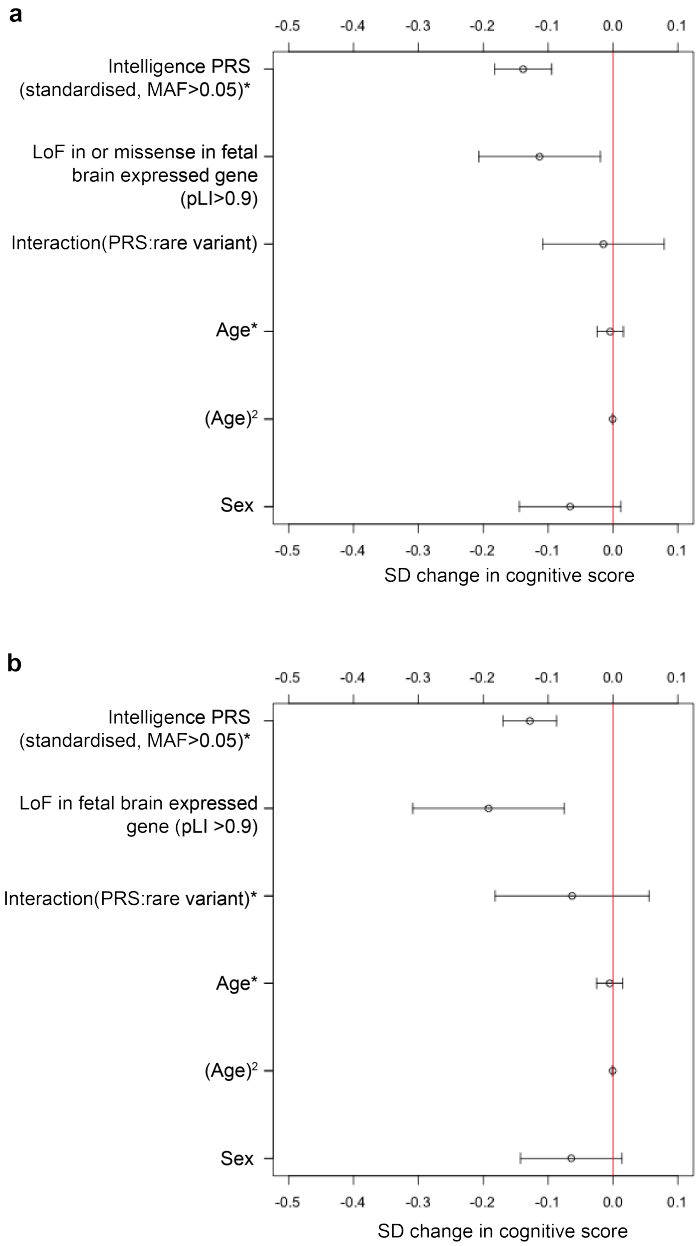


Figure 4.10: Rare and common variants affect cognitive scores in INTERVAL (N=1,906) (intelligence polygenic scores). Results from regression of cognitive scores on standardised polygenic score for intelligence, age, age^2, sex and 10 ancestry PCs (not shown) and **a**. deleterious LoF and missense variants in pLI>0.9 fetal expressed genes. **b**. LoF variants only. Sex variable labelled males=1, females=2. *Effect sizes were multiplied by -1 to allow for easier comparison to other effects.

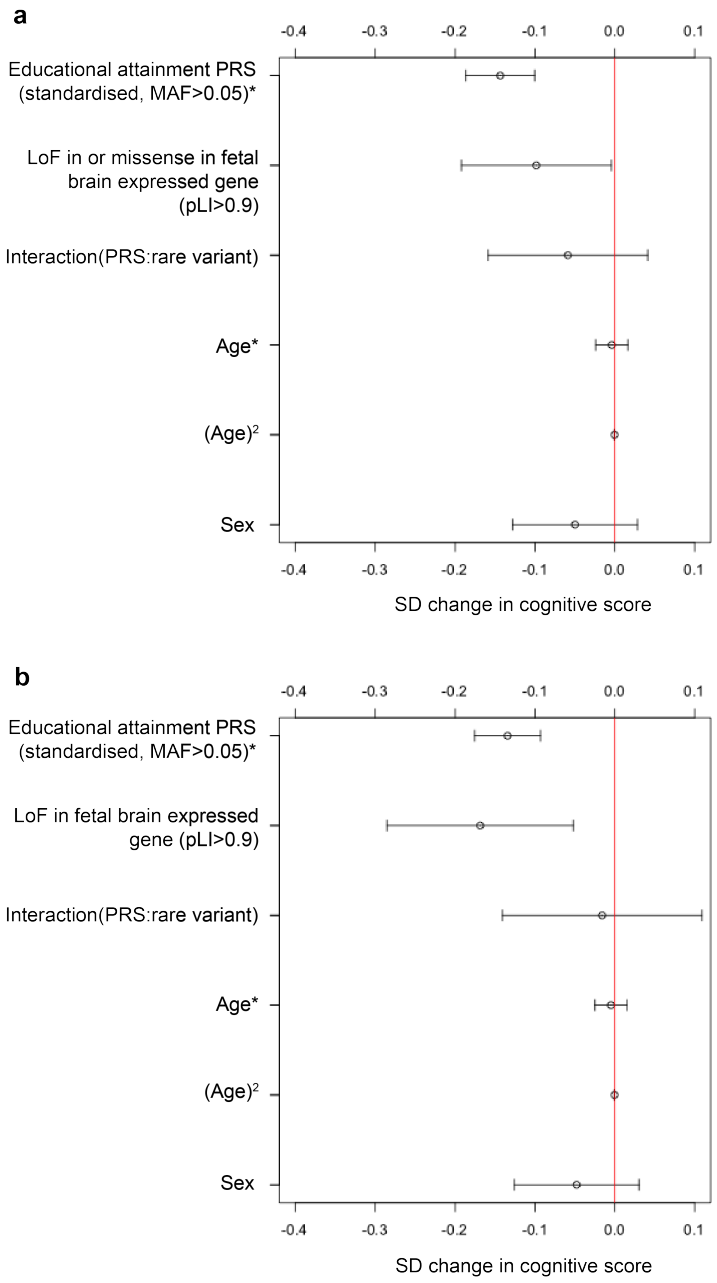


Figure 4.11: Rare and common variants affect cognitive scores in INTERVAL (N=1,906) (educational attainment polygenic scores). Results from regression of cognitive scores on standardised polygenic score for educational attainment, age, age^2, sex and 10 ancestry PCs (not shown) and **a**. deleterious LoF and missense variants in pLI>0.9 fetal expressed genes. **b**. LoF variants only. Sex variable labelled males=1, females=2. *Effect sizes were multiplied by -1 to allow for easier comparison to other effects.

Assessing power to detect a significant interaction

In our regression analysis, we did not detect a significant interaction between rare variants and the polygenic scores, unlike the (Ganna et al., 2016) study. One potential reason for this is lack of power, since the Ganna study had over five times as many samples we did in our study. We therefore wanted to assess our power to detect a significant interaction in our study. We would like to get an estimate of the relative magnitude of the interaction term effect and the polygenic score or rare variants from the Ganna et al. study. However, the effect size of the interaction was not reported in their analyses of educational attainment. This meant that we did not have any prior for how large an effect size the interaction might have if our data were comparable to the Ganna et al. study.

We therefore decided to test what power we had to detect an interaction at a range of effect sizes (Figure 4.12). We reasoned that the effect of the interaction would be no greater than that of the polygenic score alone, which had the strongest effect on cognitive scores in our data. I then plotted a power curve for a sample size range, assuming an effect size for the interaction that ranged from magnitude equivalent to the polygenic score beta (beta_prs), down to one 200th of the the beta_prs. The partial r^2 from both regression analyses on LoF + missense and LoF only in brain expressed pLI>0.9 genes was roughly the same (0.0241 and 0.0243 respectively), so I thus used beta_prs=0.025 as a starting point to to plot the power curves.

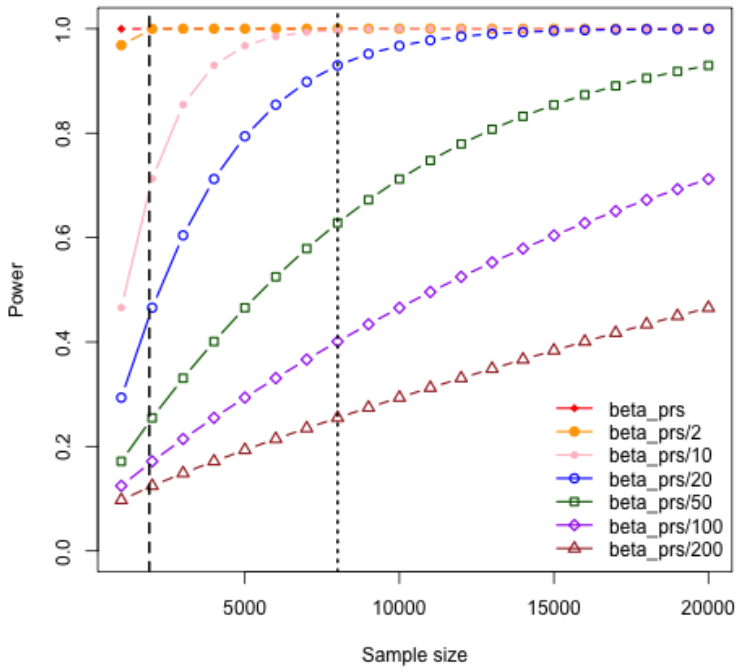


Figure 4.12: Power for detecting a significant interaction effect. Our power to detect a significant association for the interaction term between rare variants and the polygenic score in a regression analysis on cognitive scores, based on partial r^2 of the interaction term. The figure shows power for a range of values for partial r^2, with the maximum value being equal to partial r^2 of the intelligence polygenic score (beta_prs=0.025). Long dashed line shows power at our current sample size 1,906 INTERVAL participants (with WES and cognitive data). Short dashed line shows power that we could obtain after adding approximately 6,000 more samples with WGS and cognitiv e data from INTERVAL.

