Investigating penetrance of rare genetic variants using population cohorts

Submitted by Rebecca Kingdom
to the University of Exeter as a thesis for the degree of
Doctor of Philosophy in Medical Studies
June 2023

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Abstract

The same genetic variant found in different individuals can cause a spectrum of phenotypes, with some individuals showing no signs of any clinical illness, and some displaying severe illness. Variants that cause this can be said to show incomplete penetrance, where the related genotype either causes clinical disease or not, or they can be said to display variable expressivity, in which the clinical symptoms can vary across a spectrum. Incomplete penetrance and variable expressivity are both thought to be influenced by a large number of factors, including genetic modifiers, epigenetics, and environmental factors.

Many thousands of genetic variants have been identified as causal of monogenic disorders, mostly determined through small clinical studies, and thus the penetrance and expressivity of these variants may be overestimated when compared to their effect in the general population. With the wealth of population cohort data currently available, the penetrance and expressivity of such genetic variants can be investigated across a much wider contingent, potentially helping to reclassify variants that were previously thought to be completely penetrant.

This thesis aims to investigate the penetrance and expressivity of rare genetic variants in large population cohorts, and to potentially identify any genetic modifiers that could also affect the phenotypic effect of these variants, including the presence of other rare variants, and the aggregation of small effect common variants. We show that putatively damaging variants in a large number of genes are present at a higher rate than previously expected in healthy population cohorts. Furthermore, we show that as an aggregate, individuals who carry one of these variants have sub-clinical phenotypes related to the traits seen in clinical disease cases with variants in similar genes. We also show that the penetrance and expressivity of these rare variants can be modified by the presence of other rare variants in similar genes, and through common genetic variant, aggregated as polygenic scores. We then investigate methods of identifying rare non-coding variants that could be potential genetic modifiers.

Table of Contents

List of Tables

List of Figures

List of Appendices

Acknowledgements

I especially want to thank my supervisors Professor Caroline Wright and Dr Michael Weedon at the University of Exeter for their continual support and encouragement throughout this project. It has been a great privilege to get to work within such an excellent group and I am incredibly grateful for their patience and inspiration throughout the last few years.

Furthermore, I would like to thank the wider Exeter common and rare variant groups, especially Dr Robin Beaumont for support with REGENIE single gene testing, and Professor Timothy Frayling for their ideas and encouragement. Furthermore, Dr Marcus Tuke, for performing the CNV calling in UKB.

I would also like to thank the University of Exeter for the funding and for the chance to work on this project. Similarly, I would also like to greatly thank the Turing Institute at Exeter, and Dr Andrew Wood for the funding and opportunity to investigate machine learning options in genomics.

I thank all the patients, academics, and staff involved in the organisation and development of the UK Biobank project. Research such as this would not be possible without the time and support of everyone involved in this study.

Finally, thank you to my family and friends who have supported me through my academic journey, especially my parents, and my ever-helpful friend Dr Marc Kvansakul.

Publications

1) Incomplete penetrance and variable expressivity: from clinical studies to population cohorts.

   Rebecca Kingdom & Caroline F. Wright.

   Published in Frontiers in Genetics, 2022

   https://doi.org/10.3389/fgene.2022.920390

2) Rare genetic variants in dominant developmental disorder loci cause milder related phenotypes in the general population.

   Rebecca Kingdom, Marcus Tuke, Andrew R. Wood, Robin N. Beaumont, Timothy M. Frayling, Michael N. Weedon, Caroline F. Wright.

   Published in AJHG, 2022

   https://doi.org/10.1016/j.ajhg.2022.05.011

   This work was also presented at the international Curating the Clinical Genome conference in 2021.

3) Genetic modifiers of rare variants in monogenic developmental disorder loci.

   Rebecca Kingdom, Robin N. Beaumont, Andrew R. Wood, Michael N. Weedon, Caroline F. Wright.

   Manuscript under review Nature Genetics, 2023.

   Published Medrxiv, 2022.

   https://doi.org/10.1101/2022.12.15.22283523

Abbreviations

AC – Allele Count

ACMG – American College of Medical Genetics

ADHD – Attention Deficit Hyperactivity Disorder

ASD – Autism Spectrum Disorder

BMI – Body Mass Index

BPD – Bipolar Disorder

CADD – Combined Annotation Dependent Depletion

CNV – Copy Number Variant

DD – Developmental Disorder

DDD – Deciphering Developmental Disorders

DDG2P – Developmental Disorder Gene to Phenotype

EA – Educational Attainment

GoF – Gain of Function

GWAS – Genome Wide Association Study

HCM – Hypertrophic Cardiomyopathy

HES – Hospital Episode Statistics

ID – Intellectual Disability

IGV – Integrative Genomics Viewer

Indel – Insertion/Deletion

LoF – Loss of Function

LOFTEE – Loss Of Function Transcript Effect Estimator

MAC – Minor Allele Count

MODY – Maturity Onset Diabetes of the Young

NDD – Neurodevelopmental Disorder

NGS – Next Generation Sequencing

NMD – Nonsense Mediated Decay

OR – Odds Ratio

PGS / PRS – Polygenic Score or Polygenic Risk Score

pLI – Probability of Loss of Function Intolerance

PTV – Protein Truncating Variant

REVEL – Rare Exome Variant Ensemble Learner

RME – Random Monoallelic Expression

SCZ – Schizophrenia

SNP – Single Nucleotide Polymorphism

SNV – Single Nucleotide Variant

TAD – Topological Associated Domain

TDI – Townsend Deprivation Index

UKB – UK Biobank

uORF – Upstream Open Reading Frame

UTR – Untranslated Region

VAF – Variant Allele Frequency

VEP – Variant Effect Predictor

WES/WGS – Whole Exome Sequencing / Whole Genome Sequencing

WT – Wild Type

1. Chapter one: Aims and objectives

1.1 Introduction

Approximately 80% of all rare diseases are genetic in origin, and most of these are thought to be monogenic in nature[1]. Rare, deleterious variants are known to cause thousands of different genetic disorders in humans[2,3], and while the molecular basis of over 6000 monogenic diseases has been uncovered[4], with more than 350,000 pathogenic variants described[5], the underlying genetic basis of most rare disorders remains to be determined. With advances in next generation sequencing (NGS), and the increasing availability of whole exome/genome sequencing (WES/WGS), the study of genotype-phenotype relationships has become more widespread, as determining how genotype causes a phenotype is a fundamental step towards understanding disease pathology[6]. Protein-coding variants that are associated with disease phenotypes directly link DNA variation to altered protein function or dosage and to the phenotypic outcome, and so much of what we know about the genotype-phenotype relationship is based on the study of rare variants that cause monogenic disease[7]. Monogenic genotypes and phenotypes can be highly predictive for specific individual disorders, but sometimes this relationship can be complicated, with some damaging dominant monogenic variants not following expected Mendelian inheritance patterns[8]. Individuals with the same genotype can display distinctly different clinical phenotypes[9–12], including being clinically asymptomatic (i.e. incompletely penetrant). There are currently gaps in translating how individual genomic variation affects phenotypic presentation, and how genetic variants exert their functional impact to cause disease.

1.2 Rare variant interpretation

The study of genetic disease has often been divided into rare monogenic forms of disease, and more common polygenic complex disorders[13]. Rare variants are generally defined as those with an allele frequency below 1%, although many known deleterious variants have a frequency in the general population far lower than that. Current evidence suggests that the causes of rare and complex disease may be more overlapping than previously thought, as the genetic variation present across the genome highlights the complexity underlying

phenotypic presentation. There are both rare variants in individual genes that cause monogenic forms of complex disease[14,15], as well as common variants that affect the severity of monogenic disease[9,16]. Such complexity makes investigating genotype-phenotype relationships more complicated, which is only exacerbated by erroneous variant associations due to study design problems[17]. Human genetic diversity displays considerable variability, with individual genomes differing from the reference at 4.1-5 million sites[18]. Although most variation is common and predicted to be functionally neutral[19], each individual has on average 85 heterozygous and 35 homozygous protein truncating variants (PTVs)[20]. Population cohort studies have shown that the average genome contains around 200 very rare coding variants (gnomAD frequency of <0.1%) per person[21] and 54 variants previously reported as disease-causing, including 7.6 rare non-synonymous coding variants in monogenic disease genes[20,22]. Variant interpretation is an ongoing challenge within diagnostic medicine, making understanding the phenotypic consequences of underlying genetic variation a key aim of genomics research.

1.3 Using population cohorts

While small clinical studies that are based on a specific presentation of disease can overestimate the penetrance of any rare variants identified, large population studies will tend to underestimate the penetrance of variants due to largely consisting of generally healthy individuals, the "healthy volunteer effect"[17,23]. However, large population cohorts can give us the ability to investigate previously defined 'highly penetrant' variants in healthy individuals, and identify variants that may have a much lower level of penetrance than previously suggested[24]. Furthermore, within rare variant studies, very large sample sizes are often needed to identify such variants that contribute towards disease[25]. Population cohorts increasingly consist of detailed clinical and biological information in addition to genetic data, and the aggregation of all this data gives us the ability to research gene-phenotype associations, and potentially identify variants behind disease mechanisms[26]. These large datasets of genomic information paired with deep phenotypic has given researchers the ability to identify and characterize genetic and phenotypic relationships[27].

1.31 UK Biobank

The UK Biobank (UKB) is a voluntary population-based cohort from the UK, with deep phenotypic data and genetic data for ~500,000 individuals aged 40-70 at the time of recruitment[28,29]. Participants provided a variety of information via self-report questionnaires, cognitive and anthropometric measurements, and linked medical information through Hospital Episode Statistics (HES) data, including ICD9 and ICD10 codes. This data currently includes whole exome sequencing on 450,000 individuals (data made available October 2021), and whole genome sequencing on 200,000 individuals, along with the hospital record data, medical data, self-report questionnaire results, and additional test data. Detailed sequencing and variant detection methodology for UKB is available at https://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170. The UKB resource was approved by the UK Biobank Research Ethics Committee and all participants provided written informed consent to participate[28]. While such a large and phenotypically defined cohorts gives us a great ability to investigate genetic relationships, as previously mentioned, population cohorts tend to be healthier than the average individual, and participants in UKB are known to be healthier and wealthier than the average individual in the UK population[23].

1.4 Monogenic developmental disorders

Developmental disorders are a collection of severe neurological conditions that manifest from birth or early childhood, and have been shown to be caused by rare deleterious variants in a large number of genes, or large CNVs that overlap these regions. Approximately 2-5% of children are born with major congenital malformations, or develop severe neurodevelopmental disorders during early childhood[30–33]. Developmental disorders are a heterogenous group of conditions that can affect brain development and function, and can result in issues with behaviour, language, motor functioning, and impaired cognition[34,35]. Many of these disorders are caused by dominant *de novo* variants in developmentally important genes[30], and have been identified by large scale studies such as the Deciphering Developmental Disorders (DDD) study[36], and this information has led to resources such as the developmental disorder gene to phenotype database (DDG2P)[37], which provides clinically curated information on genes and variants that are reported to be associated with related disorders[38].

Monogenic developmental disorders are an interesting collection of conditions in which to study penetrance for several reasons. Firstly, because they are extremely genetically heterogenous, large numbers of genes are linked with monogenic conditions and large numbers of pathogenic variants have been identified, making them statistically tractable[30]. Although they are also phenotypically heterogenous, developmental delay and intellectual disability are a very common part of many rare syndromes[39]. Secondly, their phenotypic effect occurs from early childhood (or before), so a phenotype should be apparent at almost any adult age, and would therefore be expected to have a lifelong effect on many cognitive or cognitive-related traits. Thirdly, although many of the genes in which these damaging variants cause disease have been discovered through small clinical cohorts, there are several large-scale clinical cohorts (such as the DDD study) in which genes and causal variants have been systematically evaluated using a more statistical approach[36]. Finally, although some examples of incomplete penetrance has previously been observed within families[40–42], many of these disorders were (until recently) believed to be fully penetrant, and therefore potentially pathogenic variants were not anticipated to be present in population cohorts such as gnomAD or UK Biobank. The presence of plausibly pathogenic variants in these 'healthy' cohorts therefore provides an excellent opportunity to study their likely pathogenicity, penetrance and expressivity in a clinically-unselected cohort.

1.5 Aims

The aims of this project were:

- To review our current understanding of penetrance and expressivity of rare genetic variants (Chapter 2)
- To explore the penetrance of predicted pathogenic rare variants in a population cohort using developmental disorders as an example (Chapter 3)
- To investigate potential genetic modifiers of penetrance in a population cohort (Chapter 4)

- To evaluate new analytic approaches to finding novel *cis*-modifiers, using a single gene as an example (Chapter 5)

2. Chapter two: Incomplete penetrance and variable expressivity

2.1 Introduction

The same genetic variant found in different individuals can cause a range of diverse phenotypes, from no discernible clinical phenotype to severe disease, even among related individuals. Such variants can be said to display incomplete penetrance, a binary phenomenon where the genotype either causes the expected clinical phenotype or it doesn't, or they can be said to display variable expressivity, in which the same genotype can cause a wide range of clinical symptoms across a spectrum. Both incomplete penetrance and variable expressivity are thought to be caused by a range of factors, including common variants, variants in regulatory regions, epigenetics, environmental factors, and lifestyle.

This chapter examines our current knowledge of the penetrance and expressivity of genetic variants in rare disease and across populations, as well as looking into the potential causes of the variation seen, including genetic modifiers, mosaicism, and polygenic factors, among others. We also consider the challenges that come with investigating penetrance and expressivity.

2.11 Incomplete penetrance and variable expressivity

A deleterious genotype should be no more prevalent in the population than the disease that it causes[43]. However, the same genetic variant can result in different disease presentations in different people, from clinically asymptomatic to severely affected, even among members of the same family[44]. The proportion of individuals who possess a particular genotype and exhibit the expected clinical symptoms is defined as the penetrance of that genotype[45,46]. If everyone with the genotype presents with clinical symptoms by a particular age then it is said to be fully penetrant, whereas if it falls below this it is said to exhibit reduced or incomplete penetrance. Genotype-phenotype relationships can also display variable expressivity, where the severity of the phenotype caused by the

genotype can vary among affected individuals[46] (**Table 2.1**); this differs from pleiotropy, where variants in the same gene can cause different, potentially unrelated phenotypes that may even be categorised as different diseases[47] (**Figure 2.1**). Although penetrance, expressivity, and pleiotropy are three distinct concepts, biological reality means that their overall effects often overlap, especially in population cohorts where it is difficult to identify the cause of the phenotypic diversity. Multiple distinct phenotypes, in aggregate, could either be classified as a single more severe phenotype or different disease subtypes. As these three are likely to be caused by overlapping or similar mechanisms[48], especially in genetically heterogenous conditions, we will discuss them together in this review.

Causal Gene	**Severe Phenotype**	**Milder Phenotype**
HOXD13	Synpolydactyly (extra fused digits)[49]	Short digits[50,51]
KCNQ4	Deafness[52]	Mild hearing loss[50]
SGCE	Myoclonus Dystonia[53]	Dystonia / Writer's cramp[50,54]
KRT16	Pachyonychia congenita[55]	Blistered Feet[50,56]
FLCN	Birt-Hogg-Dube Syndrome[57]	Mild fibrofolliculomas[50]
SFTPC	Lung Disease[58]	Abnormal lung diffusion capacity[50,59]
FBN1	Severe Marfan syndrome[60,61]	Mild Marfan phenotypes (tall, thin, slender fingers)[62]
ERCC4	Xeroderma pigementosum[63]	Higher likelihood of sunburn[17]
FLG	Ichthyosis vulgaris[64]	Eczema[17]
POLG	Childhood onset Alpers-Huttenlocher[65]	Deterioration of eye muscles[66]

Table 2.1: Examples of variable expressivity in monogenic diseases. Deleterious variants in these genes are known to cause a spectrum of phenotypes, from severe disease to mild subclinical effects.

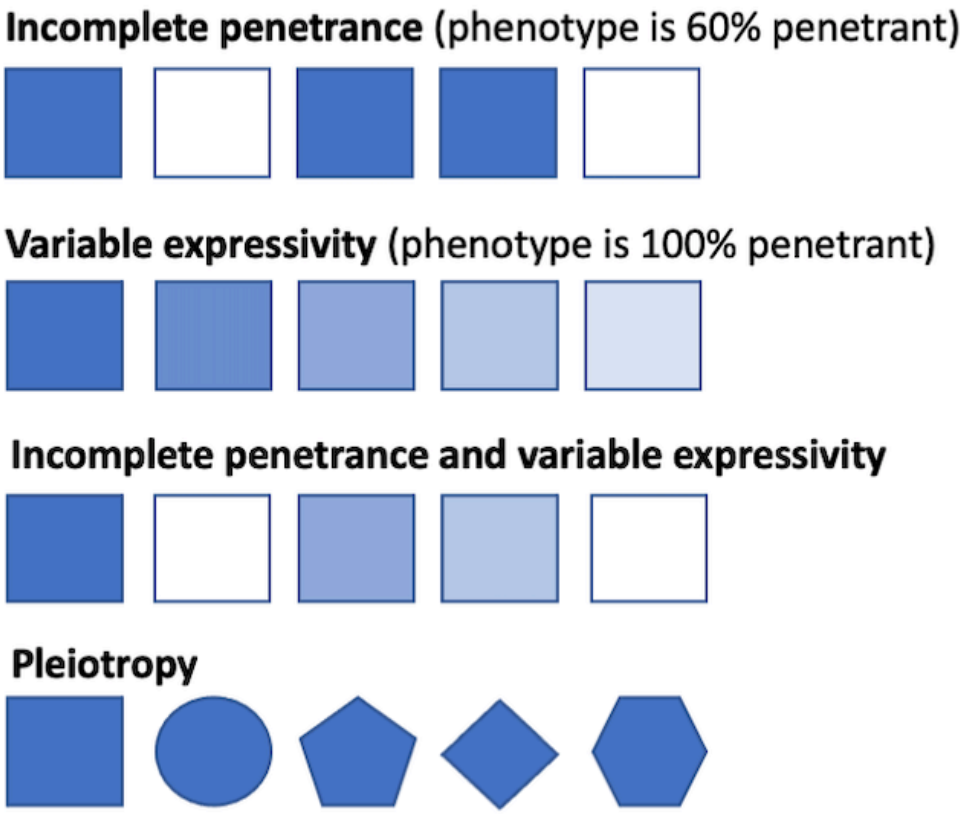
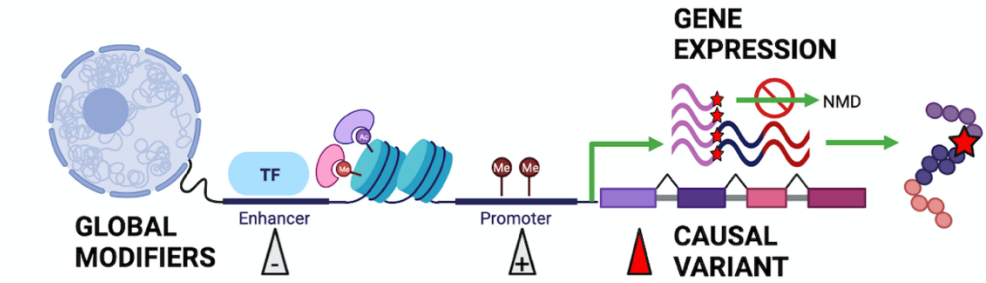


Figure 2.1. **Conceptual representation of penetrance, expressivity, and pleiotropy.** Squares represent individuals with the same genotype, with shaded squares indicating the individual displays the related phenotype, and non-shaded squares indicating the individual does not display the related disease phenotype. Line one shows incomplete penetrance, where 60% of the individuals display the related phenotype. Line two shows that all individuals display the related phenotype, from severe manifestations to milder presentations. Line three shows incomplete penetrance and variable expressivity, where the genotype varies both in severity of presentation, and in penetrance across the population. Line four shows pleiotropy, whereby different phenotypes are caused by variants (represented by different shapes) in one gene.

Incomplete penetrance can be observed in both dominant and recessive conditions. However, the cause of variability in genotype-phenotype correlations can be difficult to elucidate – phenotypic variation has been observed in mice with identical environmental and genetic backgrounds, including variability in lethality for knockout genes despite the introduction of identical variants[67]. Establishing that a variant identified is the sole (or primary) cause an individual's clinical phenotype can be difficult[68], which is an important concern when it comes to diagnosis and providing accurate genetic counselling, and such difficulties can lead to incorrect or delayed diagnosis[69]. The widespread presence of incomplete penetrance and variable expressivity through many overlapping mechanisms (**Figure 2.2**) can explain why apparently unaffected parents can pass on pathogenic variants to affected offspring[12], and why seemingly healthy individual's genomes can contain a large number of putatively damaging variants and yet not suffer any obvious adverse effects[70].

A



B

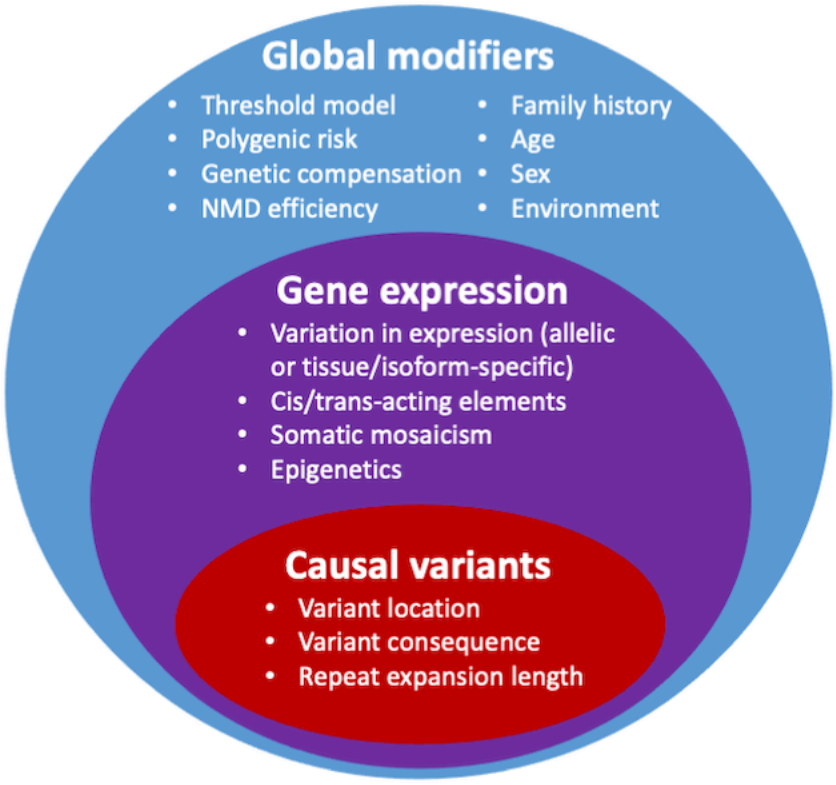


Figure 2.2. Factors affecting penetrance and expressivity. (a) Examples of different biological mechanisms that can affect the overall penetrance and expressivity of monogenic disease-causing genetic variants. Figure created with BioRender.com. **(b)** Summary of factors affecting penetrance and expressivity across the genome, from global modifiers that can have wide-ranging overall effects, to expression of the gene containing causal variants, to specific causal variants that have more distinctive effects.

While databases of clinically identified variants in affected individuals are useful for assessing pathogenicity[71], population-based datasets that include WES/WGS alongside phenotypic and medical information are increasingly important for investigating the penetrance and expressivity of these variants. Large population cohort studies have shown the occurrence of apparently pathogenic variants is much higher than previously estimated through small clinical or familial cohort studies[17,71,72], and their frequency highlights either the incomplete penetrance, variable expressivity, or misclassification of such variants. The existence of PTVs (protein truncating variants) in dosage-sensitive genes in healthy individuals also remains problematic when it comes to determining pathogenicity[73]. The potential for genomic technologies and WGS to detect individuals at risk of genetic disease is enormous, but incomplete penetrance and variable expressivity present a challenge for clinicians, especially when an incidental finding occurs without any prior clinical indication, leading to uncertainty over whether a clinical phenotype will develop, and if so, when. This problem is highlighted when testing unselected population cohorts, who may or may not have phenotypes of relevance to genomic findings at the point of testing. To understand how genetic disorders develop, we need to consider how deleterious variants interact with the rest of variation in the genome, and how variation can affect phenotypic presentation. This may also identify targets that help prevent disease progression[74]. The presence of putatively pathogenic variants in asymptomatic adults also highlights the possibility that there are disease-resistance mechanisms we can identify through the sequencing of general population cohorts.

2.12 Clinical versus population cohorts

Traditionally, rare pathogenic variants were identified in small phenotypically enriched clinical cohorts of individuals and families with similar monogenic disease. Population cohorts allow us to utilize the information from small clinical studies to investigate the penetrance of variants in the general ‘healthy’ population, where such severe monogenic phenotypes are likely to be depleted, as well as the potential to identify the causes of clinical heterogeneity. Ascertainment bias can occur with any study design, with volunteer population cohorts tending to be healthier than the average individual[23], and clinical cohorts tending to have more severe phenotypes. Estimates of the maximum

and minimum variant effect sizes across different ascertainment contexts are needed to avoid falsely predicting that a significant proportion of the healthy population are at risk for a monogenic condition[75]. The proportion of individuals affected and the average age of onset (i.e., age-dependent penetrance) can vary depending on ascertainment context (**Figure 2.3**). For example, individuals with putatively pathogenic variants in *HNF1A* and *HNF4A*, known for causing maturity onset diabetes of the young (MODY), develop diabetes significantly later or not at all when tested outside of the context of clinical referrals for suspected MODY[76].