The power curves (Figure 4.12) show that we had good power in our analysis to detect interaction effect sizes equivalent to the effect size of polygenic scores in our regression model. We should also have ∼70% power to detect an effect size down to one 10th of a the polygenic score effect. This suggests that if a significant interaction existed in our cohort, the effect would likely be much smaller than that of the polygenic score alone.

The INTERVAL project is currently sequencing 12,000 samples using whole genome sequencing technology. Of these additional samples, approximately half will have

cognitive data available, and all will have been genotyped on the same DNA chip as the exome sequenced samples. The majority of WES and WGS samples do not overlap, and WES and WGS cohorts are currently being jointly called. This means that in the future we will be able to combine our current cohort of 1,906 with ∼6,000 additional samples. With these additional samples, we will have 60% power to detect an interaction with an effect size down to ∼1/50 of the effect of the polygenic scores.

4.6 Discussion

In this chapter, I set out to characterise the interplay between common and rare deleterious variants in the healthy population. I specifically look at their effects on cognitive ability, using data from healthy blood donors in the INTERVAL cohort. I found that there is no significant difference in polygenic scores for intelligence and educational attainment for individuals carrying deleterious rare variants in brain expressed, highly constrained genes versus those without such variants. We plan to reassess this with more samples in the near future. My second analysis found that common and rare variants both contribute to explaining variance in cognitive scores in INTERVAL. However, I did not observe an interaction effect between the two types of variants, although again our power was limited at the samples sizes I had. In addition, since we are looking at healthy individuals, it is likely that we would observe less penetrant variants that do not have as strong effects on the phenotype as would variants enriched in disease cohorts (Wright et al., 2018a).

When comparing to the results of our second analysis to the paper by Ganna et al, I found similarities but also differences in our data. Similarly to Ganna et al., I find a common variant and rare variant effect in our cohort. However, my analyses did not replicate the interaction between common and rare variants. From the power analysis, it seems evident that if there was an interaction that modified the effect of polygenic scores in the presence of a rare variant in INTERVAL, the effect would not be of the same magnitude as the effect of polygenic scores alone. Ganna et al. did not describe the effect size of the interaction in their combined

regression model. However, other studies that have found an interaction between common and rare variants have reported a very small effect from the interaction. For example, a paper by Barrett et al. (2009) studying type 1 diabetes, found that individuals carrying high risk HLA genotypes had a decreased risk from other loci associated with the disease compared to individuals who did not carry the HLA genotypes. With the additional INTERVAL samples with WGS data soon to be released, we should soon have better power to reassess the interaction between common and rare variants.

A difference between our results and those reported by Ganna et al. include that their paper found that the effect of common variants was approximately three times as large as the effect of having a rare variant, whereas in our study we find that the rare variant effect is almost as strong as 1SD change in polygenic scores in the LoF+missense analysis, or even stronger than the polygenic score in the LoF-only analysis. This difference may be the result of several factors, including real differences in the ascertainment of the individuals in the two studies, or the phenotypes measured (years of schooling versus general cognitive ability). For polygenic scores, Ganna et al. used the same P-value threshold for pruning variants for their educational attainment polygenic scores as we did for our score, but we used a better powered, larger GWAS variant effects to construct the scores. However, the genetic correlation between the two GWASs used is approximately 1 (analysis not shown), so therefore it is unlikely that there is a difference in the polygenic architecture of the two GWAS which the scores were constructed from. However, there are multiple differences in the rare variant filters we used. These differences include restricting to fetal brain-expressed genes and using newer annotation tools to filter LoFs and missense variants. In the near future, we plan to replicate the filtering used by Ganna et al. for better comparison.

As mentioned above, a major difference in our analysis compared to Ganna et al. was that we restricted our rare variants to those in fetal brain-expressed genes in our the combined regression, whereas Ganna et al. did not. For the analysis testing association between rare variants and educational attainment (Figure 2 in Ganna et al.), the authors show that when LoF and missense variants are split into variants in brain- and non-brain-expressed genes, the effect of LoFs is much stronger than

the effect of missense variants in both sets of genes. The negative effect of LoFs in brain-expressed genes was very close to the effect size of polygenic scores in their combined regression, which is in effect what we see in our analysis of LoF variants and polygenic scores. Variants in non-brain expressed pLI>0.9 genes, on the other hand, did not have a significant effect on educational attainment. It therefore seems likely that in their combined analysis, both not restricting to brain-expressed genes and including missense mutations, the authors may have diluted the effects of rare variants on educational attainment. In addition, their combined analysis included rare CNVs, which we do not have data for. However, our future analyses will include these data for INTERVAL, as CNV calling is currently underway.

Overall, these analyses show that even in a relatively small cohort, we can find significant genetic modifiers of general cognitive performance in the INTERVAL cohort. Further work will be required to carry out more extensive analyses into the interplay between common and rare variants, and hopefully with a boost in sample size we will be able to detect more subtle genetic effects. Expanding this analysis framework to larger datasets such as the UK Biobank (who have 50,000 exomes to be released), which has data on cognition but also educational attainment and potentially other relevant phenotypes will be of great interest.

Chapter 5

Discussion and future directions

5.1 Common variants contribute to neurodevelopmental disorders

In chapters 2 and 3, I have described the largest GWAS to date of rare disorders which had been presumed Mendelian. Studying these disorders using tools from complex trait genetics field was challenging, particularly because our patient cohort comprised of individuals with extremely heterogeneous phenotypes, and therefore also likely different genetic aetiologies. Unsurprisingly (albeit initially disappointingly), we did not see any individual common variant signals our GWAS. However, when we looked closer we did find a significant overall contribution to the risk of these disorders that was attributable to inherited common genetic variation. This polygenic burden shared common variant effects with other neuropsychiatric and cognitive traits, which implied that there may be shared underlying biology between these disorders, and that some common variants on the one hand confer risk to common disease and at the same time increase risk of rare neurodevelopmental disorders.

One of the important aspects of our work was that the results were reproducible, as both the overall risk (over-transmission in trios) and the genetic overlap with other traits (polygenic scores in Australians) replicated in independent samples. These

findings justify further work in the field, as they imply that leveraging data for rare neurodevelopmental disorder patients from across the globe could be fruitful in furthering our understanding of the genetic architecture of these disorders. In addition, we found no significant differences in polygenic scores between DDD patients who had a diagnostic rare variant and those who we had not yet identified one for. This suggested that patients without a monogenic diagnosis may not be solely responsible for the polygenic contribution to neurodevelopmental risk that we observe.

5.2 Impact in the clinic

The findings from our study challenge the typical view in medical genetics that rare, severe neurodevelopmental disorders are simply single-gene disorders. These findings may encourage clinicians to consider the possibility of a polygenic contribution to a patient's disorder, or to particular phenotypes observed in the patient. In the future, clinicians could be made aware of the possibility of common variants contributing to disease, by providing them with informative polygenic scores. This could result in clinicians reassessing and attempting to differentiate which phenotypes are more likely to be associated with the monogenic disorder and which may be partially explained by inherited common variants. As an example, if the patient's short stature was partially explained by a very low polygenic score for the trait, and their parents were also short of stature, the indication could be that at least part of the patient's height phenotype was explained by common variants. To an extent, clinicians are already evaluating the possibility of a common variant contribution in clinic when they meet the parents of affected children in clinic. Clinicians will often be able to notice unusual characteristics in the parents, particularly if they are shared with the child. However, this is not always possible to do, and polygenic scores would provide a way of assessing the common variant contribution even when meeting the parents in person is not possible.

The incorporation of polygenic scores to a patients' clinical data is something to be considered carefully. Certainly for the DDD Study participants this would

be possible, since information on polygenic scores could be incorporated into DECIPHER database for clinicians to view. There are options to how the data could be presented. For example, one might choose to report which quartile or decile of the distribution of scores a patient lies within, with respect to the rest of the DDD cohort. Another possibility would be to compare scores against a representative reference population cohort, for scores relevant to neurodevelopmental disorders. One would then generate polygenic scores for the reference individuals, and assess where the patient lies with respect the reference distribution.

Availability of information about where a patient lies in the polygenic score distribution could be of interest for clinicians when they assess the genetic data for the patient, particularly if there is evidence that the parents might be affected to some extent, or if there are multiple affected siblings with different severities of symptoms. However, it should be ensured that if such data were available for clinicians to view, a clear explanation of the implications of the polygenic scores would be required e.g. what is the expected distribution of IQ, or height, for people in this decile of the distribution.

As a real life experiment for using polygenic scores together with other clinical data, the DDD study is considering employing a clinician to have a closer look at specific patients' data, where the phenotype does not fully match the monogenic diagnosis. This would involve first identifying such cases from the individuals for whom exome sequencing has yielded a diagnostic variant in a developmental disorder associated gene. If the patients' abnormal phenotypes included any growth, cognitive or neuropsychiatric symptoms that are known to be affected by common variants in the general population, we could then supply polygenic scores for these traits for that individual. The clinician could then assess whether any of the unexplained symptoms could in part be explained by common variants acting in that individual. This could serve as a pilot for incorporating polygenic scores into DECIPHER for DDD clinicians to access for their patients. If incorporating polygenic scores into clinical data proved feasible in practice, in the long run these could be used to adjust recurrence risk estimates for families who are given genetic counselling. Although it may be difficult to infer what the exact increase or decrease in risk for families would be, particularly without the parental genotypes, it could still

be somewhat informative in cases where the scores are at the extreme ends of the relevant distribution.

5.3 Expanding to other cohorts

As our study has demonstrated, different cohorts of neurodevelopmental disorder patients will likely have different overall genetic burden due to differences in ascertainment. However, there is likely some overlap between common variant effects to be found in various cohorts of patients. One of the next logical steps for future work in the field would be to expand these analyses to other cohorts of neurodevelopmental disorder patients, and to perform a large meta-analysis to better understand the underlying biology of the common variant component to these disorders. Based on the results from our genetic correlation analyses in DDD and polygenic scores in Australians, it seems likely that a substantial proportion of the SNP heritability we found is shared with cognitive functioning (intelligence) and related traits. We could therefore consider also including our analyses to other cohorts of specifically intellectual disability cases on top of mixed neurodevelopmental disorder patient cohorts.

The DDD Study is a rather unique cohort in the sense that the genotyping of patients with rare disorders was done systematically in batches of ∼1,000 trios and then a larger cohort consisting of the remainder of patients. This approach greatly reduces biases and produces cleaner data for analysis. In addition, the DNA chips used for DDD were regular genotyping chips that have been used in other GWAS studies. For this reason, we were also able to find suitable controls for our discovery GWAS of neurodevelopmental disorder risk. We have heard through personal communication of other cohorts where intellectual disability cases or other neurodevelopmental patients have been genotyped on a DNA chip. However, these data were generated mainly to search for large CNVs for diagnostic purposes. Therefore the DNA chips used were often older and the variants on these do not overlap well with modern GWAS chips. In addition, we found that most of the data consisted of singleton patients or patient-parent duos, and not complete trios,

making it impossible to do trio analysis. Therefore, the main limiting issue in using genotype data on patients from these older chips often is that no healthy controls were genotyped using them. Using these samples in a large meta-analysis may therefore prove challenging due to the biases introduced by genotyping cases and controls on completely different chips, as well as from imputing data with only a small number of overlapping variants between different datasets.

Another aspect to consider if embarking on an effort to create a large consortium for studying common variant effects in rare, severe and heterogeneous disorders, is how much such an effort would make an impact to the patients. The main purpose of GWAS in human disease is finding significantly associated loci in order to hone in on potential drug targets. With more patients, it is possible that we might find genome-wide significant hits. It seems likely though that even if we were to find such loci, many may be shared with risk for conditions such as intelligence, educational attainment or schizophrenia, for which there are already well-powered GWAS. In addition, severe childhood onset disorders tend to be largely incurable conditions, where only the symptoms can be managed through drugs and therapies. Efforts to finemap loci to find drug targets from association studies may not the most effective approach for severe neurodevelopmental disorder treatment. If the main goal for studying common variants in neurodevelopmental disorders was to refine recurrence risks, an alternative approach could be to use polygenic scores for other larger and better-power traits that share polygenic architecture with these disorders (such as educational attainment or schizophrenia scores). Nevertheless, if this type of approach could give some benefit to patients, such a path of work could be considered.

5.4 Issues of differential ancestry

One major caveat of the work presented in this thesis is that it only considers individuals with European ancestry. In our study, the genotyped DDD cohort included a few thousand patients of non-European ancestry, but with the lack of parental genotypes and suitable controls, we were unable to include these individuals

in our discovery GWAS. For the downstream analyses using genetic correlation and polygenic scores, we utilise summary statistics from previously published GWAS which have also been performed mainly using individuals of European-ancestry. This means that even if we had been able to include non-Europeans in our study, we would still likely have been unable to use these samples in our downstream analyses, at least using these methods. This is because the causal variants for diseases and traits are not necessarily tagged by the same variants in different populations due to differences in LD structure. It has previously been shown that polygenic scores constructed using summary statistics from a single-population do not translate well in other populations (Martin et al., 2017a; Weiner et al., 2017). The authors of Martin et al. (2017a) also showed that scores generated in a single population explained the most variance in a target population from the same ancestral background as the original study. The study found that for example, polygenic scores from a European GWAS predicted Europeans to be taller than West Africans, despite the fact that West Africans are phenotypically no shorter than Europeans. In order to better understand the genetic architecture of neurodevelopmental disorders globally, we would need to consider how to better design our studies to include more diverse populations.

5.5 Continuing the search for common variant modifiers in health and disease

In Chapter 4, I described analyses of rare and common genetic variation in the healthy blood donor cohort INTERVAL. We found that both common variant polygenic scores and rare deleterious variants (particularly LoFs) in highly constrained genes affected the general intelligence scores in this cohort. Our sample size was too small to conclusively say whether polygenic scores were acting in a protective manner in individuals carrying these rare variants. However, our results suggest that further analyses are warranted once the extra ∼6,000 samples (individuals with cognitive data) from whole sequencing are ready.

The question of penetrance in genes, specifically those associated with known developmental disorders, is of great interest to the DDD analysis group. Current work undertaken by Kaitlin Samocha has uncovered an enrichment in DDD patients of inherited rare LoF and missense variants in known developmental disorder genes and in highly constrained genes. This raises the question whether something in the parental genotypes is protecting carrier parents from expressing the disease phenotype. If we were able to generate genotype data for DDD parents (at least those carrying inherited rare variants), we could then try to assess whether the polygenic scores for parents are systematically more protective than those of their affected children.

Another example of possible analyses taking on investigating variable penetrance further, would be to assess whether common variants are contributing to the phenotypes in DDD patients by affecting the expression of developmental disorder genes. A recent paper by Castel et al. (2018) showed that in the general population, deleterious variants were depleted on highly expressed haplotypes, decreasing their penetrance. Conversely, in individuals with disease, these variants were more likely to be on the more highly-expressed haplotype. To learn about haplotype expression in the DDD, one could identify eQTLs for genes associated with developmental disorders, and assess the expression of the haplotype that the diagnostic variants recide on. For *de novo* variants, one could expect to find that patients with more severe phenotypes have deleterious variants on the more highly expressed haplotype. With inherited or recessive variants the picture may be more complicated, as the unaffected parents carrying rare LoFs or deleterious missense mutations will have their rare variant on the same haplotype as the patient. However, in these cases one potential explanation could be to do with a shift in the relative expression of the haplotype with versus without the deleterious rare variant between unaffected parents and affected children. More analyses utilising data from the DDD could potentially yield interesting information on mechanisms by which common variants may (or may not) be playing a role in severe neurodevelopmental disorders and in the healthy population.

On the whole, we previously knew that genetic contributors to rare severe neurodevelopmental disorders included deleterious variants (inherited and *de novo*) within

the protein-coding regions (Deciphering Developmental Disorders Study, 2017; Martin et al., 2017b), in splice-site regions (Lord et al., 2018) and in regulatory elements (Short et al., 2018) of genes associated with these disorders. Here, I have described work that has uncovered a new contributor to our understanding of the genetic architecture of these severe disorders.

Appendix A

Partitioned SNP heritability

Results from partitioned SNP heritability analyses for discovery neurodevelopmental disorder risk GWAS

Table A.1: SNP heritability for discovery neurodevelopmental disorder risk GWAS, partitioned by cell type groups. LDSC baseline model was used to estimate enrichment of neurodevelopmental disorder discovery GWAS SNP heritability in cell type groups. Enrichment is defined as the proportion of SNP heritability in the category divided by the proportion of SNPs in the category. Results are ordered by coefficient z-score for cell type groups. P-values are uncorrected, two-sided and from z-score distribution.