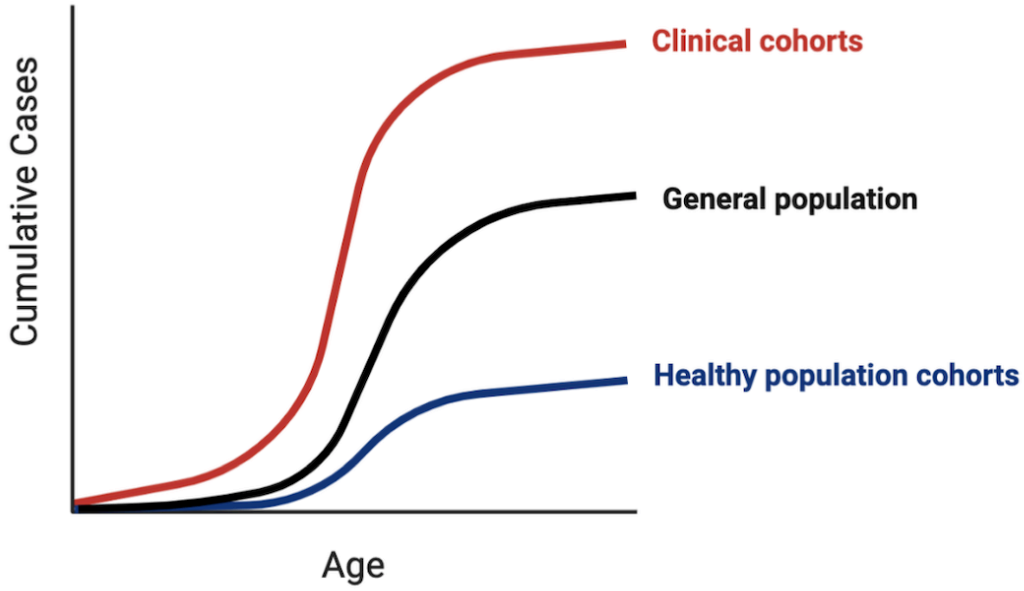


Figure 2.3. Penetrance in clinical versus population cohorts. Penetrance of genetic variants identified in clinical cohorts tends to be higher than the same variants identified in population cohorts, which can manifest as earlier disease onset, less severe disease, or a larger proportion of affected individuals. Due to inherent ascertainment biases in both types of cohorts, the penetrance of variants in the general unselected population is likely to lie somewhere in-between.

For almost all human genetic disease, individual variability in phenotype is influenced by background variation in the genome. As genetic testing becomes more widely available, both through healthcare systems[77] and direct-to-

consumer testing[78], our understanding of how genomic variation affects disease progression and prevalence becomes significantly more important, both for clinical utility[68] and for our functional understanding of disease[79]. Variation in the genome can predispose individuals to disease through traditional monogenic variants that disrupt physiological pathways and exert a large effect on phenotype, or through the accumulation of polygenic effects that involve many variants of small effect sizes in different pathways[80], or as is increasingly becoming clear, through their combined effect.

Within population cohorts, penetrance estimates for monogenic variant carriers average 60% or lower for most conditions[16], illustrating that there are many individuals who have apparently highly penetrant, pathogenic variants in known monogenic disease-causing genes who never develop the corresponding phenotype[81]. For example, one in 75 (1.3%) of healthy elderly individuals in the APSREE trial carried a previously identified pathogenic variant, including in Lynch Syndrome and familial hypercholesterolemia genes, without having the phenotype associated[72]. These cases demonstrate that carrying such pathogenic variants does not necessarily equate to disease and that other mechanisms may contribute towards the protection of human health, including genetic modifiers that 'rescue' individuals from a disease phenotype.

2.2 Causal variants

2.21 Variant location and consequence

For genetically heterogenous monogenic diseases, the penetrance and expressivity can vary between different genes or variants, with the same phenotype potentially caused by numerous different variants across multiple genes[82]. Even within the same gene, some deleterious variants in known monogenic disease genes may exhibit complete penetrance, while others show incomplete or low penetrance. Variation can be due to functional redundancy of genes, or the location and type of variant, with missense and PTVs in the same gene often causing different phenotypes. For example, hereditary angioedema can show great phenotypic diversity, even among members of the same family, and individuals with missense variants in *SERPING1* typically display a milder and later onset of disease compared to patients with PTVs[83]. In contrast,

missense variants in *BMPR2* cause earlier and more severe pulmonary hypertension compared to PTVs in the same gene[84].

Pathogenic PTVs typically cause disease through loss of function (LoF) due to degradation of the RNA by nonsense mediated decay (NMD)[85]. NMD is an mRNA surveillance pathway that recognizes and degrades damaged mRNA transcripts that would produce misfolded or shortened proteins that can accumulate in the cell and initiate the endoplasmic reticulum (ER) stress response[86]. However, production of variant protein can either exacerbate disease severity through accumulation of toxic proteins in the cell[87], or alleviate it through providing residual function that protects against haploinsufficiency-mediated disease[88–90], meaning the occurrence of NMD can affect phenotypic severity depending on the mechanism of disease. PTVs may also cause LoF through aberrant splicing[73], which is also regulated by NMD[91]. In some cases, the location of NMD boundaries at the 5' and 3' ends of genes containing causal variants can explain phenotypic variation between individuals with different PTVs in the same gene[92,93]. For example, PTVs located outside of the region that triggers NMD in *SOX10* escape NMD and produce proteins that have dominant negative activity, causing the severe complex neurological disorder PCWH, whereas PTVs located within the NMD region produce transcripts that are recognised by NMD and removed, causing the relatively milder WS4 syndrome via haploinsufficiency[94,95]. This variability in penetrance or expressivity could potentially be classed as distinct subtypes of disease, with different variants causing different mechanisms of disease and producing distinct syndromes. Pathogenic variants in *KAT6B* show a similar disease manifestation, with two distinct syndromes depending on whether NMD is triggered or not[96]. Variants in *KAT6A* cause severe intellectual disability (ID) and neurodevelopmental disorders (NDD), with late PTVs more likely to cause a severe phenotype, compared to 60% of early PTVs which conferred a mild phenotype[88], potentially due to whether NMD is activated or not. The position of the PTV within the gene has also been seen to modulate the severity of clinical phenotypes in Marfan Syndrome[97] and Charcot Marie Tooth disease[98]. Disease due to *SFTPB* variants typically presents in neonates as respiratory distress syndrome, resulting in death within the first few months; variants that allow partial production of the SP-B protein confer longer survival times and later

onset of disease, whereas the variants that cause complete deficiency of SP-B due to NMD cause fatal neonatal respiratory distress syndrome[99].

Missense variants can also result in LoF due to substantially reduced protein function or stability[100]. Although many missense variants have little or no effect, they can result in conformational changes, increased protein misfolding, and aberrant protein trafficking, which can lead to intracellular retention or accumulation, increased ER stress, activation of the unfolded protein response, or increased pro-apoptotic signalling and apoptosis[99]. Some missense variants, small insertions/deletions and gene duplications can also result in gain of function (GoF) effects due to increased activity[101], increased protein production[102], or via protein products that gain a new damaging function[103]. Some GoF variants can exhibit a more severe phenotype than LoF variants in the same gene; for example GoF variants in *KCNA2* were associated with more severe epilepsy phenotypes than LoF variants[104]. Where in a gene a variant is located can affect mechanism of disease, as well as penetrance and expressivity through molecular subregional effects[105]; the impact of a variant depends on whether it is located at sites that undergo post-translational modification, within sites that are critical for tertiary and quaternary structure, at protein-protein interaction interfaces or ligand binding sites, or inside or outside of functional domains[106]. For example, missense variants in *GRIN2A* located in transmembrane or linker domains were more frequently associated with severe developmental phenotypes than those located elsewhere, such as within amino terminal or ligand-binding domains[107], with a wide range of phenotypes observed from normal, to mild epilepsy, to severe developmental and epileptic encephalopathy[108]. Similarly GoF variants in highly conserved regions of the potassium channel of *KCNA2* were associated with more severe epileptic encephalopathy phenotypes than variants located elsewhere[109]. An improved understanding of protein structure and the functionality of interacting domains will help elucidate specific variant effects on resulting phenotypic presentation[110].

Finally, there are a small but increasing number of pathogenic non-coding variants that have been identified as causes of monogenic diseases. These variants operate either through LoF or GoF mechanisms by altering gene or

isoform expression[111]. For example, biallelic variants in the *PTF1A* enhancer are a well-established cause of recessive pancreatic agenesis through tissue-specific LoF[112], *de novo* LoF variants in the 5' untranslated region (UTR) of *MEF2C* have been shown to account for around a quarter of developmental disorder diagnoses in this gene[113], and a single GoF variant that creates a novel promoter has been shown to cause α-thalassaemia[114]. However, establishing the pathogenicity of non-coding variants is often much more challenging than coding variants, and thus studies of penetrance and expressivity of these variants are likely to lag behind.

2.22 Size of repeat expansions

Repeat expansion disorders are caused by genomic expansions of short tandem repeat (STR) sequences that either affect gene expression or protein sequence[115], with the penetrance and expressivity affected by the number of repeats (**Table 2**). Anticipation is often observed in families due to molecular instability around the repeats; in each generation the repeat length can increase, resulting in earlier onset of disease and increased severity. For example, Fragile X Syndrome is caused by expansion of over 200 repeats in the CGG motif in the 5'UTR of *FMR1* on the X chromosome, resulting in hypermethylation of the promoter, silencing the gene[116]. Fragile X exhibits incomplete penetrance and reduced expressivity, with 100% of males and 60% of females presenting with ID, and 50-60% of males and 20% of females diagnosed with autism spectrum disorder (ASD)[117]. Wild type (WT) alleles contain <44 CGG repeats while full mutations in affected individuals typically have >200 repeats. Those with premutation alleles of 55-200 repeats have milder phenotypes compared to full mutation carriers, although they have an increased risk of Fragile X associated tremor/ataxia syndrome[118] and primary ovarian insufficiency prior to age 40[119] compared to WT. Monotonic dystrophy shows a similar mechanism, with unaffected individuals having 5-37 CTG repeats in the 3'UTR of *DMPK* and fully affected individuals having >80 repeats (although repeats of >1000 have been seen in congenitally affected children[120]), with the number of repeats correlating with earlier age of onset.

Disease	Gene	STR	Non-penetrant	Intermediate Penetrance	Full Penetrance
Spinocerebellar Ataxia 8	*ATXN8OS* / *ATXN8*[121]	CTG/CAG	<91	92-106	>107
Spinal Muscular Atrophy	*SNM1*[122]	CAG	<34	35-46	>47
Fragile X	*FMR1*[116]	CGG	<44	45-200	>200
Huntington's	*HTT*[123]	CAG	<36	37-39	>40
ALS	*C9orf72*[124]	GGGGCC	<23	24+	>700
Friedrich's Ataxia	*FXN*[125]	GAA	<34	35-99	>100

Table 2.2: Trinucleotide repeat disorders with varying penetrance depending on the number of repeats present.

While the number of repeats accounts for a large proportion of variable expressivity, there are still missing genetic factors accounting for differences in age of onset. For example, in Huntington's disease, a lower number of n-terminal CAG repeats in *HTT* is associated with reduction in penetrance and later onset of clinical symptoms[123] but while the number of repeats is inversely correlated with the age of onset of motor symptoms they only account for 70% of the variability[126]. The remaining unexplained variance displays a high degree of heritability, suggesting further genetic modifiers[127]. Additional genetic variants in the DNA mismatch repair pathway have been linked with anticipation and overall severity of disease, and functional studies showing the knockout of base-excision repair or transcription-coupled repair pathways in animal and cellular models of nucleotide repeat disorders can inhibit the expansion and reduce the phenotypic severity[128,129]. Variants in the DNA repair gene *MSH3* have also been linked with differences in disease severity through somatic instability[130]. As non-penetrant individuals will not necessarily come to clinical attention, and large triplet repeats are hard to genotype accurately using NGS[131], it is suspected that individuals with fewer than 41 CAG repeats in *HTT* may exist at a higher frequency than previously expected in the general asymptomatic population[123].

2.3 Gene expression

2.31 Variation in allelic expression

It has been hypothesized that the differential expression of alternative alleles in the gene containing causal variants could affect the presentation of phenotypic traits in individuals with identical genotypes. This mechanism has been proposed primarily for dominantly inherited conditions where haploinsufficiency is the cause of disease[132,133], including Lynch Syndrome[134] and hypertrophic cardiomyopathy (HCM)[135], where an allelic imbalance could cause either higher expression of the wild-type allele (thus compensating for the haploinsufficiency and resulting in reduced penetrance), or lower expression of the WT allele (thus exacerbating the haploinsufficiency and resulting in higher penetrance). Significant allelic imbalance has been observed in up to 88% of genes in human tissues, potentially caused by genetic modifiers or stochastic factors[136], and has been identified as both tissue-specific and genome-wide in mouse models[137]. Structural variants such as duplications that are in *trans* with a pathogenic LoF variant can alleviate the potential clinical phenotype when disease would be caused by haploinsufficiency, by providing an additional WT copy of a gene, thus resulting in a normal level of gene expression[138], as has been observed in DiGeorge syndrome[139]. Additional variants in the untranslated regions of mRNA can also affect translational efficiency, and gene expression can also vary widely across tissues[140], highlighting the importance of sequencing disease-relevant tissue in interpretation of genetic variation[141]. Compared to synonymous variants, rare missense variants show a significant reduction in allelic expression across many tissues in proportion to their predicted pathogenicity, suggesting deleterious variants are depleted from highly expressed haplotypes[142]. Some highly differentially expressed genes have been shown to contain fewer disease-associated variants[143], which are less likely to accumulate on haplotypes that are highly expressed, or in high-penetrance combinations[142]. For example, genetically heterogenous monogenic eye disorders display both incomplete penetrance and variable expressivity and also display significant variability in gene expression levels throughout the population[144]. Differential expression of alleles has also been shown to play a role in the variable expressivity of the dominant condition Marfan's syndrome[145] and inherited eye disorders[144]. In HCM, the proportion of sarcomeric proteins

produced by variant alleles can vary with allelic expression, and 30-80% of sarcomere structure can be made up of proteins with reduced function[146,147], causing variation in overall phenotypic severity. Differential expression of alleles can also potentially cause recessive conditions to present in a dominant fashion. For example, Zellweger spectrum disorder (ZSD) is an autosomal recessive disorder caused by deleterious variants in any of 13 PEX genes, with the most common cause being variants in *PEX1* or *PEX6*. Affected heterozygous carriers have been identified with ZSD, despite lacking a second pathogenic allele, with all affected heterozygotes presenting with allelic overexpression of the variant allele compared to WT, and a common polymorphism has been linked to this allelic overexpression[148].

Stochastic variation within normal cellular and developmental processes can potentially be amplified by disease-causing variants, and thus play a role in incomplete penetrance and variable expressivity[149]. Random monoallelic expression (RME) is the transcription of only one allele from a homologous pair, and can be constitutive, with all cells expressing the same allele throughout (as seen in imprinted genes), or somatic, with individual cells showing variation in expression levels[150]. Overall levels of RNA in cell populations tends to be stable, but dynamic allelic fluctuation through RME can present variability in gene expression. Genes that show little RME are mostly housekeeping genes which have higher expression levels[150]. Although no variation in disease trait has yet been definitively linked to somatic RME, conceptually it could explain phenotypic variation either through alteration of gene dosage or higher expression of a variant allele. RME during embryonic development has been tentatively linked with variation in developmental disorders such as Holt-Oram syndrome[151]. Model organism research has suggested stochastic variation in gene expression can affect the expressivity of variant genotypes, with 20% of genes causing variation in phenotypes in two different isolates with defined genetic backgrounds in *C. elegans*[152]. Phenotypic variability has also been observed in inbred mice with a defined genetic background[67], as well as in monozygotic twins[153] suggesting the influence of stochastic molecular events in variable expressivity.

2.32 Variation in isoform expression

Production of different transcripts of genes may also lead to differential expression of traits and explain why potentially deleterious variants in haploinsufficient genes are found in population cohorts. Annotations based on transcription levels of different isoforms in haploinsufficient genes identified 23% of LoF variants are in under-expressed exons, and had similar effect sizes to synonymous variants[73]. In monogenic cardiomyopathies caused by LoF variants in the giant muscle protein titin, studies of *TTN* expression levels indicate that LoF variants found in unaffected population cohorts occur predominantly in exons that are absent from the most highly expressed transcripts, and thus do not cause the phenotypic effect associated with deleterious variants[154,155]. Similarly, haploinsufficiency of *TCF4* causes the highly penetrant Pitt-Hopkins syndrome[156,157], and unaffected individuals identified with PTVs in this gene were all found to be located in minimally expressed exons[136], suggesting that functional protein can be made in the presence of these variants. The expression of tissue-specific isoforms can also affect the penetrance of a genotype, potentially resulting in distinct disease subtypes. For example, *CACNA1C* has two clinically important isoforms with mutually exclusive exons that explain two different forms of Timothy Syndrome; pathogenic variants across the widely expressed transcript produce a multi-system disorder (type 1), while pathogenic variants in the alternative exon of a transcript predominantly expressed in the heart are much rarer and result in more severe cardiac-specific defects and fewer syndromic phenotypes (type 2)[158]. Further examples are likely to be uncovered through large-scale analysis of isoform expression in different tissues and at different times.

2.33 Cis and trans acting genetic modifiers

Variants in regulatory regions can affect the phenotypic presentation of disease by altering gene expression, and through modulation of deleterious genetic variants found in associated protein-coding regions[159], potentially affecting the penetrance and expressivity of the monogenic variant. Cis acting elements are DNA sequences located on the same haplotype as the gene they affect, whereas trans-acting factors are proteins or elements that bind to the cis-acting sequences to affect gene expression. Variants in these non-coding regions can have multiple downstream effects, through interactions with other genetic

features, or through effects on monogenic variants[160]. Small changes within transcription factor binding or expression can lead to dysregulation that affects multiple genes within the same regulatory network[160], and therefore could potentially alter the final phenotypic presentation. Cis-regulatory variants have been36oncept36ed that modify the penetrance of coding variants, and therefore contribute to disease risk or presentation. Pathogenic coding variants are depleted from higher-expressed haplotypes with cis-regulatory variants in the general population[142], suggesting that individuals who present with a disease phenotype may have an enrichment of cis regulatory variants that increase the expression of the pathogenic allele, compared to individuals who are asymptomatic who have an enrichment of 'protective' regulatory variants that decrease the expression and therefore penetrance of the pathogenic allele[142].

Upstream open reading frames (uORFs) are tissue-specific *cis*-regulators of protein translation found in the 5'UTR of protein-coding genes, and variants that alter uORFs can affect whether a deleterious protein-coding variant causes a disease phenotype or not, and may alter the phenotypic presentation of the disease[161]. Active translation of a uORF can reduce downstream protein levels by up to 80% via several mechanisms, including production of a peptide that stalls the translating ribosome[162], and termination at a uORF stop codon that can trigger NMD[163]. Variation that either introduces or removes uORF start or stop codons can therefore affect phenotypic presentation, and peptides created by uORF variants may also have a role in disease pathology[164]. Variants in the downstream 3'UTR may also play a role in regulation of gene expression through altering mRNA stability or translational efficiency[140,165,166]. For example, a common SNP downstream of *GATA6* has been shown to reduce *GATA6* expression, potentially resulting in a more severe pancreatic agenesis phenotype when found in trans with a LoF variant in the same gene[167]. Similarly, polymorphisms in the 3'UTR region of *KCNQ1* have been suggested to alter expression of the *cis* allele, either increasing the severity of disease or reducing it through uneven expression of WT or variant alleles[168]. However, an attempt to replicate this in a diverse group of population cohorts found no association between the identified polymorphisms and the severity of disease[169], highlighting the difficulties with trying to identify non-coding modifiers of rare disease, both in clinical cohorts and population studies.

Approximately 400,000 candidate enhancer regions have been identified in the human genome, an average of around 20 enhancers per gene[170,171]. Non-coding variants within enhancer regions can be a cause of phenotypic diversity through alterations in gene expression, therefore affecting overall disease phenotype presentation[172]. Although identifying non-coding variants that affect disease presentation can be very difficult, there are some notable examples. A large study identified a SNP in an intronic enhancer of the *RET* gene that appeared to increase penetrance of Hirschsprung disease in patients with rare *RET* coding variants[173]. Intronic variants have also been suggested to affect the penetrance of coding variants in patients with Stargardt disease, where a deep intronic variant has been shown to be a major cis-acting modifier of the most common pathogenic variant in *ABCA4*[174,175]. A small study also suggested that SNPs in promoter regions affect severity of arrhythmias among individuals with LoF variants in *SCN5A*[176]. Variants that create novel binding sites for transcription factors have been implicated in affecting penetrance through altering gene expression, including a common non-coding polymorphism that alters the hepatic expression of *SORT1*[177], contributing to myocardial infarction. Further WGS research is needed to identify non-coding variants that affect gene expression levels.

Genes are often associated with more than two *cis* regulatory elements through topologically associated domains (TADs)[178]. These domains are thought to affect gene expression and mediate the effects of *cis*- and *trans*- regulatory factors through the 3D conformation of chromatin, and therefore variants in these domains can affect penetrance and expressivity of genotypes[179,180]. Although expression of some genes has been shown to be unaffected by changes in TADs[181], the creation of new TADs has been implicated in the pathogenicity of rare duplications[182]. Alterations to 3D chromatin structure within and between TADs can lead to mis-alignment of genes, enhancers, and silencers, affecting transcriptional control of gene expression[183]. Variants in TAD loops may have no effect on healthy individuals but could affect disease-presentation in those with an underlying monogenic variant[184]. Common genetic variants in *cis* regulatory domains can affect gene expression, and rare variants have been identified that disrupt the structure of the domain[160,185], and both

could contribute to varying phenotypic expressivity of identical protein-coding sequences by causing changes in upstream mechanisms of gene regulation. Structural changes that affect transcription factor binding can lead to functional gene expression changes[180], as seen in the *EPHA4* locus, where deletions or duplications that overlap the TAD boundary can cause severe limb malformations[186], while deletion of the entire locus does not[187], thought to be due to differential gene-enhancer associations.

2.34 Somatic mosaicism

Postzygotic *de novo* mutations that occur during cell division can result in somatic genetic variation that differs between cells, leading to mosaicism[188]. Monogenic disease is usually less severe in mosaic individuals compared with those who have the same variant constitutively and, depending upon which cells or tissues contain the pathogenic variant, mosaicism can result in non-penetrance or reduced expressivity[189]. Somatic mosaicism is suspected to be more widespread than is usually detected, especially when testing only a single tissue sample that may or may not contain the clinically relevant variant(s), although NGS is making it easier to identify lower-level genetic changes[190,191].

Mosaic somatic variants have been suggested to be more representative than germline variants of the true diversity and range of potential variation in human disease, as genotypes that are lethal in constitutive form can be identified when present as mosaic[192,193]. These include variants that cause osteogenesis imperfecta, where a mosaic father presented with mild symptoms, but the constitutive form was incompatible with life[194], Proteus Syndrome[195] and CLOVES Syndrome[196], two overgrowth disorders that are lethal in constitutive form, and various mosaic aneuploidies[197]. Alternatively, mosaic individuals can display different or milder phenotypes compared to those with germline variants in the same gene. For example, mosaic individuals with a variant in *HRAS* present with benign keratinocytic epidermal nevi (“woolly hair”)[198], whereas those with the same constitutive variant have the more severe Costello Syndrome[199]. Other diseases that have been demonstrated to show a milder phenotype when caused by somatic mosaicism include telangiectasis[200] and polycystic kidney disease[201]. Mosaic genotypes can also display varying phenotypes that include segmental forms of the constitutive disease, such as

segmental neurofibromatosis type 1, where clinical manifestations are only seen in certain parts of the body[202]. As well as presenting with variable expressivity, mosaic variants can also be incompletely penetrant. In individuals with primary immunodeficiencies, 80% of mosaic individuals were clinically asymptomatic, with the remaining 20% exhibiting partial clinical symptoms[48,203]. Similarly, mosaic chromosomal aneuploidy has been shown to be incompletely penetrant in population cohorts, with women who had 45X,46XX mosaicism presenting with normal reproductive lifespan and birth-rate, and no cardiovascular complications, compared to those with the non-mosaic genotype[204]. Unaffected parents with mosaic pathogenic variants can pass their genotype onto their offspring as a constitutive germline variant, so an incompletely penetrant or milder disease in one generation can cause a completely penetrant disease in the next[205–209].

Somatic mosaicism can also rescue an individual from disease, through cellular reversion that reduces the expressivity of a phenotype. For example, somatic reversions have been observed in several cell lineages from individuals with immunodeficiency caused by biallelic variants in *DOCK8*, including SNVs that correct or remove germline PTVs, and recombination events that attenuate or remove the deleterious variant from one allele. These somatic reversions improve overall survival time, but they are unable to completely eliminate the disease phenotype[210]. Somatic reversion has been observed in other primary immunodeficiencies[211,212] and may partially explain incomplete penetrance[48]. Reversion of clinical phenotype in individuals with recessive dystrophic epidermolysis[213] and Fanconi anaemia[214,215] has also been identified. Remarkably, long-term remission from WHIM syndrome, caused by GoF variants in *CXCR4*, was seen in an adult who had undergone chromothripsis of chromosome 2 resulting in deletion of the disease allele in a single haematopoietic stem cell, leading to repopulation of the bone marrow with the haploinsufficient CXCR4 cells[216,217].