Cell type group	Proportion of SNPs	Proportion of h^2	Standard error (h^2)	Enrichment	Standard error (enrichment)	P-value (enrichment)	Coefficient	Standard error (coefficient)	z-score (coefficient)
CNS	0.149	0.616	0.241	4.140	1.622	0.025	9.5×10^{-8}	4.4×10^{-8}	2.142
DDG2P	0.063	0.175	0.071	2.787	1.133	0.064	4.0×10^{-8}	2.5×10^{-8}	1.598
Cardiovascular	0.111	0.489	0.277	4.401	2.489	0.142	8.1×10^{-8}	8.1×10^{-8}	1.004
Other	0.203	0.740	0.324	3.652	1.600	0.068	6.1×10^{-8}	6.2×10^{-8}	0.997
GI	0.168	0.529	0.303	3.154	1.807	0.214	3.7×10^{-8}	5.9×10^{-8}	0.627
Connective or Bone	0.115	0.340	0.236	2.956	2.050	0.313	1.9×10^{-8}	5.7×10^{-8}	0.341
Kidney	0.043	0.140	0.183	3.288	4.284	0.592	2.9×10^{-8}	1.0×10^{-7}	0.290
Liver	0.072	0.202	0.186	2.804	2.579	0.477	1.6×10^{-8}	6.0×10^{-8}	0.268
Skeletal and Muscle	0.104	0.228	0.216	2.193	2.082	0.561	6.1×10^{-9}	7.3×10^{-8}	0.083
Immune	0.233	0.592	0.302	2.539	1.293	0.211	-6.3×10^{-9}	5.6×10^{-8}	-0.114
pLI$\geq$0.9	0.128	0.186	0.084	1.454	0.655	0.484	-2.1×10^{-9}	1.9×10^{-8}	-0.109
Adrenal or pancreas	0.094	0.132	0.229	1.411	2.452	0.866	-4.2×10^{-8}	7.1×10^{-8}	-0.595

Table A.2: SNP heritability for discovery neurodevelopmental disorder risk GWAS, partitioned by functional categories. LDSC baseline model was used to estimate enrichment of neurodevelopmental disorder discovery GWAS SNP heritability in overlapping functional categories. Enrichment is defined as the proportion of SNP heritability in the category divided by the proportion of SNPs in the category. Results are ordered by enrichment P value for functional categories. P-values are uncorrected, two-sided and from z-score distribution. The LDSC model adds 500bp regions around annotations and 100bp regions around ChIP-seq peaks to prevent upward bias in the estimate from enrichment in nearby regions. Studies used for functional categories in the LDSC baseline model are described in Finucane et al. (2015).

Functional category*	Proportion of SNPs in the category	Proportion of h^2 in the category	Standard error (h^2)	Ratio (enrichment)	Standard error (enrichment)	P-value (enrichment)
Conserved_LindbladToh	0.026	0.646	0.292	24.795	11.191	0.009
H3K4me1_Trynka	0.427	1.645	0.535	3.857	1.254	0.012
SuperEnhancer_Hnisz.extend.500	0.172	0.459	0.144	2.672	0.841	0.030
DHS_Trynka	0.168	1.706	0.811	10.167	4.833	0.035
DGF_ENCODE.extend.500	0.542	1.369	0.435	2.528	0.803	0.044
DHS_Trynka.extend.500	0.499	1.404	0.518	2.814	1.038	0.071
H3K4me3_peaks_Trynka	0.042	-0.571	0.400	-13.661	9.564	0.095
SuperEnhancer_Hnisz	0.168	0.399	0.150	2.371	0.893	0.101
PromoterFlanking_Hoffman	0.008	0.259	0.177	30.785	20.953	0.120
Repressed_Hoffman.extend.500	0.719	0.441	0.191	0.614	0.266	0.122
H3K4me3_Trynka.extend.500	0.255	0.701	0.321	2.743	1.255	0.142
DHS_peaks_Trynka	0.112	0.945	0.650	8.455	5.819	0.179
H3K4me1_Trynka.extend.500	0.609	0.989	0.268	1.623	0.440	0.183
TFBS_ENCODE.extend.500	0.343	0.944	0.458	2.748	1.333	0.190
H3K27ac_Hnisz	0.391	0.672	0.216	1.717	0.552	0.195
H3K27ac_PGC2	0.269	0.750	0.404	2.784	1.499	0.200
Intron_UCSC.extend.500	0.397	0.555	0.137	1.399	0.345	0.207
Intron_UCSC	0.387	0.574	0.177	1.482	0.458	0.264
Enhancer_Andersson	0.004	0.128	0.127	29.633	29.209	0.317
FetalDHS_Trynka	0.085	0.553	0.498	6.519	5.872	0.345
CTCF_Hoffman.extend.500	0.071	-0.180	0.286	-2.526	4.023	0.360
TSS_Hoffman.extend.500	0.035	0.168	0.165	4.812	4.744	0.407
TSS_Hoffman	0.018	0.177	0.196	9.696	10.756	0.407
H3K27ac_Hnisz.extend.500	0.423	0.644	0.266	1.524	0.630	0.411
Coding_UCSC.extend.500	0.065	-0.061	0.172	-0.948	2.663	0.462
Enhancer_Hoffman.extend.500	0.154	0.383	0.328	2.490	2.130	0.480
TFBS_ENCODE	0.132	0.479	0.49	3.615	3.702	0.482
Enhancer_Andersson.extend.500	0.019	-0.077	0.144	-4.013	7.558	0.498
Transcribed_Hoffman	0.345	0.121	0.363	0.350	1.051	0.551
Conserved_LindbladToh.extend.500	0.333	0.163	0.303	0.489	0.910	0.560
Enhancer_Hoffman	0.063	-0.137	0.349	-2.168	5.512	0.561
H3K27ac_PGC2.extend.500	0.336	0.523	0.335	1.557	0.998	0.571
H3K9ac_Trynka.extend.500	0.231	0.062	0.317	0.268	1.373	0.582
Promoter_UCSC.extend.500	0.039	0.113	0.140	2.914	3.616	0.587
WeakEnhancer_Hoffman.extend.500	0.089	0.244	0.290	2.740	3.261	0.591
Repressed_Hoffman	0.461	0.253	0.452	0.547	0.980	0.635
CTCF_Hoffman	0.024	-0.073	0.277	-3.044	11.626	0.727
UTR_3_UCSC	0.011	-0.031	0.121	-2.779	10.962	0.728

Table A.3: SNP heritability for discovery neurodevelopmental disorder risk GWAS, partitioned by functional categories (continued).

Functional category*	Proportion of SNPs in the category	Proportion of h^2 in the category	Standard error (h^2)	Ratio (enrichment)	Standard error (enrichment)	P-value (enrichment)
H3K4me3_Trynka	0.133	0.011	0.361	0.086	2.706	0.737
H3K4me1_peaks_Trynka	0.171	0.345	0.565	2.015	3.301	0.755
Transcribed_Hoffman.extend.500	0.763	0.676	0.285	0.886	0.373	0.758
DGF_ENCODE	0.138	0.336	0.668	2.443	4.856	0.766
FetalDHS_Trynka.extend.500	0.285	0.425	0.490	1.491	1.719	0.776
WeakEnhancer_Hoffman	0.021	0.091	0.268	4.333	12.687	0.792
H3K9ac_Trynka	0.126	0.210	0.347	1.665	2.754	0.808
Promoter_UCSC	0.031	0.075	0.190	2.420	6.097	0.814
UTR_3_UCSC.extend.500	0.027	0.057	0.134	2.126	4.981	0.822
PromoterFlanking_Hoffman.extend.500	0.033	0.080	0.213	2.404	6.362	0.824
UTR_5_UCSC.extend.500	0.028	0.003	0.134	0.100	4.803	0.852
UTR_5_UCSC	0.005	-0.005	0.100	-1.005	18.434	0.914
Coding_UCSC	0.015	0.029	0.164	1.982	11.209	0.930
H3K9ac_peaks_Trynka	0.039	0.018	0.327	0.462	8.440	0.949

Appendix B

Data on UK population highest level of qualification achieved

UK census data 2011

Source: UK census 2011.

The highest level of qualification is derived from the question asking people to indicate all types of qualifications held. People were also asked if they held foreign qualifications and to indicate the closest equivalent.

There were 12 response options (plus 'no qualifications') covering professional and vocational qualifications, and a range of academic qualifications.

These are combined into five categories for the highest level of qualification, plus a category for no qualifications and one for other qualifications (which includes vocational or work related qualifications, and for foreign qualifications where an equivalent qualification was not indicated):

No Qualifications: No academic or professional qualifications

Level 1 qualifications: 1-4 O Levels/CSE/GCSEs (any grades), Entry Level, Foundation Diploma, NVQ level 1, Foundation GNVQ, Basic/Essential Skills

Level 2 qualifications: 5+ O Level (Passes)/CSEs (Grade 1)/GCSEs (Grades A*-C), School Certificate, 1 A Level/ 2-3 AS Levels/VCEs, Intermediate/Higher Diploma, Welsh

Baccalaureate Intermediate Diploma, NVQ level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First/General Diploma, RSA Diploma Apprenticeship

Level 3 qualifications: 2+ A Levels/VCEs, 4+ AS Levels, Higher School Certificate, Progression/Advanced Diploma, Welsh Baccalaureate Advanced Diploma, NVQ Level 3; Advanced GNVQ, City and Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma

Level 4+ qualifications: Degree (for example BA, BSc), Higher Degree (for example MA, PhD, PGCE), NVQ Level 4-5, HNC, HND, RSA Higher Diploma, BTEC Higher level, Foundation degree (NI), Professional qualifications (for example teaching, nursing, accountancy)

Other qualifications: Vocational/Work-related Qualifications, Foreign Qualifications (Not stated/ level unknown).

Table B.1: UK census data from 2011 on highest level of qualification achieved.

		England and Wales
All categories	Persons	45,496,780
No qualifications	Percentage	22.7
Level 1	Percentage	13.3
Level 2	Percentage	15.3
Apprenticeship	Percentage	3.6
Level 3	Percentage	12.3
Level 4 and above	Percentage	27.2
Other qualifications	Percentage	5.7

UK Labour Market statistics 2012

Source: Novis labour market statistics.

Qualifications data are only be available from the APS for calendar year periods, for example, Jan to Dec 2005. The variables show the total number of people who are qualified at a particular level and above, so data in this table are not additive. Separate figures for each NVQ level are available in the full Annual Population Survey data set (Query data).