2.35 Epigenetics

Epigenetic modifications are molecularly heritable changes that alter gene expression without altering the DNA sequence itself, including DNA methylation, histone modifications, and microRNA (miRNA) expression[218]. Differential

epigenetic modifications between individuals carrying the same pathogenic genotype can potentially account for incomplete penetrance and variable expressivity of phenotype. DNA methylation is important in the control of alternative splicing, prevention of cryptic initiation of transcription from alternative promoters, and X chromosome inactivation, all of which have been shown to affect progression of disease[219]. Studies of monozygotic (MZ) twins that are discordant for disease phenotypes have highlighted how epigenetic mechanisms could affect the penetrance or expressivity of disease[220]. For example, MZ twins with neurofibromatosis, caused by variants in *NF1*, showed significant discordance in the presence of tumours, and severity of scoliosis, suggesting that additional non-hereditary factors were modifying their phenotypes[221]. Similarly, one MZ twin with a pathogenic homozygous variant in *GBA* was diagnosed with Gaucher disease, while the other was clinically asymptomatic[222,223], and differences in their epigenome were posited as a mechanism to explain this discordance. However, epigenetic studies are generally more challenging than genetic studies, as variation may be both tissue and time-specific, making it harder to elucidate how epigenetic mechanisms affect the penetrance of such genotypes. One suggested mechanism is that epigenetics may compensate for the presence of a deleterious variant, and this may segregate through several generations without any ill effects until the epigenetic modifications are no longer functional[224]. This has been seen in Xq24 microdeletions that are inherited from mothers with extremely skewed X-chromosome inactivation, which modifies the penetrance[224]. Skewed X inactivation is also suggested to be a cause behind the clinical heterogeneity in Klinefelter Syndrome[225]. Epigenetic mechanisms have also been suggested to partially compensate for deletions in heathy carriers of *IMMP2L* deletions, which cause ID and NDD, as reduced DNA methylation levels were seen in healthy carriers but not in affected offspring[226].

Another mechanism by which epigenetic changes may affect penetrance of monogenic diseases is via miRNAs, small non-coding RNAs that regulate gene expression[227]. One miRNA can influence multiple genes, and a gene can be affected by several miRNAs, potentially highlighting how variants in one may lead to multiple downstream phenotypic effects[228]. Differential miRNA expression can be caused by genetic variation, and variants within miRNA

could thus affect allelic expression and modify the penetrance or expressivity of monogenic diseases[229]. Expression of numerous miRNAs may affect the penetrance and expressivity in hereditary breast and ovarian cancer (HBOC)[230]; incomplete and age-dependent penetrance is common in carriers of pathogenic variants in *BRCA1* and *BRCA2,* and variation in several miRNAs that bind the 3'UTRs and downregulate expression of both genes have been linked with an increased risk of earlier onset cancer[230–234].

2.4 Global modifiers

2.41 Threshold model of disease

There may be a threshold that has to be met for manifestation of a clinical disease phenotype, and genetic and other factors may vary in their relative contribution to meeting this threshold for different diseases and in different individuals (**Figure 2.4**)[235]. Some highly penetrant monogenic disease variants may always be sufficient to push the genetic burden above the threshold of disease, although secondary variants may still contribute to severity[236]. For example, Dravet Syndrome (DS) is a highly penetrant and devastating form of childhood epilepsy caused by *de novo* loss-of-function variants in *SCN1A*[237]. Although DS displays considerable clinical heterogeneity within families, and severity may relate to background genetic variation[238], there are no known modifiers that protect against the effects of the primary causal variant; the LoF variant alone is sufficient to push the individual above the threshold for disease, and other variants can only change the severity of the phenotype above this point. Individuals with monogenic variants that are causative of disease alone, and thus are already above the threshold for disease, can be further modulated by secondary monogenic variants in related genes that also cause the same phenotype and the accumulation of these PTVs is associated with a more severe phenotype, as the burden is pushed way beyond the threshold[239]. For example, in monogenic polycystic kidney disease, individuals with a PTVs in each of the causative genes, *PKD1* and *PKD2*, present with a much more severe disease than those with just one PTV[240]. Many monogenic disease-causing variants have been found to have secondary genes or loci that affect the severity of their related clinical phenotype[236,241] (**Table 2.3**).

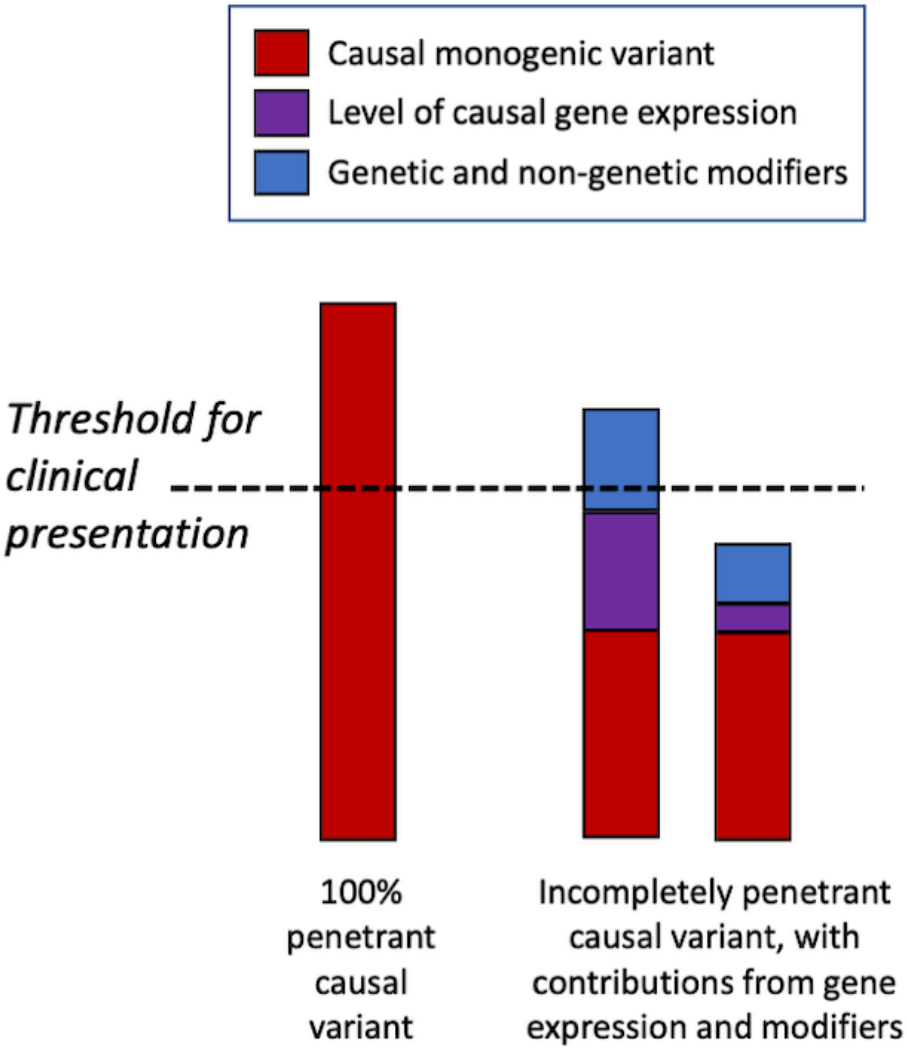


Figure 2.4. Threshold model of disease. Some deleterious monogenic variants are sufficient to cause disease alone, and do not need any genetic modifiers to cause the disease phenotype. Other monogenic variants may be incompletely penetrant, and only display a disease phenotype when accompanied by other genetic or non-genetic factors that raise them above the clinical threshold for disease presentation. In the latter scenario, individuals may have the same underlying causal variant, but have very different phenotypic presentations depending upon their modifying factors.

Disease	Causal gene	Modifier gene/loci	Phenotypic effect
Cystic Fibrosis	*CFTR*	*TGFB1*[242] *IFRD1*[243] *DCTN4*[244,245]	Increased severity of lung disease Earlier age of onset of chronic infection
Sickle Cell disease	*HBB*	*BCL11A*[246] *HBS1L-MYB*[246–249] *CLCN6*[250] *OGHDL*[250]	Prolonged production of fetal haemoglobin, reduced disease severity Decrease in disease severity
Long QT syndrome	*KCNQ1* *KCHN2* *SCN5A*	*NOS1AP*[251]	Modulate risk of arrythmias
X-linked retinitis pigmentosa	*RPGR*	*IQCB1*[252] *RPGRIP1L*[253] *CEP290*[65]	Increase in disease severity
Bardet-Biedl syndrome	*BBS10*	*MGC1203*[254]	Increase in disease severity
Joubert syndrome	*NPHP11*	*CEP290* *AHI1*[255]	Increase in disease severity (also been linked to monogenic disease alone, with conflicting results[256–258])
Spinal muscular atrophy	*SMN1*	*PLS3*[259] *SNM2*[260]	Reduction in disease severity
Fragile X syndrome	*FMR1*	*COMT*[10]	Reduction in disease severity
Spinocerebellar Ataxia 17	*TBP*	*STUB1*[261]	Changes from non-penetrant to penetrant
Phenylketonuria	*PKU*	*SHANK gene family*[262]	Protective effect on cognitive development in untreated patients

Table 2.3: Examples of monogenic conditions affected by a putative second genetic locus that modifies phenotypic expression.

In contrast, some monogenic disease-causing variants may be partially tolerated and transmitted through unaffected generations unnoticed, until they surpass the threshold for causing disease in the presence of other contributory factors. For example, large copy number variants (CNVs) are well known causes of NDDs, but some – such as recurrent 16p12.1 deletions[263] – have been widely observed to be inherited from unaffected parents. In this case, penetrance of a phenotype that is severe enough to present clinically requires an additional variant that modulates the primary genetic variant[138] supporting a "two-hit" model of NDDs[264]. Similarly, deleterious variants in *CNTNAP2* and *LRRC4C* are insufficient to cause disease alone, but together may impair development and function of synapses[265,266], suggesting a possible digenic mechanism for modulation of phenotypes[267]. In many cases, however, there are likely to be numerous factors that affect whether an individual lies above or below the disease threshold, including the overall deleteriousness of the primary causal variant(s), the level of expression of the causal gene or isoform, and other genetic and non-genetic modifiers (**Figure 2.4**). Global modifiers that might affect penetrance and expressivity include polygenic risk, genetic compensation, variation in NMD efficiency, family history, age, sex, and environmental factors.

2.42 Polygenic risk

The penetrance and expressivity of genotypes can be altered through the accumulated impact of many common genetic variants throughout the genome. The "omnigenic" model proposes that, due to their interconnected nature, variants in gene-regulatory networks that are expressed in disease-relevant cells or tissues may affect the functioning of "core" disease-related genes, due to effects on genes outside of the core pathways[268], suggesting that many unrelated variants contribute to the presentation of a phenotype. While proposed as a factor in inheritance of complex traits, this polygenic architecture could potentially also affect the presentation of monogenic conditions in a similar way, through non-coding variation that affects overall gene regulation, and many loci have been shown to additively affect expressivity and penetrance of monogenic variants in model organisms[269].

Genome-wide association studies (GWAS) have uncovered thousands of susceptibility loci for hundreds of diseases[270], suggesting that polygenic background can either predispose[80] or protect individuals from disease[271]. Polygenic background can be quantified into a polygenic risk score (PRS)[272,273] and potentially used as a tool for the prediction of overall disease risk in both monogenic and polygenic disorders[274]. PRS associations highlight the additional risk of polygenic components in affecting severity of monogenic disease, with polygenic risk being shared across monogenic variant carriers and the general population[275]. The effect of PRS has been widely explored to improve clinical interpretation of the penetrance of pathogenic variants across a range of monogenic conditions, including numerous familial cancer syndromes[276]. The penetrance estimates for individuals with a pathogenic *BRCA1* or *BRCA2* variant range from 45-85% for breast cancer, and 10-65% for ovarian cancer[277,278], some of which can be explained by polygenic background[275,279,280]. Using a PRS generated from breast cancer GWAS, it has been shown that individual carriers of monogenic variants have risk differences of over 10% between the top and bottom PRS deciles[275]. Interestingly, the majority of the SNPs identified as polygenic risk variants are common non-coding variants within regulatory regions, the target genes of which overlap with other known somatic cancer driver genes[281]. Polygenic risk can also have a large effect on phenotypic diversity, even within individuals who have a known monogenic variant, illustrating that the genetic architecture for many diseases can be viewed as a spectrum rather than a binary classification of clinically symptomatic vs asymptomatic[235]. While overall polygenic contribution to disease phenotype can be weaker in individuals with a monogenic variant[282], it can be useful in predicting overall penetrance and risk stratification.

2.43 Genetic compensation

The phenomenon of genetic compensation (or genetic buffering), where another gene or genes in a network can functionally compensate for LoF variants, has been shown in model organisms[283] and hypothesised to play a role in incomplete penetrance in humans[284]. The upregulation of related genes or pathways or the differential expression of compensating alleles can help suppress a disease phenotype[285], either through a small number of compensatory mechanisms or via a global shift in gene expression. The

functional redundancy of genes and rewiring of affected genetic networks may affect the penetrance and expressivity of corresponding phenotypes, and the consequence of a pathogenic variant may be influenced by variation across the genome[286] and explain why certain LoF variants are tolerated by some individuals but not others[287,288]. Haploinsufficiency caused by genetic variation can influence the expression of other genes in the same network, for the purposes of maintaining homeostasis or suppression of disease phenotypes[289]. The functional loss of one gene can be compensated for through functional redundancy[290]. Genes that contain high numbers of PTVs in general population cohorts and thus are less likely to cause adverse phenotypes were found to belong to larger gene families than genes that contain known pathogenic PTVs[19], suggesting functional redundancy as a mechanism affecting penetrance[291]. Further research is needed to find robust evidence of this mechanism in humans.

2.44 Nonsense mediated decay efficiency

The efficiency of NMD varies between individuals[292], which could act as a potential modifier of penetrance and expressivity of PTVs targeted by NMD irrespective of the specific causal variant(s)[293]. The variation in NMD efficiency across codons, genes, cells, and tissues can affect disease pathology[94,294,295]; in studies of model organisms, the variant alleles that caused milder phenotypes were those that exhibited more NMD, with reduction in NMD being correlated with a more severe phenotype[289]. In this case, NMD could either help trigger a compensatory response, or haploinsufficiency could produce a milder phenotype than accumulation of truncated proteins. Variants in genes that encode the NMD machinery, or that either downregulate or remove NMD activity, have been linked to several NDD and ID syndromes, including variants in *UPF2*[296], *UPF3A*[297], *EIF4A3*[298], *SMG8*[299], and *RNPS1*[300], highlighting its importance in development and phenotypic expression. Common polymorphisms within the NMD pathway have been suggested to cause differences in NMD efficiency[301,302], which could help explain differences in expressivity of diseases caused by haploinsufficiency, with severity linked to whether they trigger NMD or not. Interindividual variability in NMD efficiency has the ability to alter the expressivity of genetic variants, through converting the cause of the disease phenotype from dominant negative to haploinsufficiency,

or vice versa[303]. For example, two patients with the same PTV in the *DMD* gene displayed different clinical phenotypes, with one diagnosed with Duchenne muscular dystrophy, and the other with the milder Becker muscular dystrophy; here, the difference in phenotype was suspected to be caused by weaker NMD efficiency in the less severely affected patient, which resulted in production of the damaged but still partially functional DMD protein[304,305].

2.45 Family history

Family history can be seen as a crude but effective proxy for the combined effect of many shared genetic and environmental modifiers of disease phenotypes. In many cases, the pathogenicity and penetrance of variants in monogenic diseases has only been determined through studies of large families with multiple affected individuals, which can make it difficult to disentangle the relative contribution of different modifiers. Family history is a well-known major risk factor for hereditary cancer syndromes and the number of affected relatives increases the risk of a pathogenic variant carrier developing cancer[306]. Although the evidence base for estimating penetrance in individuals without a family history is currently very limited[307], individuals identified with a pathogenic variant for a heritable monogenic disease but without a family history of that disease may have a lower penetrance than those with a family history[17,308].

Evaluating genetic differences between affected and unaffected carriers in the same family – such as *de novo* variants or unique combinations of modifiers – can be informative for understanding penetrance. It has been shown that children with monogenic NDDs have an excess of other damaging genetic variants compared to their either mildly clinically affected or asymptomatic carrier parents, with the extra genetic burden being enriched in genes that are highly expressed within the brain and in neurodevelopmental pathways[236]. Similarly, children with 22q11.2 deletion syndrome display a wide variability in IQ scores that is highly correlated with the scores of their immediate relatives[309]. The IQ of individuals affected by 22q11.2 deletion syndrome follows a normal distribution curve, similar to that of the general population, only 30 points lower[310]. The significant association seen between parental and proband IQ[311,312] suggests inherited genetic variants associated with intelligence may alleviate some of the deleterious impact of the 22q11.2 deletion on phenotypic

presentation. The heritability of intelligence may be driven either by the cumulative effect of many common small-effect variants, similar to the heritability within population cohorts[313], or by a small number of rare high-effect variants. Similarly, individuals carrying 16p11.2 deletions present with variable phenotypic diversity[308,314], and are frequently present in 'healthy' general population cohorts[315], albeit with a range of cognitive and neuropsychiatric difficulties, despite none of them reaching traditional clinical diagnosis threshold levels[316]. Within these carrier individuals, the best overall predictor of phenotype was that of the average of their parental phenotype for the traits of interest, with individuals displaying deleterious effects relative to their phenotypic family background[317,318].

2.46 Age

It can be argued that penetrance is an almost meaningless concept without specifying an age threshold, as many diseases do not present until later in life. As we age, gene expression and chromatin structure across the genome change, which can increase the penetrance or expressivity of disease[319,320].Expression of certain genes can cause change in a predictable way throughout life, with some only being expressed in the feotus or during early childhood, and others only after this developmental period. For example, the relative proportion of two protein subunits in the NMDA receptor alters with age due to the varying expression levels of the two genes, *GRIN2A* and *GRIN2B*, which can alter phenotypic expression of deleterious variants in these genes; prenatally expressed *GRIN2B* is linked with severe cognitive defects from birth, while postnatally expressed *GRIN2A* is linked with epilepsies in childhood and schizophrenia in adults[108]. Studies of individuals who are below the age-penetrant threshold for known age-dependent diseases could explain why some pathogenic variants are found in apparently asymptomatic population cohorts. Classical examples of conditions where penetrance increases with age include cancer predisposition syndromes such as Li-Fraumeni[321], Lynch syndrome[322], and Hereditary Breast and Ovarian Cancer (HBOC)[234], where penetrance is affected by the accumulation of DNA damage over time[323]. Meta-analysis studies have shown that the cumulative breast cancer risk for *BRCA1* and *BRCA2* pathogenic variant carriers by age 70 is 57-65% and 45-49% respectively[234,324], highlighting the difficulties with predicting the course of

disease even in known pathogenic variant carriers, and the importance of considering other genetic and environmental factors[279]. Age-dependent penetrance is also seen in diseases caused by the slow accumulation of aberrant proteins, where variation can affect the rate at which the protein accumulates[325]. For example, retinitis pigmentosa, has been suggested to be caused by retention of misfolded proteins, which leads to up-regulation of genes that encode for proapoptotic machinery, and leads to apoptosis of photoreceptor cells, accumulating damage over time and eventually reaching disease threshold and causing penetrant disease[326]. Age-dependent penetrance may also be caused by gradual loss of neurons, causing the associated disease phenotype when the number of surviving cells drops below a certain threshold or overcomes brain plasticity[327]. For example, progressive and late occurring neurological manifestations in patients with *DNMT1* variants may originate from the gradual loss of DNA methylation over time, affecting adult neurogenesis[219].

The penetrance of age-dependent variants presents a diagnostic and prognostic challenge for individuals with such genotypes[328]. Previously, testing for many conditions early in life was not possible, and so little is known about long term effects of mildly deleterious variants. Variants in *HFE* cause hereditary hemochromatosis, which can lead to iron overload in adulthood, and was previously thought to be an adult-onset condition. However, healthy cohort studies of children have shown that the effects of homozygous variants in *HFE* can be seen in childhood, and that the cumulative effect of excess iron over a lifetime may affect the penetrance of numerous iron-related diseases[329]. Recent population studies of adults have also shown substantially higher morbidity in homozygous *HFE* variant carriers with increasing age[330]. The early identification of individuals at risk can help with monitoring disease progression and introducing timely interventions (such as blood donation).

2.5 Sex

Sex can affect the penetrance and expressivity of some genetic disorders, most obviously when deleterious genetic variants occur on the X chromosome, with hemizygous males more phenotypically affected than heterozygous females. Although differences in the penetrance of inherited variants based on sex have

been reported in a variety of disorders[25], mechanisms behind sex-dependent penetrance outside those that occur on the X chromosome are mostly unknown. However, there are widespread sex-biased differences in gene expression[331], so differences in penetrance of phenotypes is also likely to be common. Females are less likely to be diagnosed with neurodevelopmental disorders than men, with a fourfold increase in the number of males diagnosed with ASD compared to females[332,333], suggesting that there may be a "female protective model" which affects the penetrance of such conditions[334]. Girls diagnosed with ASD have an increased number of CNVs compared to boys with the same diagnosis, and asymptomatic mothers with children diagnosed with NDDs or ASD had a higher genetic burden of deleterious variants than fathers[317], suggesting there may be some other cause for the incomplete penetrance and variable expressivity in females compared to males. However, females are ascertained at a closer frequency to males when they are more severely affected, suggesting some bias in clinical ascertainment due to differing phenotypic presentations between the sexes[335], supported by the fact that males were more likely to be referred for genetic testing than females carrying the same autosomal variant[336].

2.6 Environment

The environment can affect disease penetrance or expressivity in both a negative and positive manner and includes diet, drugs, alcohol intake, physical activity, ultraviolet light, *in utero* exposures, education, and socio-economic status, among many others. Epigenetic factors can provide a mechanistic link between the environment and gene expression[337–339] and studies of the human microbiome can also explain some extreme variability in genotype-phenotype presentation[340]. However, although gene-environment interactions are likely to be widespread, they are often extremely hard to prove as the complete and systematic collection of an individual's environment is almost impossible, and detailed relevant exposure data are rarely available alongside genetic data.

Inborn errors of metabolism perhaps provide the simplest examples of monogenic diseases where both a pathogenic genotype and an environmental exposure are required to cause disease[341]. A clear example of dietary impact on phenotypic variation is phenylketonuria, a rare autosomal recessive disease

that is usually detected through newborn screening, whereby individuals who have damaging variants in *PAH* can be put on a low phenylalanine diet to avoid serious disease progression[342,343]. Later onset monogenic disease penetrance can also be affected by the environment, as seen in several cancer syndromes, including colorectal cancer, where inherited genetic variants interact with dietary variables and BMI to confer overall risk[344]. Cancer susceptibility can also be altered through gene-environment interactions such as smoking or sunburn, that can accelerate the accumulation of somatic variants that contribute towards tumorigenesis[345,346]. Similarly, environmental exposure to cigarette smoke, air pollution, and other airborne toxins can cause accumulation of unfolded or misfolded proteins and therefore affect the penetrance or expressivity of chronic lung disease[347]. Individuals who carry a damaging monogenic variant may also be more susceptible to some environmental exposures, which can affect phenotypic severity[348]. For example, cystic fibrosis is characterized by progressive damage to the lungs, and non-genetic factors may account for up to 50% of the clinical variation seen[349]. Environmental factors such as smoking, air pollutants, temperature and high fat diets have all been shown to affect severity and progression of disease[348–351], and the specific *CFTR* variant can also modulate how much environmental impact has on disease severity[352]. Environmental factors can also affect presentation of disease in primary atopic disorders, commonly seen as monogenic allergic disorders, where diet, microbiome at the epithelial-environment interface, presence/extent of infection, and psychological stress can all affect the penetrance or expressivity of the related phenotype[353].

2.7 Challenges within determining penetrance and expressivity

2.71 Incomplete penetrance challenges definitions of pathogenicity

Determining the penetrance and expressivity of a variant can be difficult because it is sensitive to ascertainment context, and many studies are designed to enable the discovery of causative pathogenic variants in clinically affected individuals rather than to analyse effect sizes in populations[354]. This has been demonstrated in recent studies that stress the importance of cohort background for the determination of penetrance[16,355]. Investigating clinically-classified

pathogenic variants in large population cohorts can provide additional information about penetrance and expressivity[356], or determine whether variants or genes have been misclassified[17]. However, finding low penetrance pathogenic variants in large numbers of asymptomatic individuals challenges the concept of pathogenicity, particularly in the absence of known modifiers. What does it mean to describe a genotype as pathogenic if it is frequently found in individuals without disease and no explanation as to why? Reclassification of previously reported pathogenic variants occurs frequently, with variants first classified prior to the release of large population datasets showing a higher rate of reclassification[357]. A study reappraising pathogenic variants in Brugada syndrome showed that only one gene (*SCN5A*) out of 21 could be definitively identified as causal[358], and another study has raised doubt over the involvement of 11/58 genes thought to cause inherited monogenic retinal disease[359]. Variants that show low penetrance or a wide range of expressivity can also be potentially classified as risk alleles rather than causative variants. Some *CFTR* variants have been classified this way, with variations in cystic fibrosis phenotypes from very mild to very severe, and over 1900 different genotypes reported[352,360,361]. Many genotype-phenotype associations are only reported once, or they are reported several times but with inconsistent results, due to differences in data collection, differences in methods, or differences in cohort ascertainment. Associations can also differ due to poor annotation of coding genes, lack of relevant functional information for non-coding regions, sequencing, and annotation errors, as well as varying penetrance and expressivity, making a simple binary classification of many genetic variants very difficult.

2.72 Monogenic versus polygenic disease

An overlapping genetic basis between complex traits and monogenic conditions is becoming increasingly apparent across the genome. Deleterious variants in genes causative of monogenic disease can be further dysregulated by non-coding variants that are associated with common traits, and monogenic forms of numerous common complex diseases have been identified[271,362,363]. While this can help identify and prioritize genes for further future disease analysis, it can cause considerable complexity when it comes to determining genotype-phenotype relationships[364]. The prevalence of incomplete penetrance and

variable expressivity raises questions as to what constitutes a disease state as opposed to extremes of normal phenotypic variation, especially within conditions that show significant clinical heterogeneity[308], with many traits that constitute a clinical phenotype being the extreme end of either side of the bell curve of continuous distribution in the general population. Therefore, defining the penetrance of a genotype can be difficult, especially when there is ambiguity as to what defines the “disease state”, particularly for disorders where clinical features are only identified when they reach above a certain threshold[365].