The trade apprenticeships are split 50/50 between NVQ level 2 and 3. This follows ONS policy for presenting qualifications data in publications. Separate counts for trade apprenticeships can be obtained from the full APS data set (Query data). No Qualifications: No formal qualifications held. Other Qualifications: includes foreign qualifications and some professional qualifications. NVQ 1 Equivalent: e.g. fewer than 5 GCSEs at grades A-C, foundation GNVQ, NVQ 1, intermediate 1 national qualification (Scotland) or equivalent. NVQ 2 Equivalent: e.g. 5 or more GCSEs at grades A-C, intermediate GNVQ, NVQ 2, intermediate 2 national qualification (Scotland) or equivalent. NVQ 3 Equivalent: e.g. 2 or more A levels, advanced GNVQ, NVQ 3, 2 or more higher or advanced higher national qualifications (Scotland) or equivalent. NVQ 4 Equivalent And Above: e.g. HND, Degree and Higher Degree level qualifications or equivalent.

Notes: Level and % are for those aged 16-64. % is a proportion of resident population of area aged 16-64.

Table B.2: UK labour market data from 2012 on highest level of qualification achieved.

	UK	(%)
NVQ4 and above	13,744,100	34
NVQ3	6,885,900	17.1
NVQ2	6,792,400	16.8
NVQ1	4,892,900	12.1
Other	2,545,800	6.3
Trade apprenticeship	1,476,600	3.7
No qualification	4,028,300	10

List of Tables

List of Figures

Bibliography

Abbadi, N, C Philippe, M Chery, H Gilgenkrantz, F Tome, H Collin, D Theau, D Recan, O Broux, and M Fardeau (1994). "Additional case of female monozygotic twins discordant for the clinical manifestations of Duchenne muscular dystrophy due to opposite X-chromosome inactivation". *Am. J. Med. Genet.* 52.2, pp. 198–206.

Adams, Hieab H H, Derrek P Hibar, Vincent Chouraki, Jason L Stein, Paul A Nyquist, Miguel E Rentería, Stella Trompet, Alejandro Arias-Vasquez, Sudha Seshadri, et al. (2016). "Novel genetic loci underlying human intracranial volume identified through genome-wide association". *Nat. Neurosci.* 19.12, pp. 1569–1582.

Al-Mulla, F, JM Bland, D Serratt, J Miller, C Chu, and GT Taylor (2009). "Age-dependent penetrance of different germline mutations in the BRCA1 gene". *Journal of clinical pathology* 62.4, pp. 350–356.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*

Anderson, Carl A, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan (2010). "Data quality control in genetic case-control association studies". *Nat. Protoc.* 5.9, pp. 1564–1573.

Antonarakis, Stylianos E and Jacques S Beckmann (2006). "Mendelian disorders deserve more attention". *Nat. Rev. Genet.* 7.4, pp. 277–282.

Anttila, Verneri, Brendan Bulik-Sullivan, Hilary Kiyo Finucane, Raymond Walters, Jose Bras, Laramie Duncan, Valentina Escott-Price, Guido Falcone, Padhraig Gormley, et al. (2017). "Analysis of shared heritability in common disorders of the brain".

Ardlie, Kristin G, Leonid Kruglyak, and Mark Seielstad (2002). "Patterns of linkage disequilibrium in the human genome". *Nat. Rev. Genet.* 3.4, pp. 299–309.

Astle, William J, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, et al. (2016). "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease". *Cell* 167.5, 1415–1429.e19.

BOLT-LMM v2.3.2 User Manual (2018). `https://data.broadinstitute.org/alkesgroup/BOLT-LMM/`. Accessed: 2018-4-18.

Bamshad, Michael J, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure (2011). "Exome sequencing as a tool for Mendelian disease gene discovery". *Nat. Rev. Genet.* 12.11, pp. 745–755.

Bansal, Vikas, Marina Mitjans, Casper A P Burik, Richard Karlsson Linner, Aysu Okbay, Cornelius A Rietveld, Martin Begemann, Stefan Bonn, Stephan Ripke, et al. (2018). "Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia".

Barrett, Jeffrey C, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, et al. (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes". *Nat. Genet.* 41.6, pp. 703–707.

Beaulieu, Chandree L, Jacek Majewski, Jeremy Schwartzentruber, Mark E Samuels, Bridget A Fernandez, Francois P Bernier, Michael Brudno, Bartha Knoppers, Janet Marcadier, et al. (2014). "FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project". *Am. J. Hum. Genet.* 94.6, pp. 809–817.

Becker, Kevin G (2004). "The common variants/multiple disease hypothesis of common complex genetic disorders". *Med. Hypotheses* 62.2, pp. 309–317.

Bergbaum, Anne and Caroline Mackie Ogilvie (2016). "Autism and chromosome abnormalities-A review". *Clin. Anat.* 29.5, pp. 620–627.

Bezzina, Connie R, Julien Barc, Yuka Mizusawa, Carol Ann Remme, Jean-Baptiste Gourraud, Floriane Simonet, Arie O Verkerk, Peter J Schwartz, Lia Crotti, et al. (2013). "Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death". *Nat. Genet.* 45.9, pp. 1044–1049.

Bhatia, Gaurav, Alexander Gusev, Po-Ru Loh, Hilary Kiyo Finucane, Bjarni J Vilhjalmsson, Stephan Ripke, SCZ Working Group of the Psychiatric Genomics Cons, Shaun Purcell, Eli Stahl, et al. (2016). "Subtle stratification confounds estimates of heritability from rare variants".

Botstein, D, R L White, M Skolnick, and R W Davis (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms". *Am. J. Hum. Genet.* 32.3, pp. 314–331.

Botstein, David and Neil Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease". *Nat. Genet.* 33 Suppl, pp. 228–237.

Bowie, Christopher R and Philip D Harvey (2006). "Cognitive deficits and functional outcome in schizophrenia". *Neuropsychiatr. Dis. Treat.* 2.4, pp. 531–536.

Brainstorm Consortium, Verneri Anttila, Brendan Bulik-Sullivan, Hilary K Finucane, Raymond K Walters, Jose Bras, Laramie Duncan, Valentina Escott-Price, Guido J Falcone, et al. (2018). "Analysis of shared heritability in common disorders of the brain". *Science* 360.6395.

Brook, C G D, T Gasser, E A Werder, A Prader, and M A Vanderschueren-Lodewykx (1977). "Height correlations between parents and mature offspring in normal subjects and in subjects with Turner's and Klinefelter's and other syndromes". *Ann. Hum. Biol.* 4.1, pp. 17–22.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale (2015a). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". *Nat. Genet.* 47.3, pp. 291–295.

Bulik-Sullivan, Brendan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, et al. (2015b). "An atlas of genetic correlations across human diseases and traits". *Nat. Genet.* 47.11, pp. 1236–1241.

Bustamante, Carlos D, Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, Stephen Glanowski, David M Tanenbaum, Thomas J White, John J Sninsky, et al. (2005). "Natural selection on protein-coding genes in the human genome". *Nature* 437.7062, pp. 1153–1157.

Campbell, Catarina D, Elizabeth L Ogburn, Kathryn L Lunetta, Helen N Lyon, Matthew L Freedman, Leif C Groop, David Altshuler, Kristin G Ardlie, and Joel N Hirschhorn (2005). "Demonstrating stratification in a European American population". *Nat. Genet.* 37.8, pp. 868–872.

Castel, Stephane E, Alejandra Cervera, Pejman Mohammadi, François Aguet, Ferran Reverter, Aaron Wolman, Roderic Guigo, Ivan Iossifov, Ana Vasileva, and Tuuli Lappalainen (2018). "Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk". *Nat. Genet.* P. 1.

Chen, Sining and Giovanni Parmigiani (2007). "Meta-analysis of BRCA1 and BRCA2 penetrance". *J. Clin. Oncol.* 25.11, pp. 1329–1333.

Clarke, Geraldine M, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan (2011). "Basic statistical analysis in genetic case-control studies". *Nat. Protoc.* 6.2, pp. 121–133.

Consortium, International HapMap et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". *Nature* 449.7164, p. 851.

Cooper, David N, Michael Krawczak, Constantin Polychronakos, Chris Tyler-Smith, and Hildegard Kehrer-Sawatzki (2013). "Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease". *Hum. Genet.* 132.10, pp. 1077–1130.

Cruise, Sharon Mary, Lynsey Patterson, Chris R Cardwell, and Dermot O'Reilly (2015). "Large panel-survey data demonstrated country-level and ethnic minority variation in consent for health record linkage". *J. Clin. Epidemiol.* 68.6, pp. 684–692.

Cutting, Garry R (2010). "Modifier genes in Mendelian disorders: the example of cystic fibrosis". *Ann. N. Y. Acad. Sci.* 1214, pp. 57–69.

Deary, Ian J and G David Batty (2007). "Cognitive epidemiology". *J. Epidemiol. Community Health* 61.5, pp. 378–384.

Deciphering Developmental Disorders Study (2015). "Large-scale discovery of novel genetic causes of developmental disorders". *Nature* 519.7542, pp. 223–228.

– (2017). "Prevalence and architecture of de novo mutations in developmental disorders". *Nature* 542.7642, pp. 433–438.

Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury (2011). "A linear complexity phasing method for thousands of genomes". *Nat. Methods* 9.2, pp. 179–181.

Demontis, Ditte, Raymond K Walters, Joanna Martin, Manuel Mattheisen, Thomas Damm Als, Esben Agerbo, Rich Belliveau, Jonas Bybjerg-Grauholm, Marie Bækved-Hansen, et al. (2017). "Discovery Of The First Genome-Wide Significant Risk Loci For ADHD".

Devlin, B and K Roeder (1999). "Genomic control for association studies". *Biometrics* 55.4, pp. 997–1004.

Durbin, Richard (2014). "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)". *Bioinformatics* 30.9, pp. 1266–1272.

Eichler, Evan E, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau (2010). "Missing heritability and strategies for finding the underlying causes of complex disease". *Nat. Rev. Genet.* 11, p. 446.

Finucane, Hilary K, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, et al. (2015). "Partitioning heritability by functional annotation using genome-wide association summary statistics". *Nat. Genet.* 47.11, pp. 1228–1235.

Firth, Helen V, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter (2009). "DECIPHER: Database of

Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources". *Am. J. Hum. Genet.* 84.4, pp. 524–533.

Fisher, Ronald (1918). "The correlation between relatives on the supposition of Mendelian inheritance". *Trans. R. Soc. Edinb.* 52, pp. 399–433.

Fraser, F C and A D Sadovnick (1976). "Correlation of IQ in subjects with Down syndrome and their parents and sibs". *J. Ment. Defic. Res.* 20.3, pp. 179–182.

Fuchsberger, Christian, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, et al. (2016). "The genetic architecture of type 2 diabetes". *Nature* 536.7614, pp. 41–47.

GTEx Consortium (2013). "The Genotype-Tissue Expression (GTEx) project". *Nat. Genet.* 45.6, pp. 580–585.

Ganna, Andrea, Giulio Genovese, Daniel P Howrigan, Andrea Byrnes, Mitja Kurki, Seyedeh M Zekavat, Christopher W Whelan, Mart Kals, Michel G Nivard, et al. (2016). "Ultra-rare disruptive and damaging mutations influence educational attainment in the general population". *Nat. Neurosci.* 19.12, pp. 1563–1565.

Garcia-Alonso, Luz, Jorge Jiménez-Almazán, Jose Carbonell-Caballero, Alicia Vela-Boza, Javier Santoyo-López, Guillermo Antiñolo, and Joaquin Dopazo (2014). "The role of the interactome in the maintenance of deleterious variability in human populations". *Mol. Syst. Biol.* 10, p. 752.

Genovese, Giulio, Menachem Fromer, Eli A Stahl, Douglas M Ruderfer, Kimberly Chambert, Mikael Landén, Jennifer L Moran, Shaun M Purcell, Pamela Sklar, et al. (2016a). "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia". *Nat. Neurosci.* 19, p. 1433.

– (2016b). "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia". *Nat. Neurosci.* 19.11, pp. 1433–1441.

Gilissen, Christian, Jayne Y Hehir-Kwa, Djie Tjwan Thung, Maartje van de Vorst, Bregje W M van Bon, Marjolein H Willemsen, Michael Kwint, Irene M Janssen, Alexander Hoischen, et al. (2014). "Genome sequencing identifies major causes of severe intellectual disability". *Nature* 511.7509, pp. 344–347.

Grove, Jakob, Stephan Ripke, Thomas Damm Als, Manuel Mattheisen, Raymond Walters, Hyejung Won, Jonatan Pallesen, Esben Agerbo, Ole A Andreassen, et al. (2017). "Common risk variants identified in autism spectrum disorder".

Hästbacka, J, A de la Chapelle, I Kaitila, P Sistonen, A Weaver, and E Lander (1992). "Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland". *Nat. Genet.* 2.3, pp. 204–211.

Haworth, Simon, Ruth Mitchell, Laura Corbin, Kaitlin H Wade, Tom Dudding, Ashley Budu-Aggrey, David Carslake, Gibran Hemani, Lavinia Paternoster,

et al. (2018). "Common genetic variants and health outcomes appear geographically structured in the UK Biobank sample: Old concerns returning and their implications".

Hensman Moss, Davina J, Antonio F Pardiñas, Douglas Langbehn, Kitty Lo, Blair R Leavitt, Raymund Roos, Alexandra Durr, Simon Mead, TRACK-HD investigators, et al. (2017). "Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study". *Lancet Neurol.* 16.9, pp. 701–711.

Hill, W David, Ruben C Arslan, Charley Xia, Michelle Luciano, Carmen Amador, Pau Navarro, Caroline Hayward, Reka Nagy, David J Porteous, et al. (2018). "Genomic analysis of family data reveals additional genetic effects on intelligence and personality". *Mol. Psychiatry*.

Horikoshi, Momoko, Robin N Beaumont, Felix R Day, Nicole M Warrington, Marjolein N Kooijman, Juan Fernandez-Tajes, Bjarke Feenstra, Natalie R van Zuydam, Kyle J Gaulton, et al. (2016). "Genome-wide associations for birth weight and correlations with adult disease". *Nature* 538.7624, pp. 248–252.

Hugot, J P, P Laurent-Puig, C Gower-Rousseau, J M Olson, J C Lee, L Beaugerie, I Naom, J L Dupas, A Van Gossum, et al. (1996). "Mapping of a susceptibility locus for Crohn's disease on chromosome 16". *Nature* 379.6568, pp. 821–823.

Hulzen, Kimm J E van, Claus J Scholz, Barbara Franke, Stephan Ripke, Marieke Klein, Andrew McQuillin, Edmund J Sonuga-Barke, PGC ADHD Working Group, John R Kelsoe, et al. (2017). "Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis". *Biol. Psychiatry* 82.9, pp. 634–641.

International HapMap Consortium (2003). "The International HapMap Project". *Nature* 426.6968, pp. 789–796.

International Schizophrenia Consortium, Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder". *Nature* 460.7256, pp. 748–752.

Iossifov, Ivan, Brian J O'Roak, Stephan J Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A Stessman, Kali T Witherspoon, Laura Vives, et al. (2014). "The contribution of de novo coding mutations to autism spectrum disorder". *Nature* 515.7526, pp. 216–221.

Karczewski, Konrad J and Michael P Snyder (2018). "Integrative omics for health and disease". *Nat. Rev. Genet.* 19.5, pp. 299–310.

Karczewski, Konrad. *loftee*. `https://github.com/konradjk/loftee`. Accessed: 2018-9-2.

Kemperman, Martijn H, Lies H Hoefsloot, and Cor W R J Cremers (2002). "Hearing loss and connexin 26". *J. R. Soc. Med.* 95.4, pp. 171–177.

Khera, Amit V, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, et al. (2018). "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". *Nat. Genet.*

Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure (2014). "A general framework for estimating the relative pathogenicity of human genetic variants". *Nat. Genet.* 46.3, pp. 310–315.

Knies, Gundi and Jonathan Burton (2014). "Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys". *BMC Med. Res. Methodol.* 14, p. 125.

Koch, Linda (2014). "New insights into the genetic architecture of ASDs". *Nat. Rev. Genet.* 15, p. 781.

Kosmicki, Jack A, Kaitlin E Samocha, Daniel P Howrigan, Stephan J Sanders, Kamil Slowikowski, Monkol Lek, Konrad J Karczewski, David J Cutler, Bernie Devlin, et al. (2017). "Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples". *Nat. Genet.*

Kuchenbaecker, Karoline B, Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M Domchek, Mark Robson, Amanda B Spurdle, et al. (2017). "Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers". *J. Natl. Cancer Inst.* 109.7.

Kurki, Mitja I, Elmo Saarentaus, Olli Pietilainen, Padhraig Gormley, Dennis Lal, Sini Kerminen, Minna Torniainen-Holm, Eija Hamalainen, Elisa Rahikkala, Riikka Keski-Filppula, et al. (2018). "Contribution of rare and common variants to intellectual disability in a high-risk population sub-isolate of Northern Finland". *bioRxiv*, p. 332023.

Lee, James J, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, et al. (2018). "Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals". *Nat. Genet.*

Lee, Sang Hong, Naomi R Wray, Michael E Goddard, and Peter M Visscher (2011). "Estimating missing heritability for disease from genome-wide association studies". *Am. J. Hum. Genet.* 88.3, pp. 294–305.

Leeuwen, Marieke van, Jiska S Peper, Stéphanie M van den Berg, Rachel M Brouwer, Hilleke E Hulshoff Pol, René S Kahn, and Dorret I Boomsma (2009). "A genetic analysis of brain volumes and IQ in children". *Intelligence* 37.2, pp. 181–191.

Lek, Monkol, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, et al. (2016). "Analysis of protein-coding genetic variation in 60,706 humans". *Nature* 536.7616, pp. 285–291.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, et al. (2011). "A high-resolution map of human evolutionary constraint using 29 mammals". *Nature* 478.7370, pp. 476–482.

Liu, Jimmy Z and Carl A Anderson (2014). "Genetic studies of Crohn's disease: past, present and future". *Best Pract. Res. Clin. Gastroenterol.* 28.3, pp. 373–386.

Liu, Jimmy Z, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, et al. (2015). "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations". *Nat. Genet.* 47.9, pp. 979–986.

Loh, Po-Ru, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Schizophrenia Working Group of Psychiatric Genomics Consortium, Teresa R de Candia, Sang Hong Lee, et al. (2015a). "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis". *Nat. Genet.* 47.12, pp. 1385–1392.

Loh, Po-Ru, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, et al. (2015b). "Efficient Bayesian mixed-model analysis increases association power in large cohorts". *Nat. Genet.* 47.3, pp. 284–290.

Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, et al. (2016). "Reference-based phasing using the Haplotype Reference Consortium panel". *Nat. Genet.* 48.11, pp. 1443–1448.

Loh, Po-Ru, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price (2018). "Mixed model association for biobank-scale data sets".

Lord, Jenny, Giuseppe Gallone, Patrick J Short, Jeremy F McRae, Holly Ironfield, Elizabeth H Wynn, Sebastian S Gerety, Liu He, Bronwyn Kerr, et al. (2018). "The contribution of non-canonical splicing mutations to severe dominant developmental disorders".

Luo, Yang, Katrina M de Lange, Luke Jostins, Loukas Moutsianas, Joshua Randall, Nicholas A Kennedy, Christopher A Lamb, Shane McCarthy, Tariq Ahmad, et al. (2016). "Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at ADCY7".

MacArthur, Daniel G and Chris Tyler-Smith (2010). "Loss-of-function variants in the genomes of healthy humans". *Hum. Mol. Genet.* 19.R2, R125–30.

MacArthur, Daniel G, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, et al. (2012). "A systematic survey of loss-of-function variants in human protein-coding genes". *Science* 335.6070, pp. 823–828.

Maher, Brendan (2008). "Personal genomes: The case of the missing heritability". *Nature* 456.7218, pp. 18–21.

Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Stephan Ripke, Naomi R Wray, Cathryn M Lewis, Steven P Hamilton, Myrna M Weissman, Gerome Breen, Enda M Byrne, Douglas H R Blackwood, et al. (2013). "A mega-analysis of genome-wide association studies for major depressive disorder". *Mol. Psychiatry* 18.4, pp. 497–511.

Malich, S, R H Largo, A Schinzel, L Molinari, and U Eiholzer (2000). "Phenotypic heterogeneity of growth and psychometric intelligence in Prader-Willi syndrome: variable expression of a contiguous gene syndrome or parent-child resemblance?" *Am. J. Med. Genet.* 91.4, pp. 298–304.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies". *Nat. Rev. Genet.* 11.7, pp. 499–511.

Martin, Alicia R, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny (2017a). "Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations". *Am. J. Hum. Genet.* 100.4, pp. 635–649.

Martin, Hilary C, Wendy D Jones, James Stephenson, Juliet Handsaker, Giuseppe Gallone, Jeremy F McRae, Elena Prigmore, Patrick Short, Mari Niemi, et al. (2017b). "Quantifying the contribution of recessive coding variation to developmental disorders".

Martin, Nicolas W, Sarah E Medland, Karin J H Verweij, S Hong Lee, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Grant W Montgomery, Margaret J Wright, and Nicholas G Martin (2011a). "Educational attainment: a genome wide association study in 9538 Australians". *PLoS One* 6.6, e20128.

– (2011b). "Educational attainment: a genome wide association study in 9538 Australians". *PLoS One* 6.6, e20128.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger,

Petr Danecek, et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". *Nat. Genet.*

Mina-Vargas, Angela, Lucía Colodro-Conde, Katrina Grasby, Gu Zhu, Scott Gordon, Sarah E Medland, and Nicholas G Martin (2017). "Heritability and GWAS Analyses of Acne in Australian Adolescent Twins". *Twin Res. Hum. Genet.* 20.6, pp. 541–549.

Moreno-De-Luca, Andres, David W Evans, K B Boomer, Ellen Hanson, Raphael Bernier, Robin P Goin-Kochel, Scott M Myers, Thomas D Challman, Daniel Moreno-De-Luca, et al. (2015). "The role of parental cognitive, behavioral, and motor profiles in clinical variability in individuals with chromosome 16p11.2 deletions". *JAMA Psychiatry* 72.2, pp. 119–126.

Morris, Andrew P, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segrè, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, et al. (2012). "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". *Nat. Genet.* 44.9, pp. 981–990.

Myers, Richard H (2004). "Huntington's disease genetics". *NeuroRx* 1.2, pp. 255–262.

Myers, Scott M, Chris Plauché Johnson, and American Academy of Pediatrics Council on Children With Disabilities (2007). "Management of children with autism spectrum disorders". *Pediatrics* 120.5, pp. 1162–1182.

Nance, Walter E (2003). "The genetics of deafness". *Ment. Retard. Dev. Disabil. Res. Rev.* 9.2, pp. 109–119.

Narasimhan, Vagheesh M, Karen A Hunt, Dan Mason, Christopher L Baker, Konrad J Karczewski, Michael R Barnes, Anthony H Barnett, Chris Bates, Srikanth Bellary, et al. (2016). "Health and population effects of rare gene knockouts in adult humans with related parents". *Science* 352.6284, pp. 474–477.

Natarajan, Pradeep, Gina M Peloso, Seyedeh Maryam Zekavat, May Montasser, Andrea Ganna, Mark Chaffin, Amit V Khera, Wei Zhao, Jonathan M Bloom, et al. (2017). "Deep-coverage whole genome sequences and blood lipids among 16,324 individuals".

Ng, Michael, Dipti Thakkar, Lorraine Southam, Paul Werker, Roel Ophoff, Kerstin Becker, Michael Nothnagel, Andre Franke, Peter Nürnberg, et al. (2017). "A Genome-wide Association Study of Dupuytren Disease Reveals 17 Additional Variants Implicated in Fibrosis". *Am. J. Hum. Genet.* 101.3, pp. 417–427.

Niemi, Mari E K, Hilary C Martin, Daniel L Rice, Giuseppe Gallone, Scott Gordon, Martin Kelemen, Kerrie McAloney, Jeremy McRae, Elizabeth J Radford, et al. (2018). "Common genetic variants contribute to risk of rare severe neurodevelopmental disorders". *Nature* 562.7726, pp. 268–271.

Noble, Kimberly G, Suzanne M Houston, Natalie H Brito, Hauke Bartsch, Eric Kan, Joshua M Kuperman, Natacha Akshoomoff, David G Amaral, Cinnamon S Bloss, et al. (2015). "Family income, parental education and brain structure in children and adolescents". *Nat. Neurosci.* 18.5, pp. 773–778.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, et al. (2008). "Genes mirror geography within Europe". *Nature* 456.7218, pp. 98–101.

O'Roak, B J, H A Stessman, E A Boyle, K T Witherspoon, B Martin, C Lee, L Vives, C Baker, J B Hiatt, et al. (2014). "Recurrent de novo mutations implicate novel genes underlying simplex autism risk". *Nat. Commun.* 5, p. 5595.

Okbay, Aysu, Jonathan P Beauchamp, Mark Alan Fontana, James J Lee, Tune H Pers, Cornelius A Rietveld, Patrick Turley, Guo-Bo Chen, Valur Emilsson, et al. (2016). "Genome-wide association study identifies 74 loci associated with educational attainment". *Nature* 533.7604, pp. 539–542.

Olszewski, Amy K, Petya D Radoeva, Wanda Fremont, Wendy R Kates, and Kevin M Antshel (2014). "Is child intelligence associated with parent and sibling intelligence in individuals with developmental disorders? An investigation in youth with 22q11.2 deletion (velo-cardio-facial) syndrome". *Res. Dev. Disabil.* 35.12, pp. 3582–3590.

Owen, Michael J, Michael C O'Donovan, Anita Thapar, and Nicholas Craddock (2011). "Neurodevelopmental hypothesis of schizophrenia". *Br. J. Psychiatry* 198.3, pp. 173–175.

Palla, Luigi and Frank Dudbridge (2015). "A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait". *Am. J. Hum. Genet.* 97.2, pp. 250–259.

Pardiñas, Antonio F, Peter Holmans, Andrew J Pocklington, Valentina Escott-Price, Stephan Ripke, Noa Carrera, Sophie E Legge, Sophie Bishop, Darren Cameron, et al. (2018). "Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection". *Nat. Genet.*

Pe'er, Itsik, Roman Yelensky, David Altshuler, and Mark J Daly (2008). "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants". *Genet. Epidemiol.* 32.4, pp. 381–385.

Petersen, Britt-Sabina, Broder Fredrich, Marc P Hoeppner, David Ellinghaus, and Andre Franke (2017). "Opportunities and challenges of whole-genome and -exome sequencing". *BMC Genet.* 18.1, p. 14.

Peyrot, Wouter J, Dorret I Boomsma, Brenda W J H Penninx, and Naomi R Wray (2016). "Disease and Polygenic Architecture: Avoid Trio Design and

Appropriately Account for Unscreened Control Subjects for Common Disease". *Am. J. Hum. Genet.* 98.2, pp. 382–391.

Plomin, R and I J Deary (2015). "Genetics and intelligence differences: five special findings". *Mol. Psychiatry* 20.1, pp. 98–108.

Plomin, Robert, John C DeFries, Gerald E McClearn, and Peter McGuffin (2008). *Behavioral Genetics (5th edition)*. Vol. 11. Worth Publishers, pp. 388–401.

Purcell, Shaun M, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia". *Nature* 506.7487, pp. 185–190.

Raynor, Pauline and Born in Bradford Collaborative Group (2008). "Born in Bradford, a cohort study of babies born in Bradford, and their parents: protocol for the recruitment phase". *BMC Public Health* 8, p. 327.

Reichenberg, Abraham, Martin Cederlöf, Andrew McMillan, Maciej Trzaskowski, Ori Kapra, Eyal Fruchter, Karen Ginat, Michael Davidson, Mark Weiser, et al. (2016). "Discontinuity in the genetic and environmental causes of the intellectual disability spectrum". *Proc. Natl. Acad. Sci. U. S. A.* 113.4, pp. 1098–1103.