2.73 Genetic modifiers are hard to identify

Relatively few studies have investigated low penetrant rare variants in detail or identified why such variants cause disease in one individual and not another. Despite increasing numbers of sequenced individuals, identification of genetic modifiers for monogenic conditions remains challenging. By definition, carriers of rare variants that cause monogenic conditions will be rare, with even fewer individuals having identical genetic modifiers that explain incomplete penetrance or variable expression. NGS approaches involving bioinformatic algorithms, including pathogenicity score-based prioritisations, can produce conflicting results, and often need manual curation to identify candidate variants. A computational approach that could comprehensively analyse and prioritize candidate variants and potential modifiers would be a great advantage. Even in large population cohorts genome-wide analysis of genetic interactions lacks statistical power, and can be easily affected by confounders[366]. Many genetic modifiers are likely to be located in non-coding regions, making it challenging to determine their direct functional effect on gene expression, especially as much of the genome is found to be bound by at least one transcription factor, many of which have no known function yet[171]. Improved computational approaches to identify candidate modifier gene interactions across the genome are needed[367], as well as identification of functional non-coding regions and the genes that they affect[368], and machine learning approaches such as DeepSEA and Enformer[369] could improve annotation of these regions[370].

3. Chapter three: Rare genetic variants in dominant developmental disorder loci in UKB

3.1 Introduction

Many rare diseases are caused by deleterious variants in thousands of monogenic disease genes[9]. However, not all individuals with these variants share the same clinical phenotypes; some don't appear to be affected at all, whereas others are very severely affected[48]. Monogenic variants can cause different effects in different individuals[272]. The range of phenotypes caused by deleterious variants in the same gene can be explained by pleiotropy, incomplete penetrance, and variable expressivity[17]. Penetrance (i.e. whether an individual with a disease-causing genotype displays the corresponding clinical phenotype) is generally binary; either a variant is penetrant and causes the clinical phenotype associated with that genotype, or it is not[45,48]. In contrast, variable expressivity (i.e. the range of phenotypes that can be observed in affected individuals) is generally continuous, e.g. from mild to severe[46]. Although incomplete penetrance and variable expressivity are distinct concepts, in practice they can be hard to separate, especially when considering the continuous spectrum of phenotypes in populations.

As most disease-causing monogenic variants have been identified through small clinical cohorts, including families with multiple affected individuals, penetrance of these variants is often over-estimated. Investigating the effect of these variants in the general population is therefore important to give a more accurate view of the penetrance in clinically unselected individuals and families. It has been suggested that many of the primary symptoms of rare disease are actually extremes of normally distributed phenotypes in the general population[9,371]. Large, well genotyped population cohorts with deep phenotypic data gives us the ability to investigate the spectrum of phenotypes of individuals with variants in known monogenic disease-causing genes. Phenotypic heterogeneity and variability are a major concern for rare Mendelian disorders, where they can lead to incorrect or delayed diagnoses[3,372].

Many severe developmental disorders (DD) manifest from birth or early childhood and are caused by rare damaging variants in around 2,000 genes and loci[36]. Pathogenic variants in these genes have been identified primarily through phenotype-led clinical studies of affected individuals and families[17]. Due to extensive genetic and phenotypic heterogeneity, large multigene panels are increasingly being used for diagnostic testing, often through panel-based virtual analysis of whole exome or genome sequence data. However, little is known about what effect, if any, deleterious variants in these genes have on adults in the general population or their life-long implications. In this study, using genetic and phenotypic data from UK Biobank (UKB)[28], we investigated whether adults with rare deleterious variants in genes and loci known to cause autosomal dominant forms of DD have any developmentally-relevant phenotypes.

3.2 Materials and Methods

3.21 UK Biobank cohort

UKB is a population-based cohort from the UK with deep phenotyping data and genetic data for around 500,000 individuals aged 40-70 years at recruitment. Individuals provided a variety of information via self-report questionnaires, cognitive and anthropometric measurements, and Hospital Episode Statistics (HES) including ICD9 and ICD10 codes. Genotypes for single nucleotide polymorphisms (SNPs) were generated using the Affymetrix Axiom UK Biobank array (~450,000 individuals) and the UK BiLEVE array (~50,000 individuals). This dataset underwent extensive central quality control (http://biobank.ctsu.ox.ac.uk). A subset of ~200,000 individuals also underwent whole exome sequencing (WES) using the IDT xGen Exome Research Panel v1.0; this dataset was made available for research in October 2020. Detailed sequencing and variant detection methodology for UKB is available at https://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170. The UKB resource was approved by the UK Biobank Research Ethics Committee and all participants provided written informed consent to participate.

3.22 Gene selection

We used the clinically curated Developmental Disorders Gene-to-Phenotype Database (DDG2P) to select genes known to cause monogenic DD. The

database (https://www.deciphergenomics.org/ddd/ddgenes - accessed on 27 November 2020) was constructed from published literature and provides information relating to genes, variants and phenotypes associated with DDs, including mode of inheritance and mechanism of pathogenicity[36]. We initially included all genes that had been annotated as a “confirmed” or “probable” causes of autosomal dominant DD (n=599). Further subsets of these genes were selected for sensitivity analyses, including: a panel of 325 genes that are known to cause DD through a loss-of-function (LoF) mechanism; a more stringent panel of 125 of these haploinsufficiency genes that were significantly enriched for damaging *de novo* LoF mutations in a recent analysis of 31,058 DD probands[373]; and a small panel of 25 clinically well-established genes with >30 likely pathogenic *de novo* LoF mutations in the same study[373] (see **Figure 3.1** and **Appendix Table 7.3.1**).

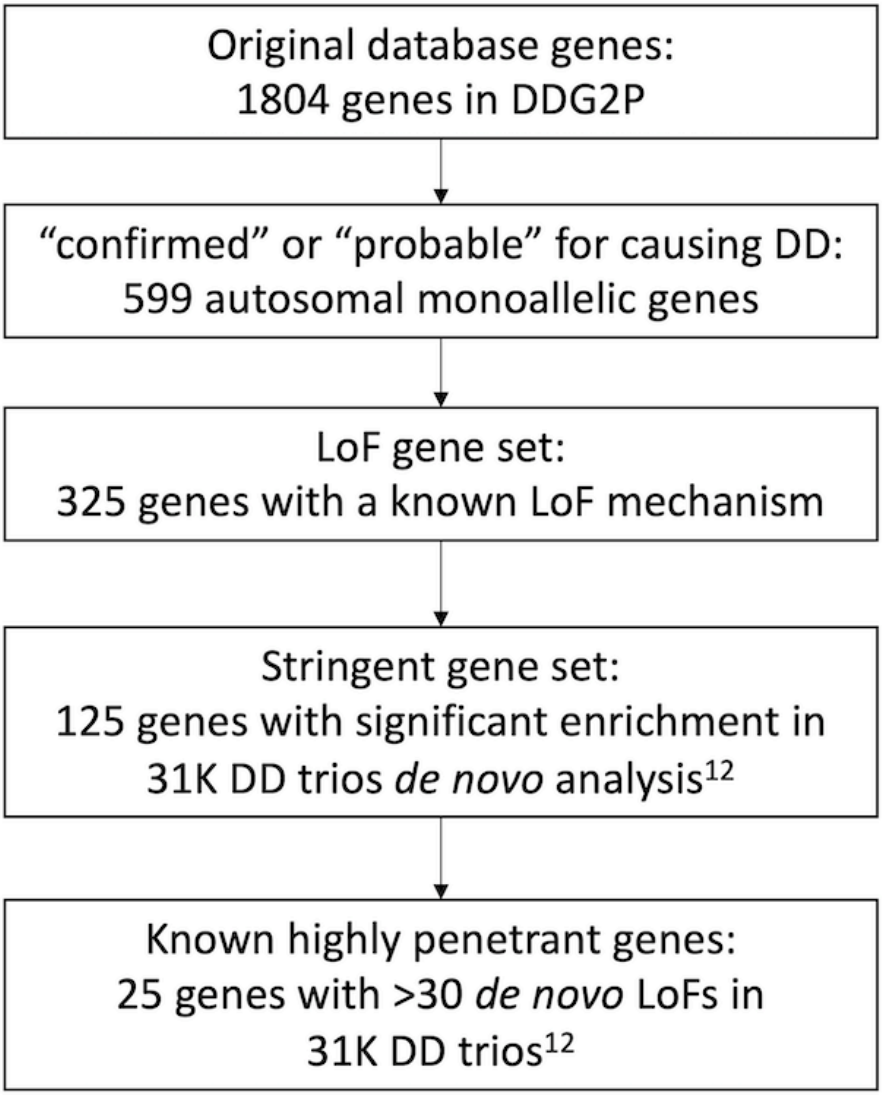


Figure 3.1. Flow diagram outlining selection process for DD genes in each subset that were used for analysis. DDG2P = Developmental Disorders Genotype-to-Phenotype database; DD = Developmental Disorder; LoF = Loss of Function variants; 31K DD trios = 31,058 parent-offspring families with developmental disorders (12: (Kaplanis et al. 2020)).

3.23 Variant selection

To investigate the penetrance of likely deleterious single nucleotide variants (SNVs) and insertions/deletions (indels) in known autosomal dominant DD genes, we used WES data from 200,632 individuals in UKB to identify individuals with a rare SNVs and/or indels in any of these genes. For most of our analyses, rare was defined as any variant that occurred in 5 or fewer

individuals in the UKB cohort; we also investigated the effect of changing this threshold to n=1, n=10, n=50 and n=100 individuals. We included variants that had individual and variant missingness <10%, minimum read depth of 7 for SNVs and 10 for indels, and at least one sample per site passed the allele balance threshold > 15% for SNVs and 20% for indels. We selected three functional classes of variant in canonical transcripts based on annotation by the Ensembl Variant Effect Predictor[37]:

(1) *Likely deleterious LoF variants:* we defined a LoF variant as one that is predicted to cause a premature stop, a frameshift, or abolish a canonical splice site; only those variants deemed to be high confidence by the Loss-Of-Function Transcript Effect Estimator (LOFTEE) were retained (https://github.com/konradjk/loftee).

(2) *Likely deleterious missense variants:* missense variants with a REVEL score > 0.7[374]. A further set of likely deleterious missense variants were identified using CADD[375], with cut offs of 20, 25, and 30**.**

(3) *Likely benign synonymous variants*.

Individuals with variants in group (1) were excluded from groups (2) and (3); individuals with variants in group (2) were excluded from group (3). Following variant selection, one gene (*DNMT3A*) was removed from further analysis as the variants in this gene – which is known to be strongly linked with blood cancer[376] – had a significantly lower allele balance, suggesting substantial somatic mosaicism (**Figure 3.2**). Other genes linked to blood cancer such as *ASXL1* and *TET3* were examined, but showed no difference in allele balance compared to the remainder of the LoF variants identified. LoF variants in the most stringent 25 gene subset were visually confirmed using the Integrative Genomics Viewer (IGV).

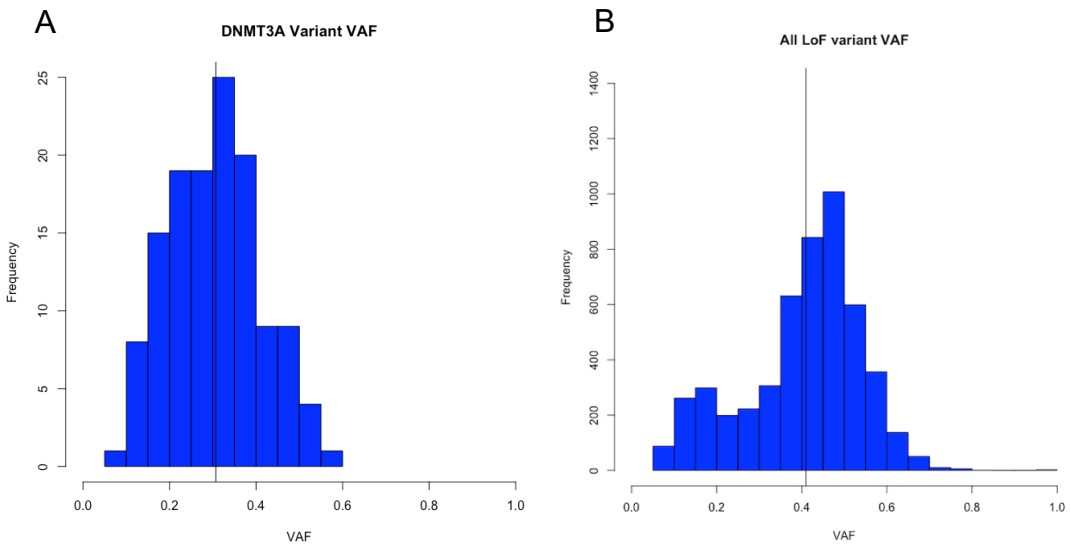


Figure 3.2: Histogram of variant allele balance, highlighting variants in *DNMT3A*. The average VAF of *DNMT3A* variants (A) is significantly below that of the average of the remaining LoF variants (B).

To investigate the penetrance of multigenic copy number variants (CNVs) overlapping known DD loci, we used SNP-array data from 488,377 genotyped individuals in UKB and PennCNV[377] (version 1.0.4) to detect multigenic CNVs overlapping 69 published CNVs strongly associated with developmental delay[378,379]. Log R ratio (LRR) and B-allele frequency (BAF) values for 805,426 genome-wide SNP probe sets were provided by UKB, and an in-house script was used to convert these data to PennCNV input signal files. The PennCNV Hidden Markov Model (HMM) transition matrix was trained using 250 random UK Biobank samples using PennCNV-train. Population Frequency B Allele reference data (PFB) were generated using 1,000 random UK Biobank samples. PennCNV-test was then used to detect regions in a duplication or deletion state in LRR/BAF Hidden Markov Model (HMM) with the generated PFB and transition matrix. An individual was classified as having a multigenic DD deletion or duplication if the region detected using PennCNV reciprocally intersected the published region by at least 50%. We plotted LRR/BAF data for each call in each of these regions, and carried out visual inspection of each event, and false positives and single gene CNVs were excluded. A list of included CNVs is provided in **Appendix 7.3.2**.

3.24 Statistical analysis

We performed both individual gene and gene panel burden tests across our different gene subsets. We grouped individuals into one of five groups depending upon the type of variant they carried (LoF, missense or synonymous variants in one or more autosomal dominant DD genes; or deletions or duplications overlapping published DD multigenic CNVs). Association tests were limited to individuals in UKB with genetically defined European ancestry that were unrelated up to third-degree relationship (184,142 with WES data; 380,029 with SNP-array data) and were controlled for age, sex, recruitment centre and 40 principal components. Variant burden association tests in gene panels and multigenic CNVs were performed using STATA (version 16.0), using linear regression for continuous phenotypes and logistic regression for binary phenotypes. Associations were tested between each group of individuals and other individuals in the UKB cohort without any of the classes of rare variation defined above. Information from HES codes, self-report questionnaires and cognitive tests taken at recruitment was used for the phenotypic information. Associations were tested for 22 UKB phenotypes selected based on their likely relevance to developmental disorders, including:

- *Medical*: epilepsy (self-reported or ICD10 codes G40); ever reported a mental health issue (self-reported through questionnaire); diagnosed with “Child DD” (including intellectual disability (ICD10 codes F70-73), epilepsy (G40), developmental disorders (F80-84) and congenital malformations (Q0-99)); or diagnosed “Adult DD” (including schizophrenia, (self-reported or ICD10 codes F20-29) and bipolar disorder (self-reported or ICD10 codes F30-F39)).
- *Reproductive*: infertility, number of pregnancies, number of stillbirths, number of children fathered.
- *Physical*: height, body mass index (BMI) (inverse normalised).
- *Cognitive*: fluid intelligence (Field ID: 20016), reaction time (inverse normalised, Field ID: 20023), time taken on pairs test (Field ID: 20131), numeric memory (inverse normalised, Field ID: 20240), age left education, number of years in education, has a degree.
- *Socioeconomic*: in employment, unable to work (both Field ID: 6142), income (Field ID: 738), Townsend Deprivation Index (TDI) (Field ID: 189).

We used the Bonferroni method to calculate the p-value for statistically significant results, to correct for multiple testing. As we tested 22 traits, our Bonferroni-corrected p-value was 0.002.

3.3 Results

3.31 Rare deleterious variants in UKB

Although each gene individually accounts for extremely rare forms of DD and has a small burden of rare deleterious variants, together they account for a large portion of DD diagnoses and have a surprisingly high burden of rare deleterious variants in UKB. In 184,477 unrelated European individuals with WES data in UKB and across 599 autosomal dominant DD genes: 9103 individuals carry a rare ($n \leq 5$) LoF variant, 25,288 individuals carry a rare missense variant with REVEL > 0.7, and 79,959 individuals carry a rare synonymous variant. As the gene panel becomes smaller and more stringent, the burden of rare deleterious variants decreases; for example, 3602, 1327 and 167 individuals in UKB carry rare LoF variants in smaller more stringent subsets of 384, 125 and 25 monogenic DD genes, respectively. In 450,274 individuals with SNP-array data in UKB and across 69 known DD loci, 4922 individuals carry large deletions, and 7054 individuals carry large duplications.

3.32 Related sub-clinical phenotypes

We performed gene panel (including 599 autosomal dominant genes) and multigenic copy number (including 53 deletions/duplications syndromes) burden tests for 22 traits in UKB selected to be of relevance (in adults) to developmental phenotypes. Bonferroni-corrected significant associations were found across most phenotypes in individuals carrying likely damaging variants compared with the rest of the UKB cohort (**Table 3.1** and **Figure 3.3** and **3.4**). Individuals carrying these variants generally had lower cognitive performance than the rest of the cohort, with reduced fluid intelligence (LoF group beta: -1.059), slower reaction times (LoF group beta: +0.043), lower numeric memory scores (LoF group beta: -0.068) and longer pairs matching times (LoF group beta: +0.122). They also completed fewer years in education, left education at an earlier age and were less likely to have a degree. Medically, individuals were more likely to have reported having a mental health issue or been diagnosed

with either a childhood DD (including mild-severe intellectual disability, epilepsy, autism, ADHD, and congenital malformations) or an adult DD-related diagnosis (including schizophrenia and bipolar disorder). Individuals were also more likely to be shorter, have a higher BMI and have had fewer children (though the latter association was only significant in men). Individuals also had significant socioeconomic disadvantages, being less likely to be employed or be able to work, having a lower income and a higher Townsend Deprivation Index (TDI). Across all phenotypes tested, we observed a trend corresponding to the likely deleteriousness of the variants; the largest effect was generally observed in the group of individuals with multigenic deletions, followed by multigenic duplications, then LoF variants and finally missense variants in one (or more) DD genes. These trends were robust to using different CADD thresholds to select missense variants (**Table 3.2**) and to removing individuals with a diagnosed childhood developmental disorder ("Child DD", as defined in Methods, n = 3,132; see **Appendix 7.3.3**). In contrast, individuals with only rare synonymous variants in DD genes showed no statistically significant difference in any phenotype compared to the remainder of the cohort, as expected for likely benign variants, suggesting that most of the confounding caused by population sub-structure was appropriately controlled.

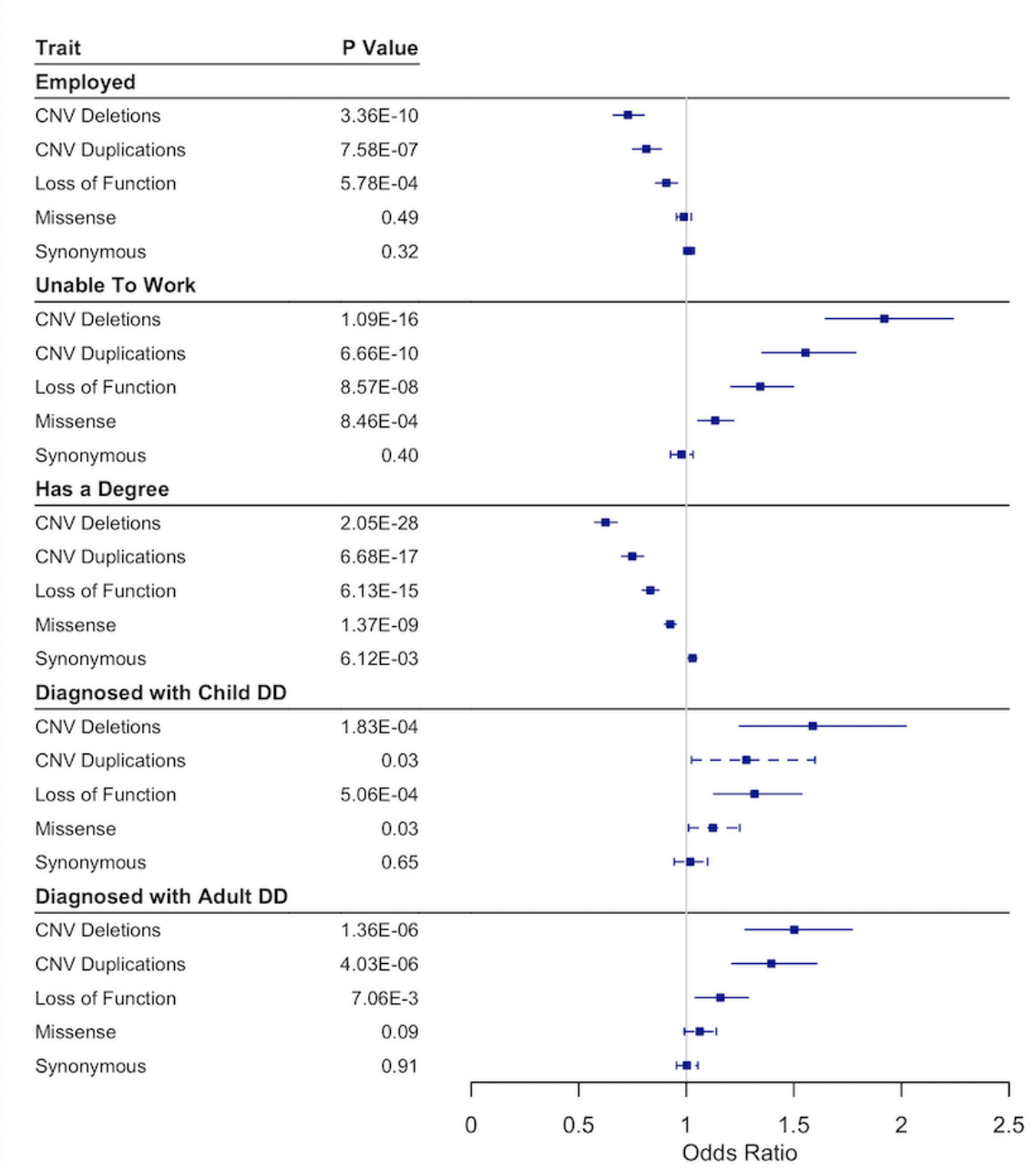


Figure 3.3. **Summary of gene panel association tests for carriers of likely deleterious variants in known autosomal dominant DD loci for binary traits.** Associations are shown for individuals carrying deletions or duplications overlapping 69 known DD syndromic loci, or rare ($n \leq 5$) LoF, missense (REVEL>0.7) or synonymous variants in any of 599 known autosomal dominant DD genes, compared with the remaining unrelated white Europeans in UKB. Lines indicate 95% confidence intervals. Unbroken lines indicate statistically significant, i.e. below Bonferroni-corrected p-value; dashed lines indicate above this value.

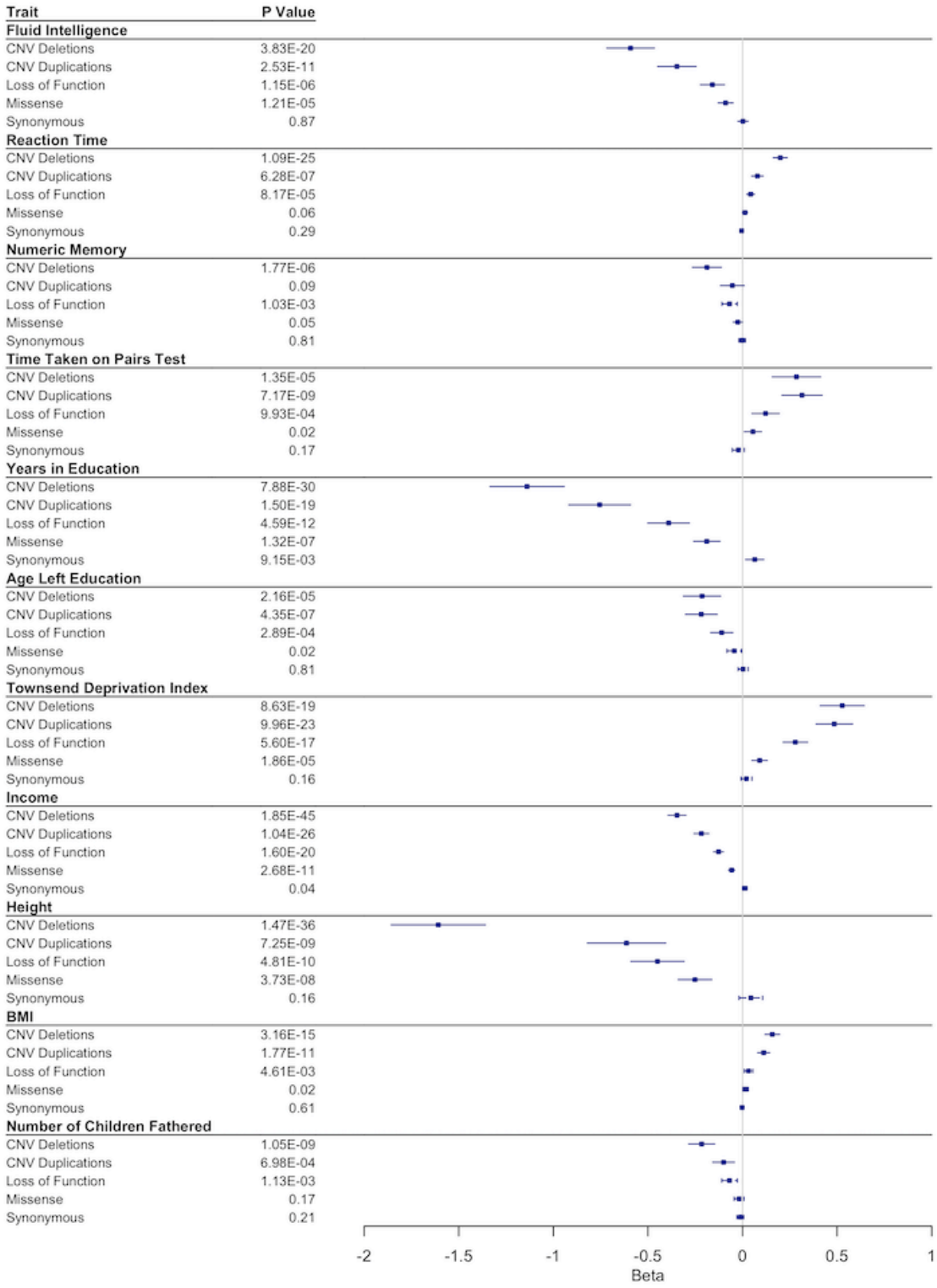


Figure 3.4: Summary of gene panel association tests for carriers of likely deleterious variants in known autosomal dominant DD loci for continuous traits. Associations are shown for individuals carrying deleterious or duplications overlapping 69 known DD syndromic loci, or rare (n≤5) LoF, missense (REVEL > 0.7), or synonymous variants in any of 599 known autosomal dominant DD genes, compared with the remaining unrelated white Europeans in UKB.

Dataset:	Deletions overlapping 69 DD loci		Duplications overlapping 69 DD loci		LoF variants in 599 DD genes		Missense variants in 599 DD genes		Synonymous variants in 599 DD genes	
Binary Traits	**OR**	**P Value**	**OR**	**P Value**	**OR**	**P Value**	**OR**	**P Value**	**OR**	**P Value**
In employment	0.728	3.356E-10	0.814	7.580E-07	0.907	5.778E-04	0.988	0.500	1.012	0.323
Have a degree	0.624	2.052E-28	0.748	6.684E-17	0.833	6.134E-15	0.925	1.368E-07	1.028	6.115E-03
Have an epilepsy diagnosis	1.689	2.179E-03	1.292	0.113	1.394	0.003	1.068	0.403	0.917	0.131
Diagnosed with Child DD	1.588	1.827E-04	1.279	0.030	1.316	5.056E-04	1.123	0.031	1.018	0.645
Diagnosed with Adult DD	1.502	1.359E-06	1.395	4.027E-06	1.158	7.061E-03	1.062	0.092	1.003	0.914
Is unable to work	1.921	1.093E-16	1.554	6.663E-10	1.344	8.573E-08	1.134	8.459E-04	0.977	0.403
Continuous Traits	**Beta**	**P Value**	**Beta**	**P Value**	**Beta**	**P Value**	**Beta:**	**P Value**	**Beta**	**P Value**
Fluid Intelligence	-0.592	3.834E-20	-0.347	2.534E-11	-0.159	1.152E-06	-0.089	1.207E-05	0.002	0.865
Number of years in education	-1.139	7.878E-30	-0.755	1.496E-19	-0.391	4.589E-12	-0.189	1.323E-07	0.064	0.009
Income	-0.346	1.850E-45	-0.217	1.042E-26	-0.127	1.599E-20	-0.058	2.675E-11	0.012	0.040
Reaction time	0.199	1.086E-25	0.079	6.277E-07	0.043	8.179E-05	0.013	0.060	-0.005	0.290
Pairs test score	0.285	1.345E-05	0.315	7.174E-09	0.122	9.928E-04	0.055	0.019	-0.022	0.172
Townsend Deprivation Index	0.527	8.628E-19	0.485	9.962E-23	0.279	5.596E-17	0.090	1.855E-05	0.020	0.160
Age left education	-0.214	2.158E-05	-0.218	4.345E-07	-0.110	2.892E-04	-0.044	0.025	0.003	0.806
Height	-1.608	1.474E-36	-0.613	7.254E-09	-0.449	4.809E-10	-0.251	3.725E-08	0.044	0.164
Reported a mental health issue	0.071	1.629E-03	0.023	0.222	0.041	1.047E-03	0.015	0.053	-0.001	0.848
Numeric memory score	-0.188	1.765E-06	-0.054	0.096	-0.068	1.032E-03	-0.025	0.053	-0.002	0.813
BMI	0.157	3.164E-15	0.112	1.766E-11	0.032	4.611E-03	0.016	0.024	-0.003	0.608
Number of children fathered	-0.216	1.048E-09	-0.100	6.985E-04	-0.069	1.135E-03	-0.018	0.168	-0.011	0.210
Number of pregnancies	-0.041	0.358	-0.039	0.292	-0.043	0.076	-0.024	0.120	0.007	0.499
Number of stillbirths	0.005	0.381	0.009	0.066	0.004	0.245	0.004	0.039	0.001	0.430

Table 3.1: Gene panel association test results: 22 phenotypes tested in individuals in UK Biobank carrying deletions or duplications overlapping 69 known DD syndromic loci, or rare (n≤5) LoF, missense (REVEL>0.7) or synonymous variants in any of 599 known autosomal dominant DD genes were tested.

	CADD > 20				CADD > 25				CADD > 30			
Phenotype	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**
Binary Traits:	**OR:**				**OR:**				**OR:**			
In employment	1.004	7.242E-01	0.981	1.027	0.992	5.362E-01	0.967	1.017	0.972	3.096E-01	0.920	1.027
Have a degree	0.969	9.256E-04	0.951	0.987	0.941	1.233E-08	0.922	0.961	0.929	1.295E-03	0.889	0.972
Have an epilepsy diagnosis	1.057	2.896E-01	0.954	1.172	1.006	9.151E-01	0.898	1.127	1.117	3.550E-01	0.883	1.414
Diagnosed with Child DD*	0.968	3.789E-01	0.902	1.040	1.010	7.990E-01	0.934	1.093	1.052	5.495E-01	0.890	1.244
Diagnosed with Adult DD*	1.007	7.833E-01	0.961	1.054	1.050	5.900E-02	0.998	1.105	1.107	6.111E-02	0.995	1.232
Is unable to work	1.055	3.448E-02	1.004	1.109	1.092	1.509E-03	1.034	1.153	1.088	1.514E-01	0.969	1.222
Continuous Traits:	**Beta:**				**Beta:**				**Beta:**			
Fluid Intelligence	-0.043	1.096E-03	-0.069	-0.017	-0.074	4.916E-07	-0.102	-0.045	-0.090	4.384E-03	-0.152	-0.028
Number of years in education	-0.087	1.922E-04	-0.132	-0.041	-0.153	2.754E-09	-0.203	-0.102	-0.207	1.817E-04	-0.315	-0.099
Income	-0.032	7.494E-09	-0.043	-0.021	-0.050	8.536E-16	-0.062	-0.038	-0.065	1.187E-06	-0.091	-0.039
Reaction time	0.017	1.113E-04	0.008	0.026	0.018	2.537E-04	0.008	0.028	0.018	9.647E-02	-0.003	0.038
Pairs test score	0.034	2.687E-02	0.004	0.063	0.042	1.347E-02	0.009	0.075	0.094	9.163E-03	0.023	0.165
Townsend Deprivation Index	0.070	3.479E-07	0.043	0.097	0.093	8.451E-10	0.063	0.123	0.136	2.778E-05	0.073	0.200
Age left education	-0.026	3.721E-02	-0.051	-0.002	-0.035	1.269E-02	-0.063	-0.007	-0.012	7.012E-01	-0.071	0.048
Height	-0.144	1.098E-06	-0.202	-0.086	-0.206	3.383E-10	-0.271	-0.142	-0.283	5.914E-05	-0.421	-0.145
Reported a mental health issue	0.000	9.633E-01	-0.010	0.010	0.009	1.013E-01	-0.002	0.021	0.026	3.312E-02	0.002	0.050
Numeric memory score	-0.012	1.563E-01	-0.028	0.005	-0.019	4.151E-02	-0.037	-0.001	-0.046	2.207E-02	-0.085	-0.007
BMI	0.009	5.943E-02	0.000	0.018	0.017	8.016E-04	0.007	0.027	0.010	3.597E-01	-0.012	0.032
Number of children fathered	-0.022	8.959E-03	-0.039	-0.006	-0.015	1.084E-01	-0.034	0.003	-0.021	2.964E-01	-0.061	0.019
Number of pregnancies	-0.023	2.334E-02	-0.042	-0.003	-0.021	5.502E-02	-0.043	0.000	-0.019	4.318E-01	-0.066	0.028
Number of stillbirths	0.002	1.565E-01	-0.001	0.004	0.002	2.058E-01	-0.001	0.005	0.001	7.943E-01	-0.005	0.007

Table 3.2: Gene panel association test results for deleterious missense variants across different CADD bins

3.33 Highly penetrant genes

We repeated our association analysis with smaller, more stringent, subsets of 325, 125 and 25 known DD genes. Interestingly, even within the most stringent subset of 25 genes that are thought to be highly penetrant causes of DD via haploinsufficiency, with >30 *de novo* LoF mutations identified in 31,058 DD probands[373], we were able to identify 167 individuals in UKB who had a high confidence LoF variant in one of these genes. We observed similar trends to the full 599 gene panel for LoF variants in smaller subsets of genes cause DD by haploinsufficiency, with the group overall exhibiting mild DD-related phenotypes, though the results were less significant due to the smaller number of individuals carrying likely LoF variants (**Table 3.3)**. Nonetheless, a Bonferroni-corrected significant result was seen across all gene subsets for shorter stature, reduced chance of having a degree and increased TDI; lower fluid intelligence, lower income, higher BMI, and an increased chance of being diagnosed with a child DD also remained nominally significant even in the 25 gene subset. We also performed single gene burden testing but were underpowered to find any significant associations for most genes due to the small number of individuals and likely mild phenotypic effects in UKB. Interestingly, despite previously reaching genome-wide significance for enrichment of damaging *de novo* mutations, *MIB1* had the largest number of individuals carrying likely LoF variants in UKB (n=260), more than the 25 most stringent genes combined, but showed no associations with any DD-related phenotypes. The gene also has almost double the number of LoF variants observed versus expected in gnomAD (https://gnomad.broadinstitute.org/gene/MIB1), and thus appears to be remarkably unconstrained and thus may not be a true haploinsufficient DD gene.

Dataset	599 Gene Set		325 Gene Set		125 Gene Set		25 Gene Set	
Phenotype	**-**	**P Value**	**-**	**P Value**	**-**	**P Value**	**-**	**P Value**
Binary Traits:	**Odds Ratio:**		**Odds Ratio:**		**Odds Ratio:**		**Odds Ratio:**	
In employment	0.907	5.778E-04	0.847	2.144E-04	0.759	1.583E-04	0.802	3.248E-01
Have a degree	0.833	6.134E-15	0.798	9.390E-10	0.763	6.424E-06	0.597	9.110E-03
Have an epilepsy diagnosis	1.394	2.690E-03	1.543	7.964E-03	0.830	6.004E-01	.	.
Diagnosed with Child DD*	1.316	5.056E-04	1.581	4.889E-05	1.612	8.029E-03	2.582	3.833E-02
Diagnosed with Adult DD*	1.158	7.061E-03	1.234	1.082E-02	1.630	3.796E-05	1.308	5.231E-01
Is unable to work	1.344	8.573E-08	1.318	1.361E-03	1.596	2.818E-04	1.481	3.504E-01
Continuous Traits:	**Beta:**	-	**Beta:**	-	**Beta:**	-	**Beta:**	-
Fluid Intelligence	-0.159	1.152E-06	-0.196	1.525E-04	-0.316	2.113E-04	-0.565	3.605E-02
Number of years in education	-0.391	4.589E-12	-0.552	4.614E-10	-0.551	1.259E-04	-0.367	4.251E-01
Income	-0.127	1.599E-20	-0.173	9.557E-16	-0.217	6.970E-10	-0.244	2.695E-02
Reaction time	0.043	8.179E-05	0.058	6.250E-04	0.087	1.624E-03	0.128	1.451E-01
Pairs test score	0.122	9.928E-04	0.145	1.270E-02	0.099	2.931E-01	-0.019	9.483E-01
Townsend Deprivation Index	0.279	5.596E-17	0.435	9.558E-17	0.663	5.178E-15	0.810	2.717E-03
Age left education	-0.110	2.892E-04	-0.143	2.396E-03	-0.149	4.935E-02	-0.324	1.824E-01
Height	-0.449	4.809E-10	-0.700	6.278E-10	-0.509	5.515E-03	-2.122	3.001E-04
Reported a mental health issue	0.041	1.047E-03	0.040	4.293E-02	0.081	1.177E-02	0.109	2.873E-01
Numeric memory score	-0.068	1.032E-03	-0.099	3.199E-03	-0.183	1.138E-03	-0.259	1.218E-01
BMI	0.032	4.611E-03	0.048	6.804E-03	0.092	1.294E-03	0.157	8.762E-02
Number of children fathered	-0.069	1.135E-03	-0.099	2.546E-03	-0.088	9.567E-02	0.187	2.338E-01
Number of pregnancies	-0.043	7.618E-02	-0.041	2.832E-01	-0.081	1.938E-01	-0.540	1.227E-02
Number of stillbirths	0.004	2.447E-01	0.004	3.932E-01	-0.005	5.212E-01	-0.029	2.987E-01

Table 3.3: Gene panel association test results for LoF variants across different gene subsets

3.34 Rare and common variants

We investigated the effect of allele count (AC) on the phenotypic effect of LoF variants in our largest gene panel (599 autosomal dominant DD genes). Specifically, we performed association tests with 16 DD-related traits that were significant in the previous analysis for groups of individuals with rare LoF variants in these genes that were present in just a single individual in UKB, compared with variants seen 5, 10, 50 or 100 or fewer times (**Figure 3.5**). The group of individuals who had the rarest variants (AC=1) had the largest phenotypic effect change compared to the rest of the cohort, though the results were generally not significant due to low numbers. However, across the phenotypes tested, both the effect size and the p-value decreased as the AC increased, suggesting either that the more common variants have a milder effect on phenotype, or that more common variants are benign and are simply diluting the effect of rare pathogenic variants. No difference was observed between the effect of LoF variants in the first or second half of genes. In addition, 295 individuals had LoF variants that were previously classified as “likely pathogenic” or “pathogenic” in ClinVar, but no significant difference was detectable in their phenotypes compared with the remainder of the LoF variant carrier group.

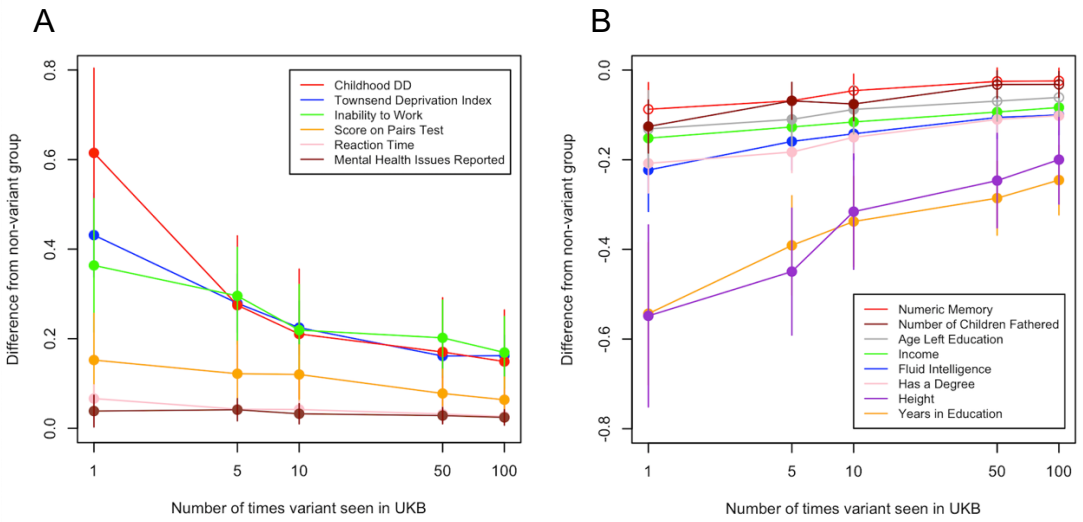


Figure 3.5: Change in phenotype associations for individuals with a LoF variant in 599 known autosomal dominant DD genes versus different minor allele counts. Associations are grouped by whether the effect of MAC = 1 LoF variants either (a) decreases or (b) increases the phenotype.

3.4 Discussion

We have shown that rare, potentially damaging variants in genes and loci known to cause autosomal dominant DD are present in adults in UK Biobank and result in a mild developmental phenotype. Individuals carrying these variants have notably reduced cognitive abilities and a lower socioeconomic status. Gene panel association tests suggest a strong and consistent trend for increasing phenotypic effects with rarer and more damaging variants. Although it is Impossible to disentangle incomplete penetrance and variable expressivity in a population study, our findings are consistent with similar studies[16,17,204,380–382] showing reduced penetrance of rare damaging variants in monogenic forms of DD in clinically unselected population cohorts. Moreover, our results are robust to removal of individuals diagnosed with a childhood developmental disorder, suggesting that fully penetrant individuals are not driving the signal.

We note that the variants identified in UKB are not necessarily the same ones that have been identified previously in clinical cases, and indeed very few of those we identified had previously been annotated in ClinVar[383]. We also note that our dataset likely includes some predicted LoF variants that do not actually result in a loss of function (either due to technical false positives or biological

rescue through translation re-initiation, alternative splicing, *etc*). Nonetheless, these issues are common to any clinical or research scenario where variants are prioritised from WES data, and our findings were robust when limited to likely LoF variants in a subset of 384 DD genes that act via a haploinsufficiency mechanism. The fact that our findings are robust to smaller, more stringent subsets of genes also suggests that the low effect sizes cannot simply be explained by a subset of low penetrance (or non-causal) DD genes. Furthermore, rare predicted LoF variants were found in individuals in genes that were thought to be fully or nearly fully penetrant causes of very well-established developmental syndromes, but without the full clinical phenotype that would be expected, suggesting that there is a range of penetrance and expressivity in the general population.

Despite the large size of UKB, we were limited by the number of individuals of European ancestry carrying rare damaging variants in these genes, which meant some of our analyses were under-powered to show a significant effect. We were also limited by the clinical and phenotypic data available on these individuals, all of whom were over 40 years of age at recruitment; evaluation and diagnosis of DD was much less routine when these individuals were children, and is less likely to be recorded in the HES codes of older adults. Nonetheless, when found in an appropriate clinical paediatric setting, rare damaging variants in these genes are widely considered diagnostic for DD, and thus they might not be expected to be present in a population cohort. Our results suggest that, although the penetrance of variants across these genes is lower than would be expected from previous clinical studies, they do still exert a phenotypic effect on adults in the general population who are nonetheless healthy enough, and have sufficient capacity, to volunteer to participate in a biobank.

Genes and loci that cause monogenic DD have historically been identified almost exclusively through clinical cohorts of affected children and families, and their effect on adults in the general population has not previously been evaluated. While clinical studies may overestimate the penetrance of such rare variants, population cohorts like UKB are likely to underestimate the penetrance, due to ascertainment bias towards healthy individuals[23]. The

penetrance and expressivity of variants in these genes could be affected by a number of different modifiers, including genetic variants in other genes, regulatory variants affecting gene expression, somatic mosaicism, and accumulated environmental factors[45]. The latter is particularly relevant when considering the effect of damaging variants in DD genes on adults. It is interesting to note that, unlike most traits, the heritability of intelligence (i.e. general cognitive ability) increases dramatically with age[384], suggesting a major role for gene-environment interactions as individuals become better able to select, modify and optimise their environment. Further research is needed into the penetrance of rare, damaging variants in the general population using larger datasets, which may allow modifiers to be investigated to help explain why some individuals are more severely affected by particular genetic conditions than others.

4. Chapter four: Genetic modifiers of rare genetic variants in UK Biobank

4.1 Introduction

Ascertaining whether rare genetic variants cause a monogenic phenotype can be challenging due to incomplete penetrance and variable expressivity[385]. Many rare variant studies use clinical or familial cohorts that can overestimate the penetrance of damaging causal variants[17]. The presence of such rare, putatively damaging variants in healthy population cohorts[79] can provide a lower boundary for estimates of penetrance, and individuals in both clinical and population cohorts display a spectrum of phenotypic variability caused by similar or identical variants in the same gene[385,386]. Previous research has suggested that common genetic variants can modify the penetrance or expressivity of phenotypes caused by rare genetic variants[9,387,388], potentially through the liability threshold model, which posits that a certain threshold of disease susceptibility needs to be crossed before clinically-diagnosable disease manifests[235,389–391]. Some damaging rare variants may reach this threshold alone, resulting in a monogenic disease phenotype with 100% penetrance, while other variants may need additional genetic, environmental, or other modifiers to reach this threshold[389]. In certain diseases, common variant burden has been shown to confer a risk similar to that of a deleterious monogenic variant, where the highest polygenic risk may be equivalent to that conferred by a monogenic variant[274,392]. As the effect of each individual common variant is very small[393], aggregating them together as a polygenic score (PGS) has become a widely used method for predicting overall related risk from common genetic variation[275,394], and combining PGS with rare pathogenic variant status could improve individual disease prediction[395,396].

Previously, we showed that rare, predicted loss-of-function (LoF), deleterious missense and large copy number variants (CNVs) in genes and loci linked with severe monogenic developmental disorders (DD) can have milder, sub-clinical effects in the general population[397]. Related common variant burden has been shown to affect the phenotype in carriers of such variants[9,236,398], suggesting

that the cumulative effect of common variants can modify the penetrance of rare variants in such phenotypes even if the primary cause is thought to be monogenic. While the impact of common variants on overall phenotypic expressivity has been examined for several neuropsychiatric[236,312,399] and other disease cohorts[80,282,400], the modification of rare variant penetrance by other rare genetic variants has not been widely investigated due to the very large cohort sizes required. Here, we present an analysis of common and rare variant burden in 419,854 adults from the UK Biobank (UKB)[28]. We investigate individuals carrying a rare LoF variant in genes and loci where similar variants are known to cause monogenic DD, and use related polygenic scores and additional rare variant burden to examine the effect on a number of related cognitive phenotypes and socioeconomic traits. We show that rare variant burden across these loci and PGS for Educational Attainment (EA-PGS) has an additive effect on the phenotype. Our results demonstrate that both additional rare and common genetic variants linked to relevant traits can contribute towards the variable expressivity of rare, predicted large-effect variants in known monogenic disease genes.

4.2 Methods

4.21 UK Biobank Cohort

The UKB cohort has been described in previous chapters. We used exome sequencing and microarray data from individuals in UKB who were of genetically defined European ancestry (N = 419,854) in our analysis.

4.22 Gene and variant selection

We used the clinically curated Developmental Disorders Gene2Phenotype Database (DDG2P) to select genes known to cause monogenic DD[37,397]. The database (accessed from https://www.ebi.ac.uk/gene2phenotype/ on 27 November 2020) was constructed and clinically curated from published literature and provides information relating to genes, variants and phenotypes associated with DDs, including mode of inheritance and mechanism of pathogenicity. We included all genes that had been annotated as monoallelic (i.e., autosomal dominant) with an evidence level of “confirmed” or “probable” (n=599). From this gene set we identified carriers of rare (allele count $\leq$ 5)

LoF[401] or deleterious missense (REVEL>0.7)[374] variants. Carriers of multigenic CNVs were also included where the variant overlapped known syndromic DD-related loci[378,379], as described previously[397]. For quality control purposes, anyone with a variant with a depth below 10 or a variant allele frequency (VAF) below 0.3 was removed.

We selected two functional classes of variant in canonical transcripts based on annotation by the Ensembl Variant Effect Predictor (v104)[37]:

(1) likely deleterious LoF variants: we defined a LoF variant as one that is predicted to cause a premature stop, a frameshift, or abolish a canonical splice site; only those variants deemed to be high confidence by the Loss-Of-Function Transcript Effect Estimator (LOFTEE), and fell outside of the final exon were retained (https://github.com/konradjk/loftee); and

(2) likely deleterious missense variants: missense variants with a REVEL score > 0.7.

Individuals with >1 variant within a 40bp window in the same gene were counted once.

In addition, we used SNP-array data from 488,377 genotyped individuals in UKB and PennCNV[377] (version 1.0.4) to detect multigenic CNVs overlapping 69 published CNVs strongly associated with developmental delay, as described previously[397].

4.23 PGS calculation

We calculated related polygenic scores using summary statistics and weighted allele effects from genome-wide association studies (GWAS) for every individual in UKB with European ancestry. We used previously published summary statistics containing 3952 SNPs for the Educational Attainment (EA) PGS, with data from a large cohort meta-analysis, Okbay *et al.* 2022[402] to create the EA-PGS. The EA-PGS was calculated as $\sum_i w_i g_i$, where w_i is the weight (effect size) for SNP I and g_i is the genotype (number of effect alleles, 0-2) at SNP i. The SNP weightings were the regression coefficients obtained from the most recently reported GWAS as mentioned above.

Sensitivity analysis: As the previously published large meta-cohort analysis included participants from UKB among many other cohort studies, we

calculated a secondary EA-PGS from an previous publication that did not use any participants from UKB in the identification of the 74 educational ability related SNPs and their consequential calculation of associated summary statistics[403]. Related results from the sensitivity analysis can be found in **Appendix 7.4.1** and **Appendix 7.4.2**.

Additional PGS: We calculated further PGSs related to our traits of interest, for cognitive ability, mathematical ability, and intelligence from previously published summary statistics and weighted allele effects from GWAS, again for every individual of European ancestry in UKB. Cognitive and mathematical ability weighted SNPs were obtained from Genç et al (2021)[393], and intelligence related SNPs were obtained from Savage et al (2018)[404], and calculated in the same way as the EA-PGS above. Schizophrenia and Bipolar PGS were downloaded from UKB[394].