Rietveld, Cornelius A, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, et al. (2013). "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment". *Science* 340.6139, pp. 1467–1471.

Rietveld, Cornelius A, Tõnu Esko, Gail Davies, Tune H Pers, Patrick Turley, Beben Benyamin, Christopher F Chabris, Valur Emilsson, Andrew D Johnson, et al. (2014). "Common genetic variants associated with cognitive performance identified using the proxy-phenotype method". *Proc. Natl. Acad. Sci. U. S. A.* 111.38, pp. 13790–13794.

Risch, N and K Merikangas (1996). "The future of genetic studies of complex human diseases". *Science* 273.5281, pp. 1516–1517.

Samocha, Kaitlin E, Jack A Kosmicki, Konrad J Karczewski, Anne H O'Donnell-Luria, Emma Pierce-Hoffman, Daniel G MacArthur, Benjamin M Neale, and Mark J Daly (2017). "Regional missense constraint improves variant deleteriousness prediction". *bioRxiv*, p. 148353.

Sazonovs, A and J C Barrett (2018). "Rare-Variant Studies to Complement Genome-Wide Association Studies". *Annu. Rev. Genomics Hum. Genet.*

Schaid, Daniel J, Wenan Chen, and Nicholas B Larson (2018). "From genome-wide associations to candidate causal variants by statistical fine-mapping". *Nat. Rev. Genet.* 19.8, pp. 491–504.

Scheuner, Maren T, Paula W Yoon, and Muin J Khoury (2004). "Contribution of Mendelian disorders to common chronic disease: opportunities for recognition,

intervention, and prevention". *Am. J. Med. Genet. C Semin. Med. Genet.* 125C.1, pp. 50–65.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). "Biological insights from 108 schizophrenia-associated genetic loci". *Nature* 511.7510, pp. 421–427.

Short, Patrick J, Jeremy F McRae, Giuseppe Gallone, Alejandro Sifrim, Hyejung Won, Daniel H Geschwind, Caroline F Wright, Helen V Firth, David R FitzPatrick, et al. (2018). "De novo mutations in regulatory elements in neurodevelopmental disorders". *Nature* 555.7698, pp. 611–616.

Singh, Tarjinder, Mitja I Kurki, David Curtis, Shaun M Purcell, Lucy Crooks, Jeremy McRae, Jaana Suvisaari, Himanshu Chheda, Douglas Blackwood, et al. (2016). "Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders". *Nat. Neurosci.* 19.4, pp. 571–577.

Singh, Tarjinder, James T R Walters, Mandy Johnstone, David Curtis, Jaana Suvisaari, Minna Torniainen, Elliott Rees, Conrad Iyegbe, Douglas Blackwood, et al. (2017). "The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability". *Nat. Genet.* 49.8, pp. 1167–1173.

Sniekers, Suzanne, Sven Stringer, Kyoko Watanabe, Philip R Jansen, Jonathan R I Coleman, Eva Krapohl, Erdogan Taskesen, Anke R Hammerschlag, Aysu Okbay, et al. (2017). "Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence". *Nat. Genet.* 49.7, pp. 1107–1112.

Sontheimer, Harald (2015). "Chapter 11 - Neurodevelopmental Disorders". *Diseases of the Nervous System*. Ed. by Harald Sontheimer. San Diego: Academic Press, pp. 319–347.

Spearman, C (1904). ""General Intelligence," Objectively Determined and Measured". *Am. J. Psychol.* 15.2, pp. 201–292.

Strachan, Tom and Andrew Read (2011). *Human Molecular Genetics*. Ed. by Elizabeth Owen. Vol. 4th. Garland Science, Taylor&Francis Group, LLC.

Swanson Jr, C L, R C Gur, W Bilker, R G Petty, and R E Gur (1998). "Premorbid educational attainment in schizophrenia: association with symptoms, functioning, and neurobehavioral measures". *Biol. Psychiatry* 44.8, pp. 739–747.

Talmud, Philippa J, Sonia Shah, Ros Whittall, Marta Futema, Philip Howard, Jackie A Cooper, Seamus C Harrison, Kawah Li, Fotios Drenos, et al. (2013). "Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study". *Lancet* 381.9874, pp. 1293–1301.

Teng, J and N Risch (1999). "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping". *Genome Res.* 9.3, pp. 234–241.

Theisen, Aaron and Lisa G Shaffer (2010). "Disorders caused by chromosome abnormalities". *Appl. Clin. Genet.* 3, pp. 159–174.

Trzaskowski, Maciej, Nicole Harlaar, Rosalind Arden, Eva Krapohl, Kaili Rimfeld, Andrew McMillan, Philip S Dale, and Robert Plomin (2014). "Genetic influence on family socioeconomic status and children's intelligence". *Intelligence* 42.100, pp. 83–88.

University of Essex Institute for Social and Economic Research (2014). *Institute for Social and Economic Research and National Centre for Social Research, Understanding Society: Waves 2 and 3 Nurse Health Assessment, 2010- 2012.* Vol. 3rd Edition. UK Data Service.

– (2018). *Understanding Society: Waves 1-7, 2009-2016 and Harmonised BHPS: Waves 1-18, 1991-2009*. Vol. 10th Edition. NatCen Social Research & Kantar Public.

Vilhjálmsson, Bjarni J, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. (2015). "Modeling linkage disequilibrium increases accuracy of polygenic risk scores". *The American Journal of Human Genetics* 97.4, pp. 576–592.

Visscher, Peter M and J Bruce Walsh (2017). "Commentary: Fisher 1918: the foundation of the genetics and analysis of complex traits". *Int. J. Epidemiol.*

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era — concepts and misconceptions". *Nat. Rev. Genet.* 9.4, pp. 255–266.

Weiner, Daniel J, Emilie M Wigdor, Stephan Ripke, Raymond K Walters, Jack A Kosmicki, Jakob Grove, Kaitlin E Samocha, Jacqueline I Goldstein, Aysu Okbay, et al. (2017). "Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders". *Nat. Genet.*

White, Karl R (1982). "The relation between socioeconomic status and academic achievement". *Psychol. Bull.* 91.3, pp. 461–481.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, et al. (2014). "Defining the role of common variation in the genomic and biological architecture of adult human height". *Nat. Genet.* 46.11, pp. 1173–1186.

Wray, Naomi R, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air,

et al. (2018). "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression". *Nat. Genet.*

Wright, Alan, Brian Charlesworth, Igor Rudan, Andrew Carothers, and Harry Campbell (2003). "A polygenic basis for late-onset disease". *Trends Genet.* 19.2, pp. 97–106.

Wright, Caroline F, Tomas W Fitzgerald, Wendy D Jones, Stephen Clayton, Jeremy F McRae, Margriet van Kogelenberg, Daniel A King, Kirsty Ambridge, Daniel M Barrett, et al. (2015). "Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data". *Lancet* 385.9975, pp. 1305–1314.

Wright, Caroline F, Ben West, Marcus Tuke, Samuel E Jones, Kashyap Patel, Thomas W Laver, Robin N Beaumont, Jessica Tyrrell, Andrew R Wood, et al. (2018a). "Assessing the pathogenicity, penetrance and expressivity of putative disease-causing variants in a population setting".

Wright, Caroline F, Jeremy F McRae, Stephen Clayton, Giuseppe Gallone, Stuart Aitken, Tomas W FitzGerald, Philip Jones, Elena Prigmore, Diana Rajan, et al. (2018b). "Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders". *Genet. Med.*

Wright, Caroline F, David R FitzPatrick, and Helen V Firth (2018c). "Paediatric genomics: diagnosing rare disease in children". *Nat. Rev. Genet.* 19.5, pp. 253–268.

Wright, Margaret J and Nicholas G Martin (2004). "Brisbane Adolescent Twin Study: Outline of study methods and research projects". *Aust. J. Psychol.* 56.2, pp. 65–78.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". *Nat. Genet.* 42.7, pp. 565–569.

Yang, Yaping, Donna M Muzny, Jeffrey G Reid, Matthew N Bainbridge, Alecia Willis, Patricia A Ward, Alicia Braxton, Joke Beuten, Fan Xia, et al. (2013). "Clinical whole-exome sequencing for the diagnosis of mendelian disorders". *N. Engl. J. Med.* 369.16, pp. 1502–1511.

Zelst-Stams, Wendy A van, Hans Scheffer, and Joris A Veltman (2014). "Clinical exome sequencing in daily practice: 1,000 patients and beyond". *Genome Med.* 6.1, p. 2.

Zheng, Jie, A Mesut Erzurumluoglu, Benjamin L Elsworth, John P Kemp, Laurence Howe, Philip C Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, et al. (2017). "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of

summary level GWAS data for SNP heritability and genetic correlation analysis". *Bioinformatics* 33.2, pp. 272–279.

Zhu, Xiaolin, Anna C Need, Slavé Petrovski, and David B Goldstein (2014). "One gene, many neuropsychiatric disorders: lessons from Mendelian diseases". *Nat. Neurosci.* 17.6, pp. 773–781.

Zhu, Yizhou, Cagdas Tazearslan, and Yousin Suh (2017). "Challenges and progress in interpretation of non-coding genetic variants associated with human disease". *Exp. Biol. Med.* 242.13, pp. 1325–1334.