4.24 Statistical analysis

We performed gene panel burden tests across our 599 gene subset, with association tests limited to individuals in UKB who had genetically defined European Ancestry due to the well-recognised biases in PGS performance in other ancestries[405,406].

Phenotypes of interest were selected from self-reported questionnaires or results from cognitive related tests undertaken through UKB, based on likely relevance to cognitive, behavioural, reproductive, and socio-economic effects within neurodevelopmental disorders. Medical-related phenotypes were categorized using ICD9 and ICD10 codes and self-reported questionnaire responses as follows:

- *Medical*:
    - ever reported a mental health issue (self-reported through questionnaire or ICD10 codes F40-F48, F50,F51, F53, F54, F99, G47 and R45, or ICD9 codes 300, 307-309, 311, 780.5);
    - diagnosed with “Child DD” (including intellectual disability (ICD10 codes F70-73), epilepsy (G40), developmental disorders (F80-84, F88-F95, F98, R62, R48, and Z55) and congenital malformations (Q0-99));

    - diagnosed with an “Adult Neuropsychiatric” condition (including schizophrenia, (self-reported or ICD10 codes F20-29) and bipolar disorder (self-reported or ICD10 codes F30-F39).
- *Reproductive*: never a parent, never a father, never pregnant.
- *Physical*: height
- *Cognitive*: fluid intelligence (Field ID: 20016), reaction time (inverse normalised, Field ID: 20023), time taken on the pairs matching test (averaged, Field ID: 20133), numeric memory (inverse normalised, Field ID: 20240), age left education, number of years in education, has a degree.
- *Socioeconomic*: in employment, unable to work (both Field ID: 6142), income (Field ID: 738), Townsend Deprivation Index (TDI) (Field ID: 189).

The list of ICD9 and ICD10 codes used to generate the defined groups is listed in **Appendix 7.4.3**.

We controlled for age, sex, recruitment centre and 40 principal components. Variant burden tests were performed using STATA (version 16.0), using linear regression for continuous phenotypes, and logistic regression for binary phenotypes. Associations were tested between individuals with an identified rare variant in any of these DDG2P genes and the remainder of the European UKB population. EA-PGS quintiles were defined using the entire cohort of European UKB. When testing across PGS quantiles, each group was tested against individuals in the middle quintile (i.e. 40-60% EA-PGS) who were not identified as being carriers of likely deleterious rare variants in the DDG2P gene subset. When testing associations within specific types of variants, similarly, the comparison group was those with were not identified as being carriers of likely deleterious variants. When testing smaller subgroups of individuals, those who had previously been identified as putatively deleterious variant carriers were removed from the comparison group. To define phenotypic “deviators”, we used the highest and lowest scores of fluid intelligence scores (0 and 1 versus 11, 12 and 13), and the top and bottom category for qualifications (no qualifications recorded versus having a degree).

4.3 Results

4.31 Additional rare variant burden

We first investigated whether DD-related phenotypes could be modified amongst rare DD variant carriers by the presence of additional rare LoF or damaging missense variants in the same set of DDG2P genes. In UKB, 50,395 (12%) individuals carry a single rare likely deleterious variant overlapping one of the 599 autosomal dominant DDG2P genes (12,153 LoF and 35,603 missense) or syndromic DD loci (1127 large deletions and 1512 large duplications); an additional 3831 individuals carry two rare DD variants, and 219 individuals have three or more putatively deleterious rare variants across these DD loci. The highest overall rare variant burden across the DD loci was five, which was seen in two individuals with three missense variants and two LoF variants each (**Table 4.1**). We performed regression analysis to test associations between number of rare variants in DD genes and the 15 DD-related traits and diagnoses, using linear regression for continuous traits (**Figure 4.1**) and logistic regression for binary traits (**Figure 4.2**). Increasing rare variant burden correlated with a larger change away from the average UKB participant in several DD-related phenotypes, including lower fluid intelligence, shorter stature, lower income, lower likelihood of being employed, lower likelihood of being a parent, and higher Townsend Deprivation Index (TDI). An increase in rare variant burden also correlated with a higher likelihood of having a DD-related diagnosis, and those with three or more rare DD variants were 2.1X (95% CI: 1.05-4.33, p = 0.03) and 1.7X (95% CI: 1.01=2.89, p = 0.04) more likely to be diagnosed with a child-DD or an adult neuropsychiatric-related diagnosis respectively than non-carriers (**Figure 4.2**). When we excluded those with rare missense variants and only considered LoF and large CNV carriers, we observed a larger change in phenotype, but the smaller number of individuals present in each group reduced the statistical power substantially; nonetheless, those with two or three rare variants were 2.2X (95% CI: 1.37-3.43, p = 0.0009) more likely to have a child DD related diagnosis than those without a LoF variant or CNV (**Appendix 7.4.4** and **7.4.5**).

Overall Rare Variant Burden in 599 DDG2P Genes in 419,865 UKB Individuals:	
Number of Variants	**Times Seen**
Overall	**54,445**
One	50,395
Two	3831
Three	206
Four	11
Five	2

Type of variant seen in 599 DDG2P Genes in 419,865 UKB Individuals:	
Type of Variant	**Times Seen**
CNV Deletions	2644
CNV Duplications	3348
LoF	13,989
Missense	39,211

Among individuals with one variant:	
Type of Variant	**Times Seen**
Individuals with one variant	50,395
CNV Deletions	1127
CNV Duplications	1512
LoF	12,153
Missense	35,603

Among individuals with two variants:	
Type of Variant	**Times Seen**
Individuals with two variants	3,831
Two CNV Deletions	1
CNV Deletion and CNV Duplication	1
CNV Deletion and LoF	47
CNV Deletion and Missense	123
Two CNV Duplications	0
CNV Duplication and LoF	37
CNV Duplication and Missense	164
Two LoF	344
LoF and Missense	1,268
Two Missense	1,846

Among individuals with three variants:	
Type of Variant	**Times Seen**
Individuals with three variants	206
Two CNV Deletions and one LoF	1
One CNV Deletion, One CNV Duplication, one Missense	1
One CNV Deletion, one LoF, one Missense	7
One CNV Duplication, one LoF, one Missense	3
Three LoF	9
Two Lof and one CNV Duplication	2
Two LoF and one Missense	46
Three Missense	59
Two Missense and one CNV Deletion	6
Two Missense and one CNV Duplication	11
Two Missense and one LoF	61

Among individuals with four variants:	
Type of Variant	**Times Seen**
Individuals with four variants	11
Three LoF, one Missense	2
Two LoF, two Missense	3
One LoF, three Missense	4
Four Missense	2

Among individuals with five variants:	
Type of Variant	**Times Seen**
Individuals with five variants	2
Two LoF, three Missense	2

Table 4.1: The number of individuals identified with a rare variant in any of 599 DDG2P genes in UK Biobank. The individuals are sorted by number of variants present, and variant type.

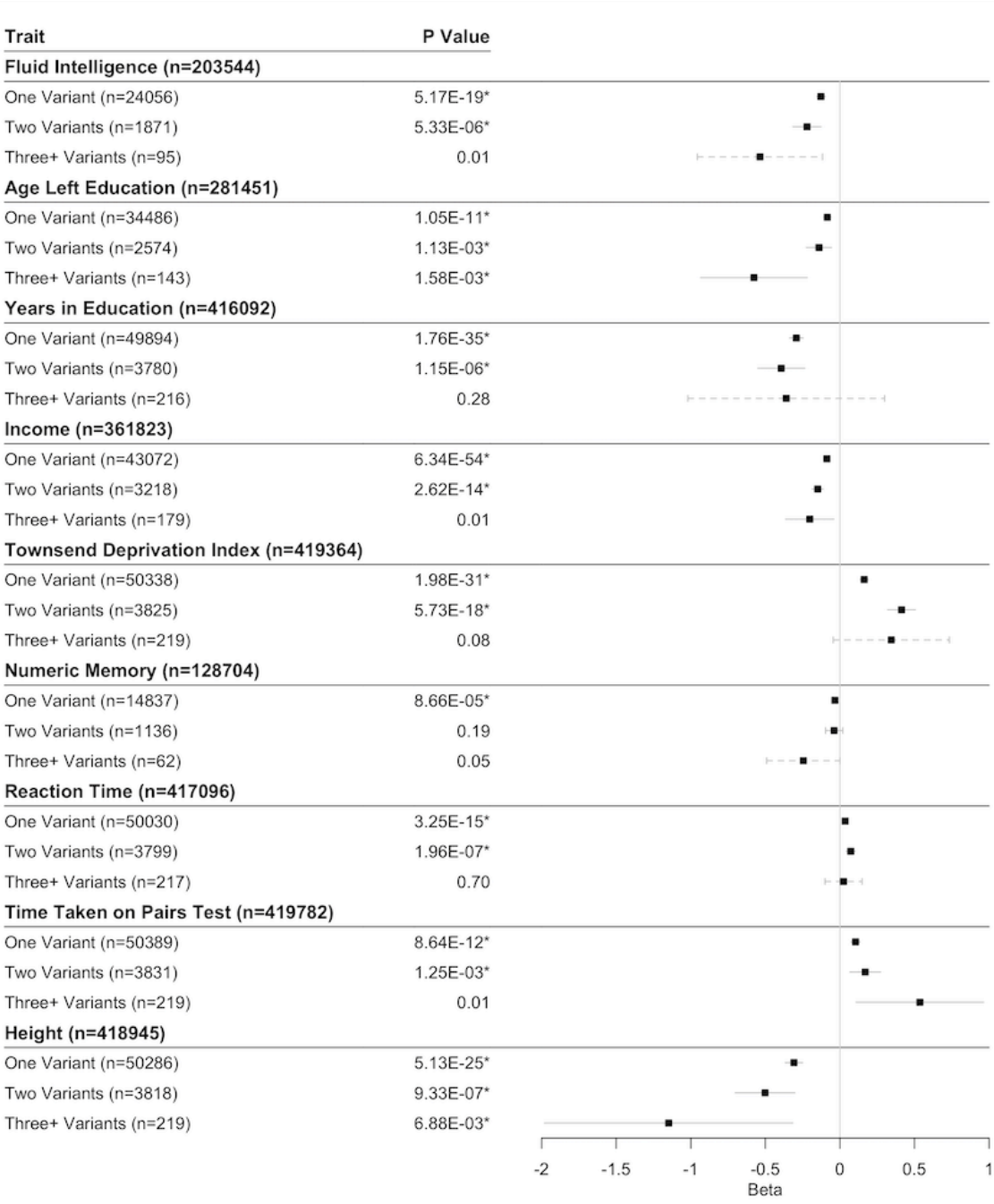


Figure 4.1: Associations of continuous traits in individuals carrying either 1, 2, or 3+ rare LoF, deleterious missense, or multigenic variants overlapping dominant DDG2P genes, compared with the rest of UK Biobank (i.e. non-carriers). Beta values were measured as follows: Fluid Intelligence = standardised unites (ranging from 0-13); Age Left Education and Years in Education are both measured in years; Height = cm; Reaction Time, Time taken on Pairs Test, Numeric Memory, Income, and Townsend Deprivation Index (TDI) = standard deviations from the mean. Bonferroni-corrected p value for multiple testing is 0.003. Lines indicate 95% confidence intervals.

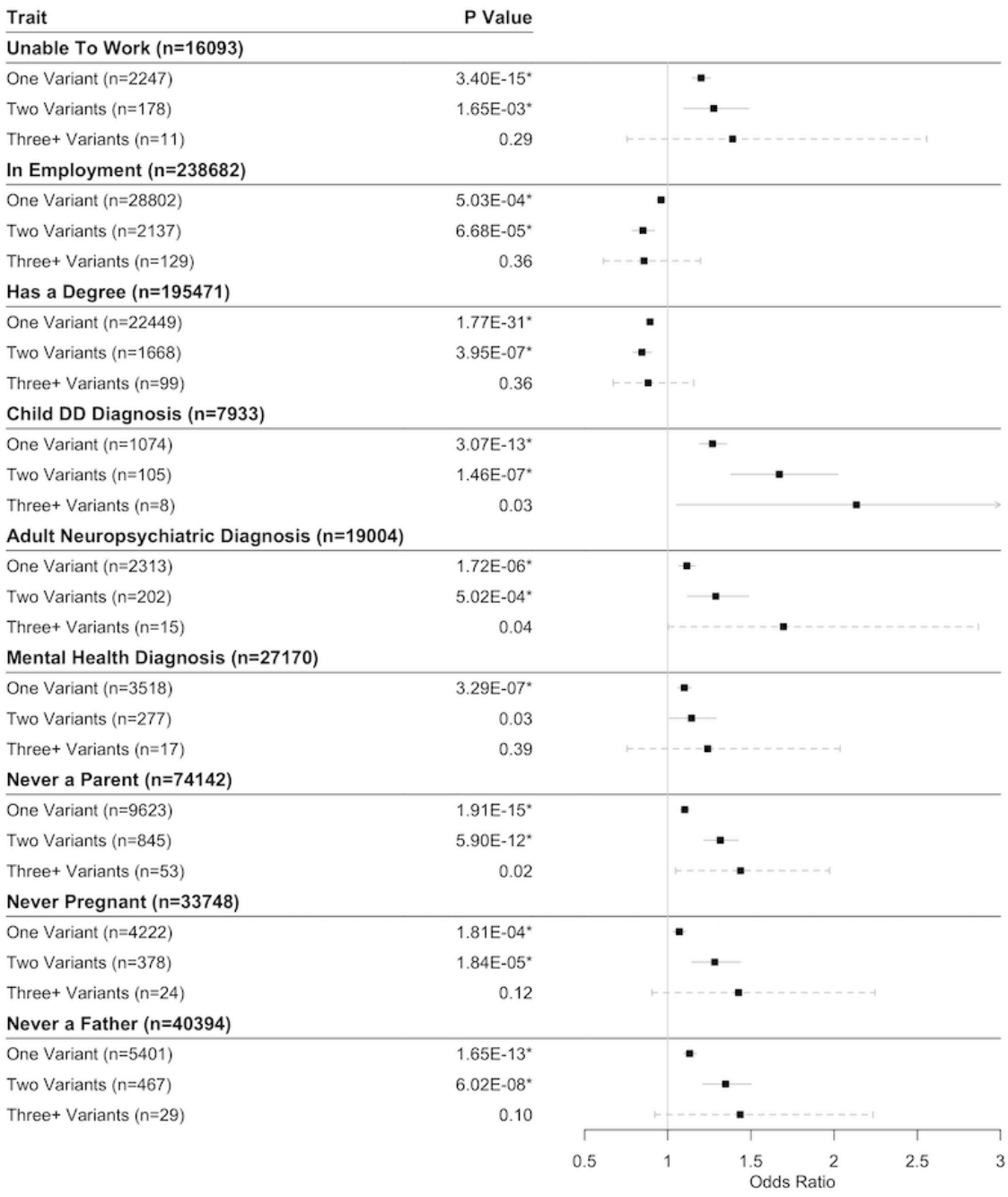


Figure 4.2: Associations of binary traits and diagnoses in individuals carrying 1, 2, or 3+ rare, LoF, deleterious missense, or multigenic variants overlapping dominant DDG2P genes, compared with the rest of UK Biobank.

4.32 Educational Attainment PGS

Next, we investigated the effect of common polygenic background on rare DD variant carriers[390]. We separated the UKB cohort into five EA-PGS quantiles and repeated the phenotype association tests with rare DD variant carrier

status. We saw a similar trend across all traits tested against the EA-PGS quintiles (**Figure 4.3**), with the direction of the PGS effect being the same in both carrier and non-carrier groups. Individuals who carried at least one rare variant showed a consistently larger change in fluid intelligence, years of education, employment and TDI across the PGS spectrum compared to the control group, with larger phenotypic effects observed in carriers of multiple rare DD variants (**Figure 4.4**). We observed similar trends when we repeated this analysis excluding missense variants (number of individuals in each group in **Appendix 7.4.5**, results in **Appendix 7.4.6**) or using a smaller subset of DD genes (**Appendix 7.4.7**), specifically those known to cause disease via haploinsufficiency (n= 325) or only those that reached genome-wide significance based on burden of *de novo* variants in ~31,000 DD cases (n=125)[373].

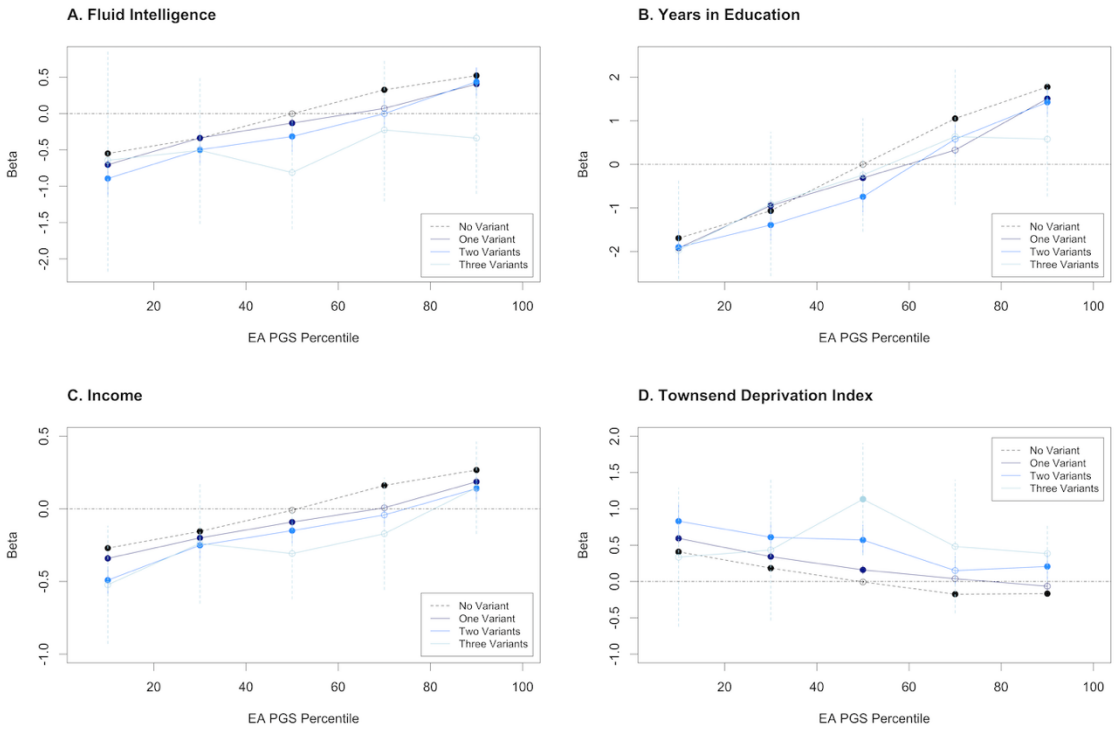


Figure 4.3: Additive effect of rare DD variant burden and EA-PGS on DD-related phenotypes. Change in (a) fluid intelligence, (b) income, (c) years in education and (d) Townsend Deprivation Index are shown versus EA-PGS quintile in UKB. Black dashed line shows the non-carriers of rare DD-related variants (n=365,409); dark/medium/light blue lines indicate carriers of 1, 2, or 3+ rare DD variants respectively (n=50,395, 3831, 219 respectively). Notably, within UKB, a high enough EA-PGS can compensate for the presence of a primary variant, and in most cases, any additional rare variant burden.

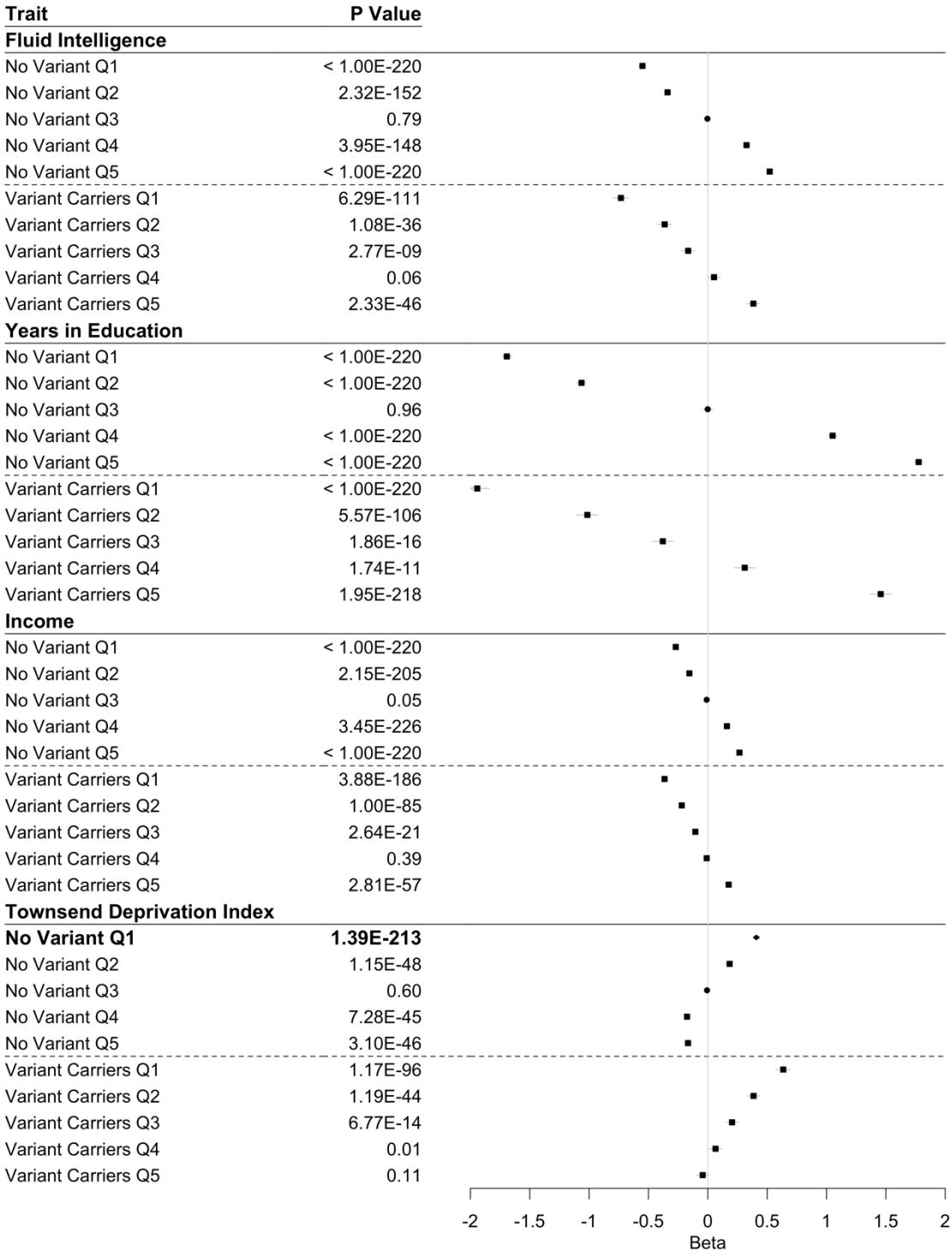


Figure 4.4: Trait results across EA-PGS quintiles for variant carriers and non-carriers. Circles represent the EA-PGS group that was the control, i.e. the comparison group for all the others.

For fluid intelligence, the difference in the mean score between the bottom and top EA-PGS quintiles equated to approximately 1 point on the 13-point scale, both for rare variant carriers and non-carriers in UKB. On average, rare DD

variant carrier status was equivalent to around a 20-percentile point decrease in EA-PGS, on average, with the result that an EA-PGS above the 70th centile was able to compensate for the effect of carrying a single rare DD variant on fluid intelligence. Importantly, rare variant carrier status and EA-PGS appear to have an additive effect when assessed against multiple related traits, with the effect of rare variants remaining similar throughout the EA-PGS spectrum. When we investigated rare variant classes within fluid intelligence scores, deleterious missense variant carriers reached parity with the control group at the 62nd EA-PGS percentile, LoF carriers at the 80th percentile and CNV duplication carriers at the 82nd percentile, while CNV deletion carriers never reached parity with the control group **(Table 4.2)**. We hypothesized that the EA-PGS could include SNPs in cis-regulatory regions of monogenic DDG2P genes, so we examined proximity between the 599 autosomal dominant DDG2P genes and 3952 SNPs included in the EA-PGS, using simulations to test whether the genes fall disproportionately close to the GWAS loci[407]. As expected, we found that found that the GWAS loci were closer to DDG2P genes than expected by chance ($p = 0.005$), suggesting that the large-effect rare variants and small-effect common variants may work through overlapping biological pathways.

Results per quintile for each variant type for fluid intelligence:					
Variant Type:	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Any rare variant:					
EA-PGS Quintile 1	-0.710	0.031	6.290E-116	-0.771	-0.650
EA-PGS Quintile 2	-0.327	0.030	7.523E-27	-0.387	-0.267
EA-PGS Quintile 3	-0.173	0.030	7.537E-09	-0.231	-0.114
EA-PGS Quintile 4	0.046	0.030	1.266E-01	-0.013	0.104
EA-PGS Quintile 5	0.387	0.029	3.971E-40	0.330	0.444
LoF variant					
EA-PGS Quintile 1	-0.837	0.059	1.053E-45	-0.953	-0.722
EA-PGS Quintile 2	-0.355	0.059	1.797E-09	-0.470	-0.239
EA-PGS Quintile 3	-0.180	0.058	1.957E-03	-0.294	-0.066
EA-PGS Quintile 4	-0.006	0.058	9.231E-01	-0.120	0.108
EA-PGS Quintile 5	0.367	0.055	1.930E-11	0.260	0.474
Missense variant					
EA-PGS Quintile 1	-0.650	0.037	1.975E-70	-0.721	-0.578
EA-PGS Quintile 2	-0.291	0.036	3.481E-16	-0.361	-0.221
EA-PGS Quintile 3	-0.145	0.035	3.575E-05	-0.214	-0.076
EA-PGS Quintile 4	0.078	0.035	2.506E-02	0.010	0.146
EA-PGS Quintile 5	0.425	0.034	2.177E-35	0.358	0.492
CNV Duplication					
EA-PGS Quintile 1	-0.757	0.124	9.829E-10	-1.000	-0.515
EA-PGS Quintile 2	-0.666	0.118	1.566E-08	-0.897	-0.435
EA-PGS Quintile 3	-0.462	0.117	7.488E-05	-0.690	-0.233
EA-PGS Quintile 4	-0.177	0.124	1.533E-01	-0.420	0.066
EA-PGS Quintile 5	0.299	0.112	7.737E-03	0.079	0.518
CNV Deletion					
EA-PGS Quintile 1	-1.272	0.141	1.653E-19	-1.548	-0.996
EA-PGS Quintile 2	-0.518	0.136	1.381E-04	-0.784	-0.252
EA-PGS Quintile 3	-0.564	0.126	8.049E-06	-0.811	-0.316
EA-PGS Quintile 4	-0.124	0.140	3.776E-01	-0.398	0.151
EA-PGS Quintile 5	-0.063	0.131	6.293E-01	-0.320	0.193
No variant					
EA-PGS Quintile 1	-0.525	0.016	2.070E-242	-0.556	-0.495
EA-PGS Quintile 2	-0.213	0.016	8.707E-43	-0.244	-0.183
EA-PGS Quintile 3	Comparison Set	.	.	.	.
EA-PGS Quintile 4	0.215	0.015	3.853E-44	0.185	0.246
EA-PGS Quintile 5	0.512	0.015	1.790E-242	0.482	0.542

Table 4.2: Rare variant association test results for individuals with a rare variant in any of the 599 DDG2P genes, grouped by the type of variant they carry, and their EA-PGS quintile.

4.33 Additional PGS

Cognitive-related PGS:

We additionally investigated alternative PGSs as other potential genetic modifiers, including cognitive abilities, mathematical abilities, and intelligence using previously published summary statistics[313,393,404], testing the additive effect on both fluid intelligence (**Figure 4.5**) and years in education (**Figure 4.6**). While we saw a similar trend among all three additional PGSs, they were less predictive than the original EA-PGS, possibly due to containing fewer numbers of SNPs, as the summary statistics used to develop them came from smaller cohort studies, and therefore may have less power than the large meta-cohort analysis used to calculate the summary statistics for the EA-PGS.

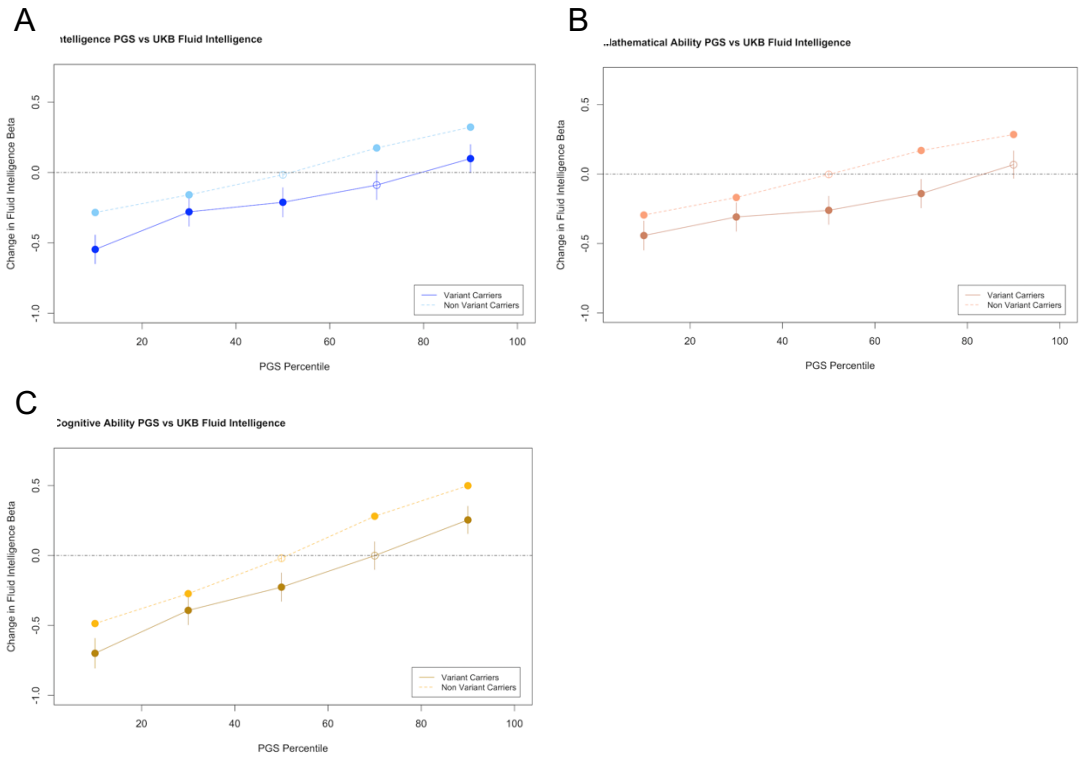


Figure 4.5: Additive effect of rare variant status and different polygenic scores on fluid intelligence test result scores: for A) Intelligence PGS, B) Mathematical Ability PGS, and C) Cognitive Ability PGS. The dashed line indicates the change in fluid intelligence results across the PGS quintiles for non-carriers, and the unbroken line indicates the change in fluid intelligence results across the PGS quintiles for any individual who carries a rare variant. Vertical lines indicate 95% confidence intervals, with filled symbols indicating statistically significant results, and unfilled symbols indicating those that did not reach statistical significance.

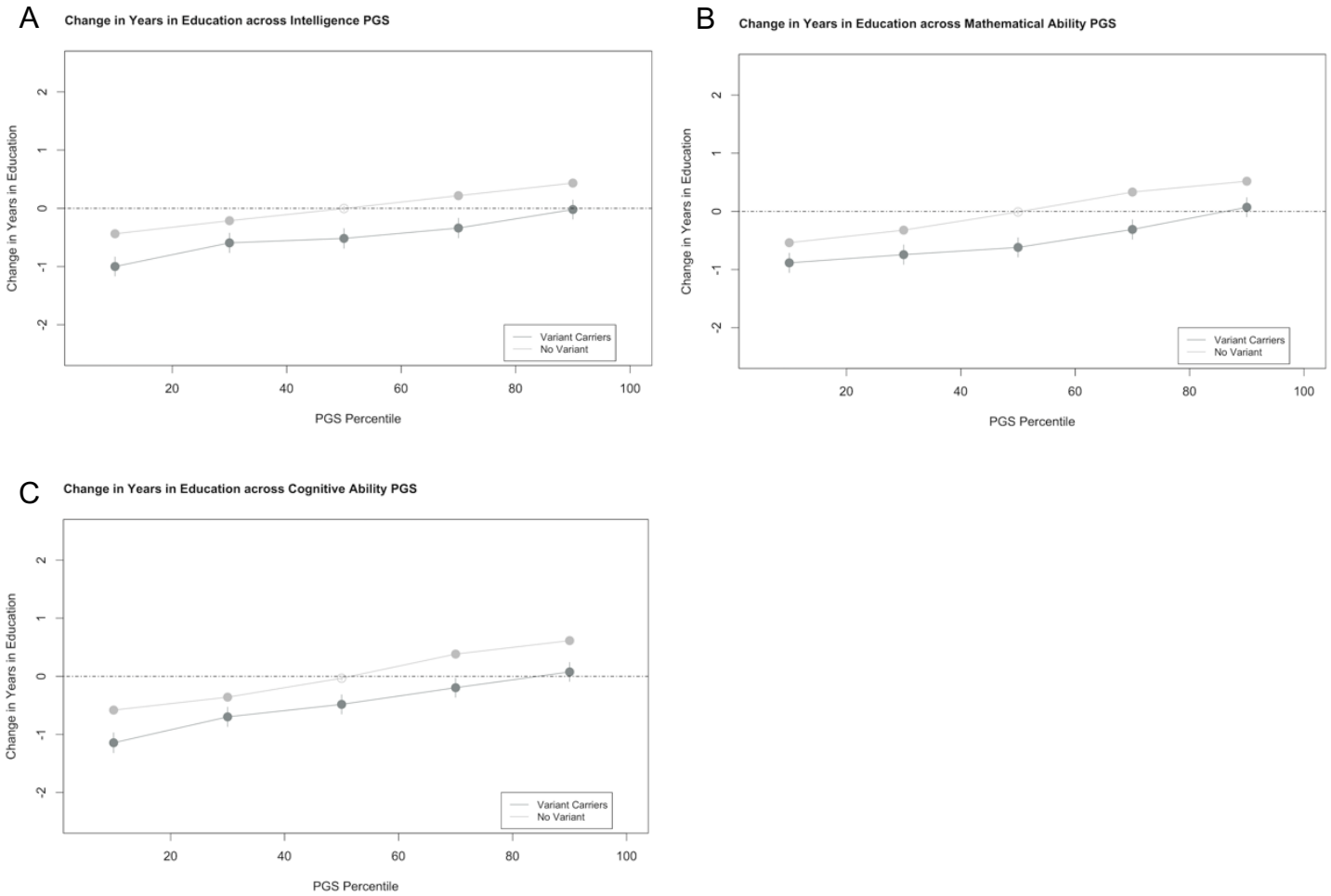


Figure 4.6: Additive effect of rare variant status and different polygenic scores on years spent in education: for A) Intelligence PGS, B) Mathematical Ability PGS, and C) Cognitive Ability PGS. The dashed line indicates the change in fluid intelligence results across the PGS quintiles for non-carriers, and the unbroken line indicates the change in fluid intelligence results across the PGS quintiles for any individual who carries a rare variant. Vertical lines indicate 95% confidence intervals, with filled symbols indicating statistically significant results, and unfilled symbols indicating those that did not reach statistical significance.

Clinically-related PGS: A large number of calculated PGS were previously released for individuals in UKB[394], of which two are related to our phenotypes of interest – PGS for bipolar disorder (BPD-PGS) and PGS for schizophrenia (SCZ-PGS).

Previous research has suggested that cognitive impairment is an important clinical component of schizophrenia[408], and tends to be present prior to onset of symptoms or diagnosis[408,409]. There was a small correlation seen between SCZ-PRS quintile and fluid intelligence score among both variant carriers and non-variant carriers in UKB, with fluid intelligence score decreasing as SCZ-PRS quintile increased, but no correlation between SCZ-PRS and years spent in education among the same group (**Figure 4.7**).

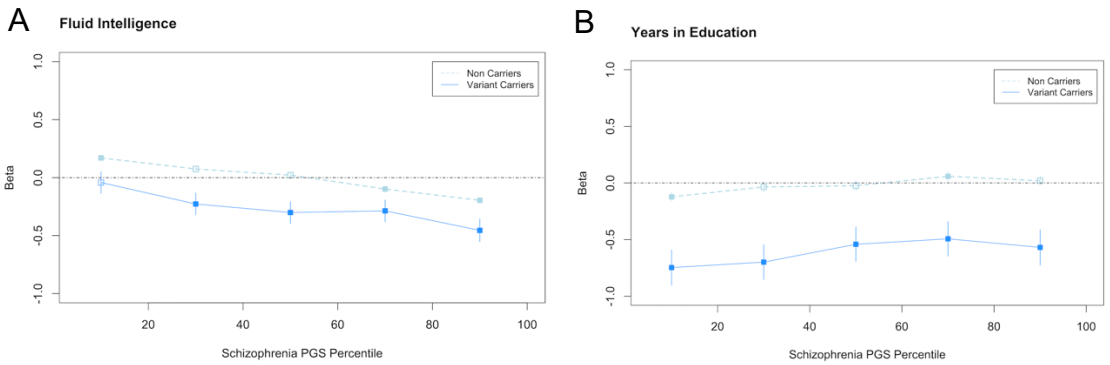


Figure 4.7: SCZ-PGS effect on A) fluid intelligence scores and B) years in education across polygenic quintiles in non-variant carriers and rare-variant carriers.

Previous research has suggested that there may not be a difference in cognitive or educational ability between individuals diagnosed with BPD prior to their diagnosis, and non-clinically affected individuals, or if there is, it may be much milder in severity when compared to individuals with schizophrenia[408,409]. Furthermore, some studies have suggested that individuals with bipolar disorder may fall at the extremes of the population when it comes to cognitive abilities or their time spent in education, and therefore taking an average does not show any cognitive differences to healthy controls[409–411]. When we tested individuals in BPD-PGS quintiles against our cognitive traits we saw no specific significant

trend in either fluid intelligence scores or years in education in individuals in UKB in either DD rare variant carriers or among non-carriers (**Figure 4.8**).

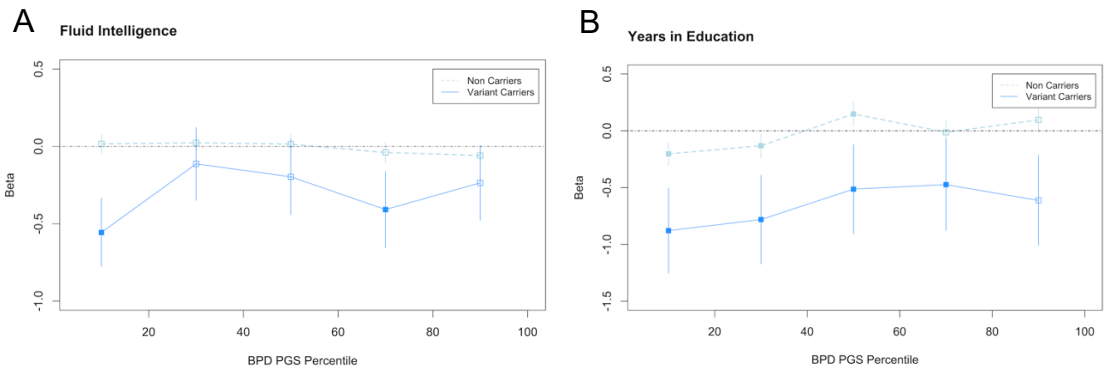


Figure 4.8: BPD-PGS effect on A) fluid intelligence scores and B) years in education across polygenic quintiles in non-variant carriers and rare-variant carriers.

We further tested both clinically-related PGS against our child and adult neuropsychiatric diagnosis groups. As our adult neuropsychiatric phenotype includes individuals with HES codes related to BPD and schizophrenia some correlation may be expected between having a higher PGS for either of these conditions and the odds ratio of being diagnosed with an adult neuropsychiatric condition in our cohort, or a higher likelihood of other or additional mental health issues, as SCZ-PRS has previously been shown to be correlated with the likelihood of mental health issues, with or without formal diagnosis[412]. However while those who fell within the top quintile of either of the scores did show a higher odds ratio of having an adult neuropsychiatric-related diagnosis, there was no obvious trend for either of the PGS groups. Both clinical PGS showed stronger correlation for diagnosis among non-variant carriers than variant carriers however (**Figure 4.9**).

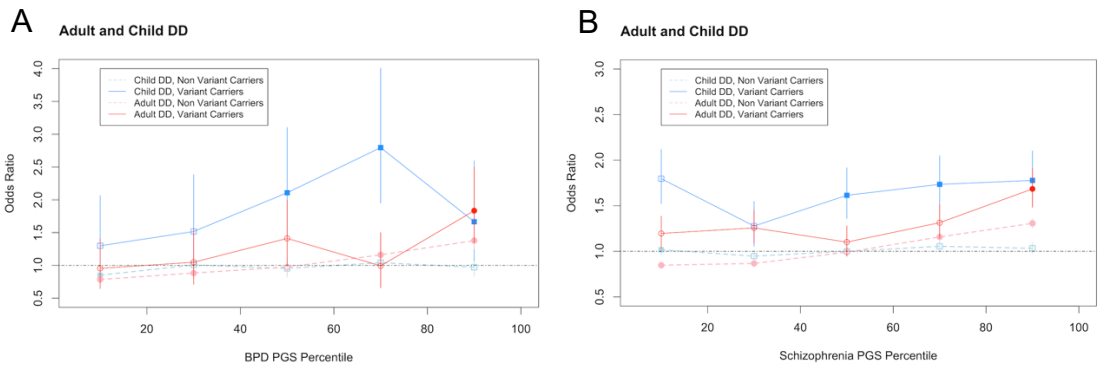


Figure 4.9: (a) BPD-PRS effect on the odds ratio of being diagnosed with either an adult neuropsychiatric-related diagnosis (pink/red), or a child DD related diagnosis (blue), when compared to the 40-60% PGS group. (b) SCZ-PRS effect on the odds ratio of being diagnosed with either an adult neuropsychiatric or child dd related diagnosis when compared to the 40-60% PGS group.

4.34 Phenotypic "deviators"

As the UKB cohort is known to be biased towards healthier, wealthier, and more educated individuals than the general population[23], we hypothesized that those individuals in UKB who carry a rare DD variant might also have a higher EA-PGS on average than the non-carrier control group, which could partially compensates for the potentially deleterious effects of the rare DD variant. Overall, we observed that individuals who carried at least one rare DD variant did indeed have a slightly higher EA-PGS percentile than non-carriers (t-test: difference =+2.1, 95%CI: 1.9-2.4, $p < 0.0005$), supporting this hypothesis. Furthermore, among the small number of individuals who achieved the top score on the fluid intelligence test (N=139), we observed that rare DD variant carriers (N=4) were depleted versus the rest of UKB (3% versus 13%, $p = 0.0002$) and had a substantially higher EA-PGS percentile than non-carriers (t-test: difference = +26.1, 95%CI: 1.8-50.3, $p = 0.04$).

Intrigued by the presence of these apparently highly intelligent rare DD variant carriers, we further investigated phenotypic "deviators" in whom the predicted genetic susceptibility is discordant with the observed phenotype[413], e.g. high EA-PGS but low fluid intelligence score and *vice versa* (**Figure 4.10**). This question has particular clinical relevance, as it has previously been suggested that individuals with familial disease could be prioritised for genetic testing

based on having a low-risk PGS, as they may be more likely to have a single large-effect causal variant than individuals with a high-risk PGS whose disease may be more polygenic[414,415]. To investigate this hypothesis, we further split the UKB cohort into deciles by EA-PGS and tested whether individuals whose low cognitive phenotype was discordant with their high EA-PGS were more likely to be rare DD variant carriers than the remainder of the UKB cohort. Individuals in the top EA-PGS decile but with low fluid intelligence (scores of 0 or 1 out of 13) were more likely to be rare DD variant carriers (OR: 1.68, 95% CI: 1.13-2.50, p = 0.01) (**Figure 4.11**), when compared to those in the same EA-PGS decile who did not have a low fluid intelligence score, as were those in the top EA-PGS decile who had no educational qualifications on record (OR: 1.22, 95% CI: 1.10-1.35, p = 0.00006) (**Figure 4.12**). When separated by rare DD variant class, we found that large multigenic deletions had a larger effect than any other type of rare DD variant (OR: 4.7, 95% CI: 1.73-12.95, p = 0.002), followed by LoF variants, and then CNV duplications (**Table 4.3**). We then investigated whether the opposite was also true, i.e., whether those with a bottom decile EA-PGS but a high fluid intelligence score (11-13 out of 13) were less likely to be rare variant carriers, and found that individuals were almost half as likely as others in the same decile to carry a rare DD variant (OR: 0.58, 95% CI: 0.38-0.87, p = 0.009).

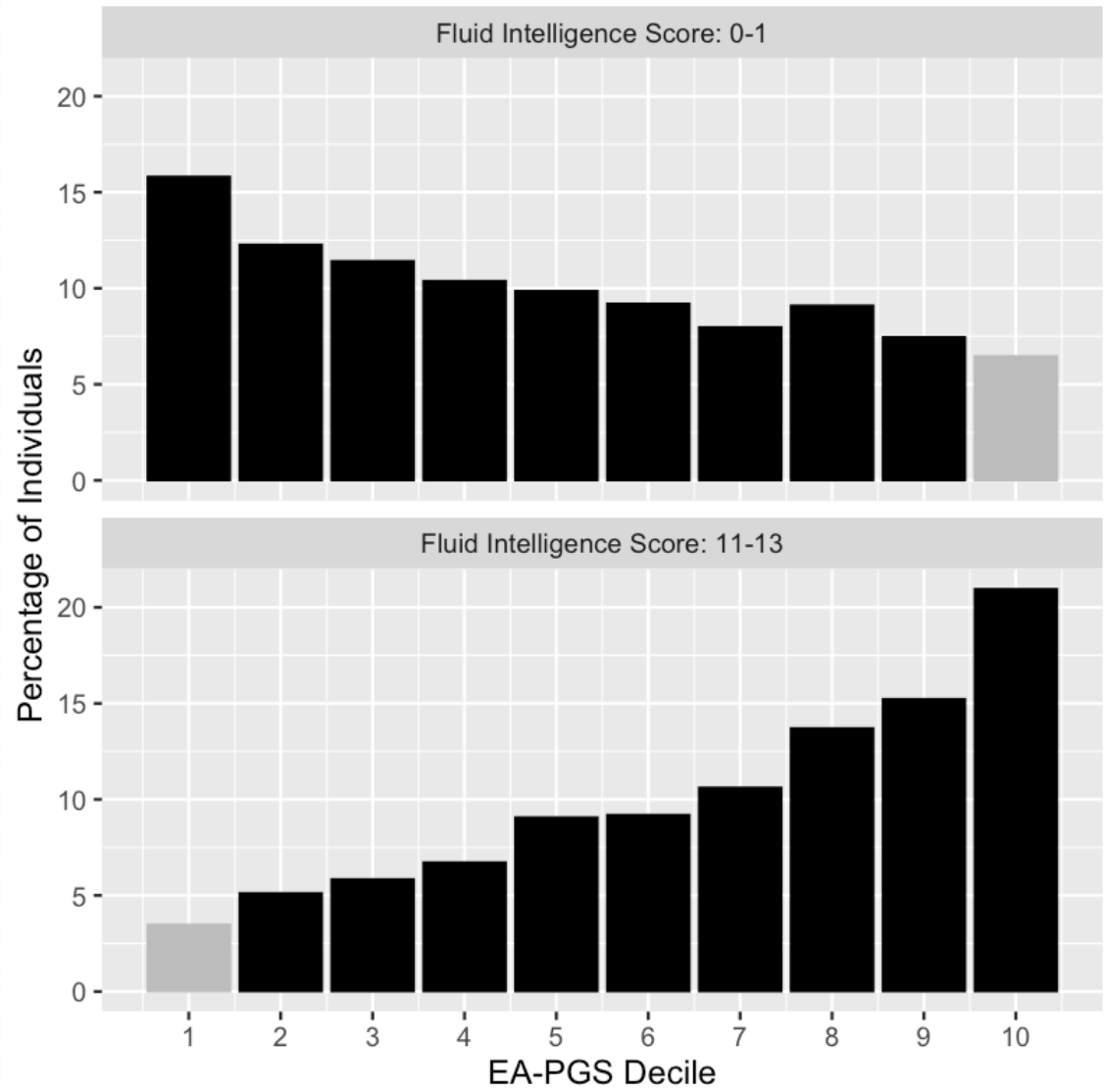


Figure 4.10: Rare DD variant carrier status of phenotypic "deviators" from EA-PGS predictions. Shows the distribution of EA-PGS and fluid intelligence within UK Biobank; phenotypic deviators are highlighted and defined as either a top decile EA-PGS and a low fluid intelligence score (0-1) or a bottom decile EA-PGS and a high fluid intelligence score (11-13).

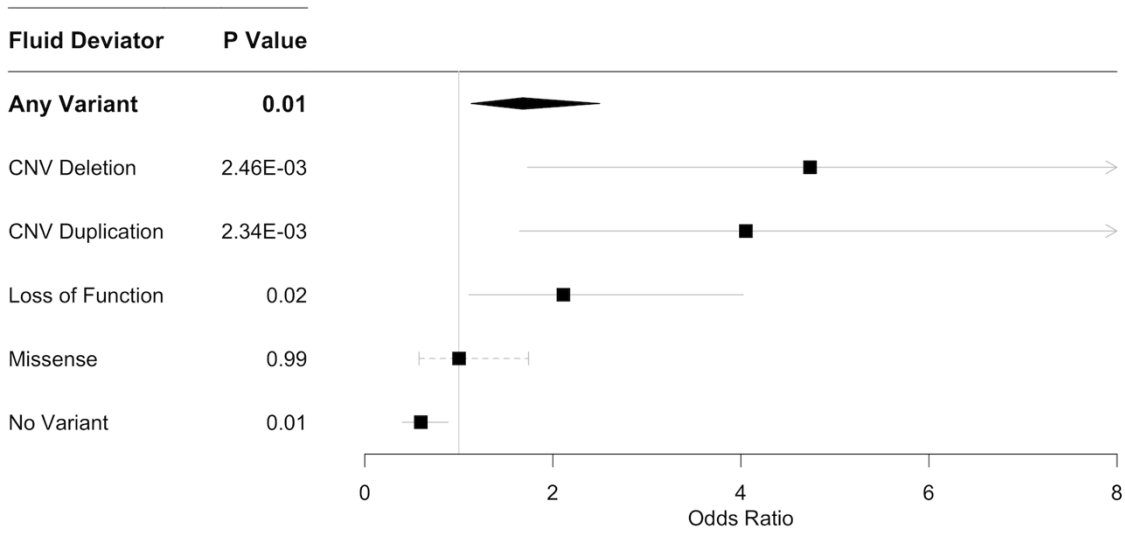


Figure 4.11: Individuals in UKB who have a top decile EA-PGS but scored low on the fluid intelligence test were more likely to be rare DD variant carriers. The comparator group is those within the same EA-PGS decile but with a higher fluid intelligence score (≥2 on the fluid intelligence test).

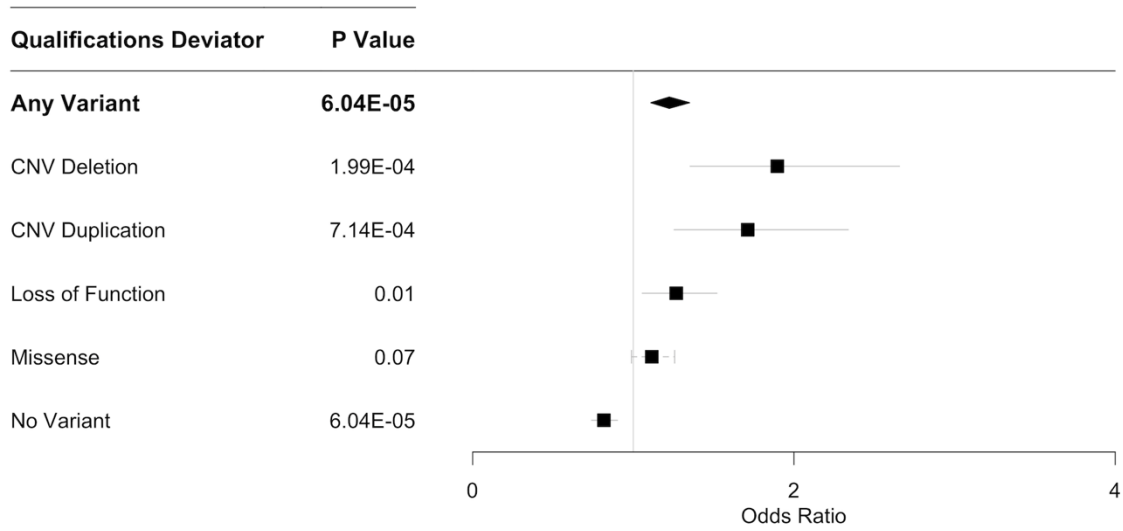


Figure 4.12: Individuals in UKB who reported having no qualifications recorded despite having a top decile EA-PGS were more likely to be rare DD variant carriers. The comparator group is those within the same EA-PGS decile but at least GCSE level qualifications.

Fluid Intelligence Deviators: (n=155)							**Qualifications Deviators: (n=3222)**					
Vs Those in the same EA-PGS Decile (10)							**Vs Those in the same EA-PGS Decile (10)**					
Variant Type	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**		**Variant Type**	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Any variant	1.681	0.340	1.032E-02	1.130	2.500		Any variant	1.224	0.062	6.039E-05	1.109	1.351
CNV or LoF Variant	2.800	0.707	4.560E-05	1.707	4.593		CNV or LoF Variant	1.438	0.107	1.130E-06	1.242	1.664
CNV Deletion	4.735	2.432	2.460E-03	1.731	12.956		CNV Deletion	1.897	0.326	1.994E-04	1.354	2.657
CNV Duplication	4.053	1.864	2.338E-03	1.646	9.981		CNV Duplication	1.713	0.272	7.136E-04	1.254	2.338
LoF	2.112	0.695	2.300E-02	1.109	4.025		LoF	1.267	0.118	1.091E-02	1.056	1.521
Missense	1.002	0.283	9.940E-01	0.576	1.742		Missense	1.115	0.068	7.559E-02	0.989	1.258
No Variant	0.595	0.120	1.032E-02	0.400	0.885		No Variant	0.817	0.041	6.039E-05	0.740	0.902

Vs Decile 1-9							**Vs Decile 1-9**					
Variant Type	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**		**Variant Type**	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Any variant	1.630	0.329	1.539E-02	1.098	2.420		Any variant	1.303	0.070	7.552E-07	1.173	1.447
CNV or LoF Variant	2.771	0.696	4.947E-05	1.694	4.533		CNV or LoF Variant	1.560	0.125	2.813E-08	1.333	1.825
CNV Deletion	4.409	2.238	3.476E-03	1.630	11.924		CNV Deletion	2.139	0.405	5.850E-05	1.476	3.099
CNV Duplication	4.281	1.949	1.406E-03	1.754	10.451		CNV Duplication	1.887	0.326	2.413E-04	1.345	2.649
LoF	2.077	0.680	2.558E-02	1.093	3.947		LoF	1.362	0.135	1.841E-03	1.121	1.655
Missense	0.969	0.272	9.113E-01	0.559	1.682		Missense	1.169	0.076	1.632E-02	1.029	1.327
No Variant	0.613	0.124	1.539E-02	0.413	0.911		No Variant	0.768	0.041	7.552E-07	0.691	0.852

Table 4.3: Rare variant association tests for phenotypic "deviators", where their EA-PGS does not correlate with corresponding phenotypes (fluid intelligence scores or their qualifications), tested against both individuals who fall in the same EA-PGS decile, or compared to the rest of UK Biobank. Odds ratio relates to the likelihood of an individual who is a phenotypic "deviator" being a carrier of a rare variant.

4.34 Clinical diagnoses among carriers

Next, we investigated whether a decrease in EA-PGS correlates with the likelihood of receiving a clinical diagnosis related to DD amongst the rare DD variant carriers we identified in UKB. The number of individuals identified within the three diagnostic categories (child-DD N=7933; adult neuropsychiatric N=19,004; and other mental health issues N=32,911) is likely to be under-estimated due to absence of, or omissions in, individual hospital records available within UKB. Therefore, while individuals in any of these diagnostic categories were more likely to be rare DD variant carriers than the rest of UKB, the majority did not carry a rare variant in any of the DD genes, and many individuals with a rare DD variant did not have a corresponding diagnosis. Despite these limitations, we found that, amongst rare DD variant carriers, those with a related clinical diagnosis across any of our three categories had a substantially lower EA-PGS than those without **(Figure 4.13)**. They also had a larger phenotypic change than other rare variant carriers without a diagnosis; individuals with a rare DD variant and a related clinical diagnosis were more likely to be unable to work (OR: 6.66, 95% CI: 6.07-7.32, p = 4.51E-308), less likely to have a degree (OR: 0.71, 95% CI: 0.66-0.76, p = 3.76E-23), and less likely to be in employment (OR: 0.33, 95% CI: 0.31-0.37, p = 2.07E-143) than those who carry a rare DD variant but do not have a diagnosis recorded in UKB **(Table 4.4)**. This suggests that both the aggregation of overall number of rare DD variants carried and a lower EA-PGS can alter the overall expressivity of the phenotype towards reaching the threshold of clinical disease.

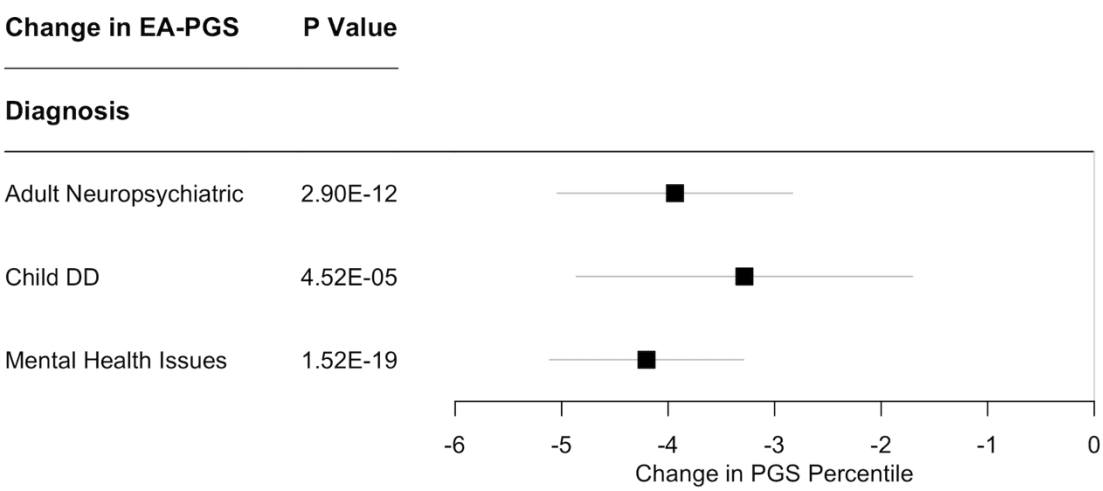


Figure 4.13: Average change in EA-PGS among rare DD variant carrier with a relevant clinical diagnosis. Amongst individuals carrying one or more rare DD variants, those who are clinically diagnosed with either child-dd or adult neuropsychiatric condition or other mental health issues have a substantially lower EA-PGS percentile than those who do not have a related clinical diagnosis recorded in UKB.

Association results between DD clinically diagnosed rare variant carrier (n=3602), and non-clinically diagnosed rare variant carriers (n= 50,843):					
Trait:	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Fluid Intelligence	-0.451	0.053	2.645E-17	-0.556	-0.347
Age Left Education	-0.281	0.041	1.274E-11	-0.362	-0.199
Years in Education	-1.095	0.084	8.852E-39	-1.260	-0.931
Income	-0.518	0.020	3.290E-142	-0.558	-0.478
Townsend Deprivation Index	1.151	0.051	1.590E-113	1.051	1.250
Numeric Memory	-0.184	0.033	3.648E-08	-0.250	-0.119
Reaction Time	0.191	0.016	8.869E-33	0.160	0.223
Time Taken on Pairs Test	0.328	0.056	4.347E-09	0.218	0.437
Height	-0.863	0.109	2.222E-15	-1.076	-0.650

Trait:	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Unable to Work	6.670	0.319	0.000E+00	6.074	7.325
In Employment	0.339	0.014	2.070E-143	0.312	0.368
Has a Degree	0.706	0.025	3.757E-23	0.659	0.756
Never a Parent	1.436	0.057	1.498E-19	1.327	1.553
Never Pregnant	1.226	0.070	4.035E-04	1.095	1.372
Never a Father	1.690	0.095	1.673E-20	1.512	1.888
(Has a mental health related diagnosis)	7.736	0.311	0.000E+00	7.151	8.369

Table 4.4: Association test results showing the difference in phenotype between individuals who carry a rare DDG2P variant and also have either an Adult neuropsychiatric or Child-DD related diagnosis, and those who carry a rare variant but do not have a clinical diagnosis.

4.35 Female protective effect

Previous research has suggested that the prevalence of neurodevelopmental disorder diagnosis is higher among males than females[30,332,416,417], and that transmission from unaffected mothers to affected sons is overrepresented. A "female protective model" could contribute towards this observation, in that females can tolerate larger, more deleterious variants in their genome before they reach the clinical diagnosis threshold[317,334,418]. To test whether this could be seen in a non-clinical population cohort, we investigated whether females with a related neurodevelopmental or neuropsychiatric diagnosis in UKB were more likely to be rare DD variant carriers than males with a similar diagnosis in

UKB. Women and men in UKB were no more likely than each other to carry a rare DD variant overall, and among those with a diagnosis, women were slightly more likely to carry a rare variant than men (7.5% of women had a variant compared to 6% of men, p = 0.10).

We found that among individuals in UKB, females with a related diagnosis were somewhat more likely to be rare variant carriers than males with a related diagnosis when compared to the rest of UKB without a related clinical diagnosis (p = 0.08) (**Figure 4.14**). While this trend followed when compared only to each other, it was not statistically significantly different (**Figure 4.15**). Among those with a diagnosis, women were more likely to be carriers of a higher impact variant, as they were more likely to carry a CNV deletion when compared to the men, which we have previously seen to cause larger phenotypic changes than other deleterious variants[397]. This trend has also been shown previously in clinical studies, where females have been seen to carry a significantly increased large CNV burden compared to males with the same diagnosis[317], and that large structural variations are more likely to be damaging than other types of variation[419], and structural variants that are associated with complex related phenotypes, such as autism or schizophrenia have been shown to affect both regions that are associated with variable phenotypes and loci that are associated with mendelian disease[419]. Within UKB, women were less likely to carry CNV duplications, and there was no difference in the likelihood of carrying a LoF. Male variant carriers showed a bigger change in phenotype when compared to the rest of UKB or the female carriers, with lower fluid intelligence, fewer years in education, and lower income as a group, whereas previous studies suggested that females with a neurodevelopmental related diagnosis were more likely to have an overall more adverse phenotype than males with the same diagnosis[317,336,420]. However, the 95% confidence intervals for males and females overlapped for all traits, despite showing a consistent trend.

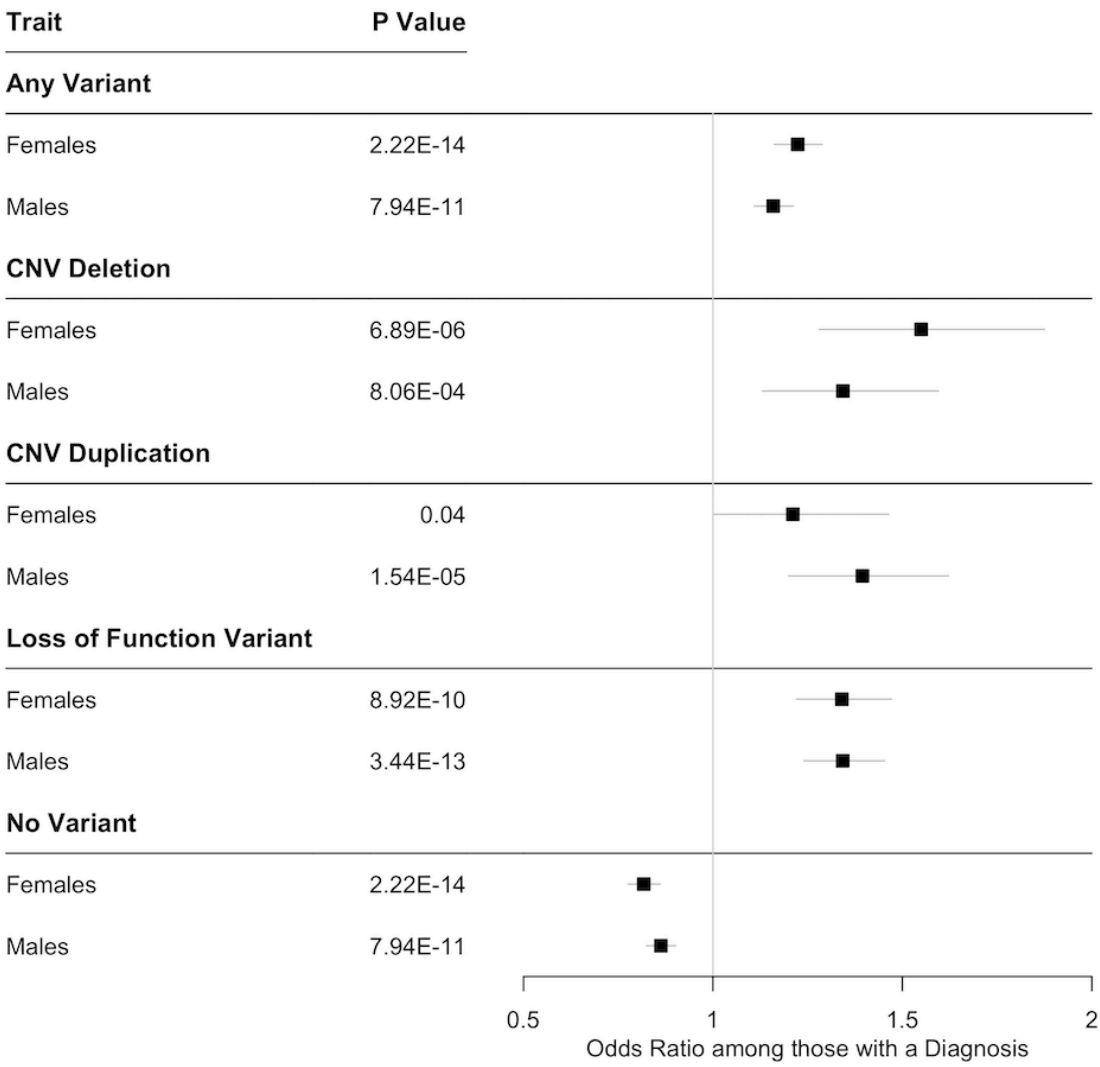


Figure 4.14: The likelihood among individuals who have a diagnosis to be carriers of a specific rare variant in a DD gene compared to the remainder of undiagnosed UKB. Females are slightly more likely to be carriers of CNV deletions, and males are slightly more likely to be carriers of CNV duplications, with no difference in the odds ratio of being a loss of function variant carrier.

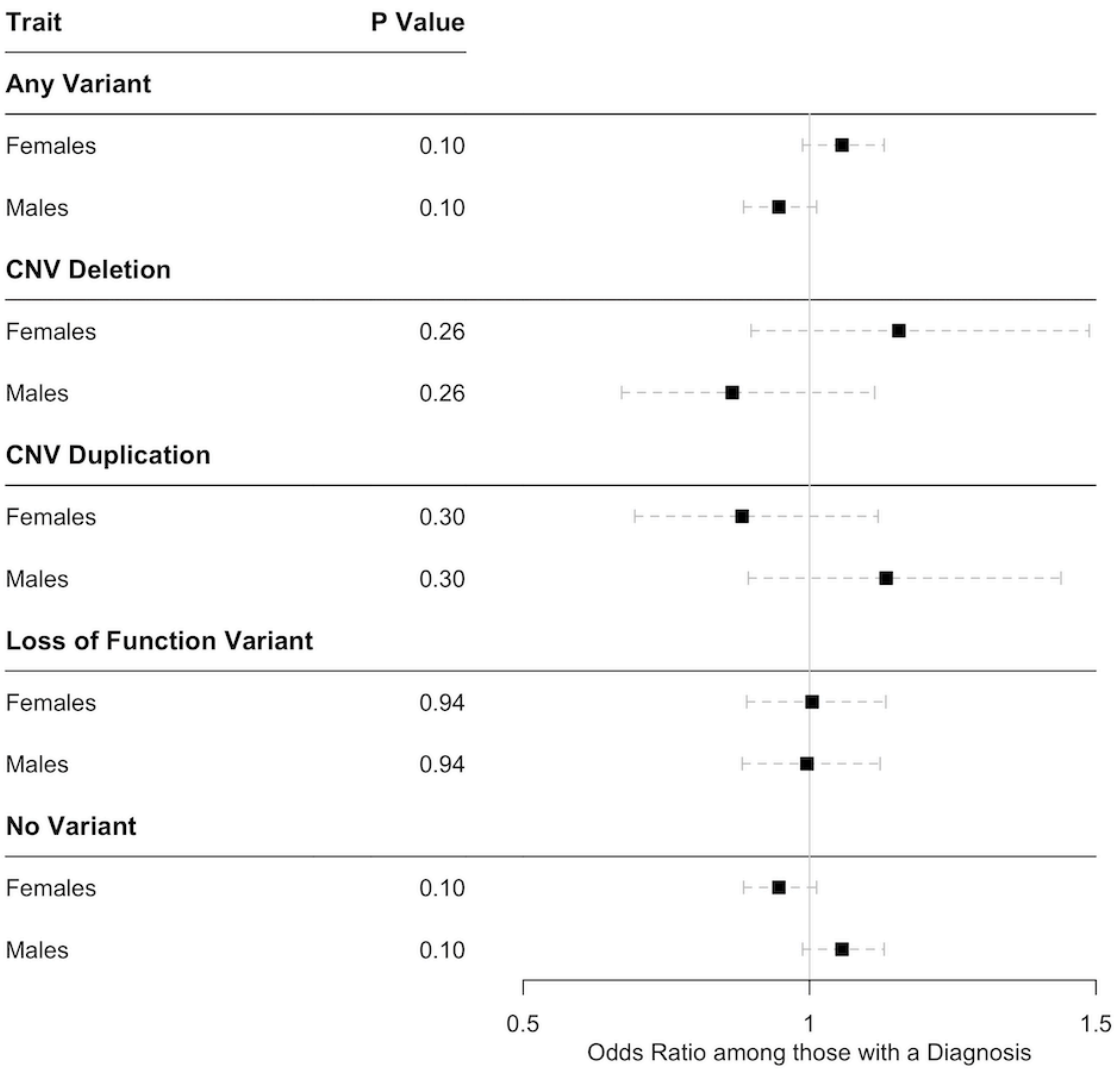


Figure 4.15: The likelihood among individuals who have a diagnosis to be carriers of a specific variant in a DD gene, when only compared to other individuals who also have a diagnosis (i.e., males vs females). When not compared to the remainder of UKB, females stay slightly more likely to be variant carriers than males, but these results do not reach statistical significance.

To see whether there was a similar additive PGS effect within the diagnosed group, we investigated whether the educational attainment PGS results were any different between males and females, and found that overall, women with a diagnosis had a lower EA-PGS percentile, whether they were also rare DD variant carriers or not (**Figure 4.16)**, again, however, the confidence intervals overlapped so the difference is not significant. Conversely, unaffected/undiagnosed women had a slightly higher EA-PGS overall when

compared to their male counterparts (+0.7% EA-PGS, CI: 0.55-0.87, p = 4.77E-18).

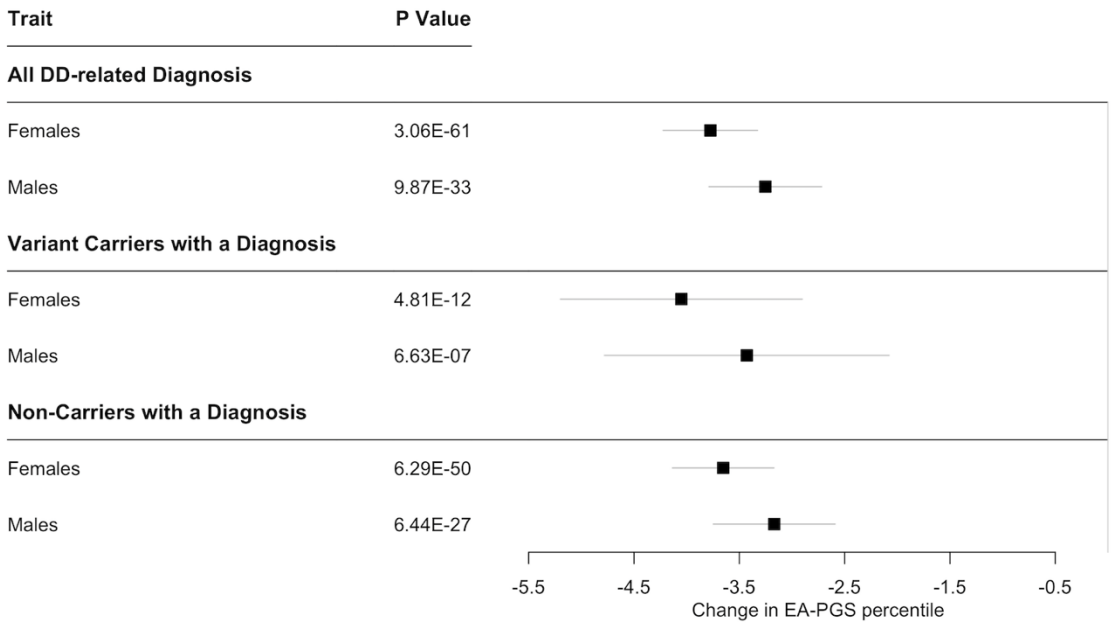


Figure 4.16: Sex differences in the change in EA-PGS among individuals who have a diagnosis, further split into those who also carry a rare DD variant in any of the 599 genes, and those who are not variant carriers. Overall, females generally have a lower EA-PGS percentile if they have a diagnosis than males. Lines show 95% confidence intervals.

4.4 Discussion

We have shown that the phenotypic effect of rare and common genetic variants is additive for a genetically heterogeneous rare disease in a population cohort. The adverse effects of carrying a single deleterious rare variant in genes wherein similar variants cause monogenic DD can be modified by additional rare variants in those genes or by common variants across the genome. Carriers of multiple rare DD variants in UKB have lower fluid intelligence, shorter stature, fewer children, lower income, higher unemployment and higher TDI compared with carriers of single rare DD variants. Additionally, our results suggest that having a higher EA-PGS can partially compensate for the negative cognitive and socio-economic effects of carrying a single or multiple rare DD variants. Moreover, a higher burden of DD-associated variants is more likely to push the phenotypic presentation over the threshold for clinical diagnosis, and correlates with a larger change in phenotype compared to individuals who carry

fewer or no variants. Our results suggest that PGS may provide some clinical utility by improving diagnostic interpretation of rare, likely pathogenic variants that cause monogenic disease.

We have also shown that PGS for similar traits to educational attainment: cognitive ability, mathematical ability, and intelligence, all show similar trends to that of the EA-PGS, both in rare variant carriers and in the remainder of the UKB population cohort, which could be expected, as these traits are all overlapping. However, there were fewer included SNPs used to calculate each of these PGS, and so they had overall less predictive value than the EA-PGS. Conversely, we didn't see the same trend within PGSs that were derived for clinical conditions among our rare variant carriers, and only a slight correlation with our traits and the non-rare variant carrier population for these PGSs.

Investigating the effect of pathogenic rare variants in the general population is important for understanding penetrance and variable expressivity of monogenic diseases. However, there are important limitations on using large-scale genetic data from UK Biobank to investigate rare disease. Firstly, some of the deleterious rare variants we identified may be benign, due to technical artefacts, erroneous pathogenicity predictions, alternative splicing or other mechanisms. Secondly, UKB is known to have an ascertainment bias towards healthier and wealthier individuals compared with the rest of the British population[23], and individuals affected by severe highly penetrant monogenic disorders are likely to be depleted from the cohort. Thirdly, complete medical histories are not available within UKB, which is a relatively old cohort, so many phenotypes of relevance to childhood DDs cannot be evaluated. Fourthly, environment influences were not assessed and may have additional effects on the overall phenotype as well as altering the penetrance and expressivity of genetic variants through gene-environment interactions. Finally, there are challenges in applying common variant PGS across a population, as the underlying summary statistics are heavily dependent upon the populations and ethnicities in which the GWAS were performed. Moreover, PGS often include GWAS results from meta-analyses that incorporate the UKB cohort, which could result in some overfitting. However, our additional sensitivity analysis with SNPs that did not include GWAS information from UKB showed a similar trend to that produced

by the meta-GWAS that did include UKB. Nonetheless, despite these limitations, our results are consistent with previous studies showing the effect of rare DD variants in non-clinical cohorts and the modifying effect of PGS on carriers of rare DD variants[9,387].

In conclusion, we have shown that common and rare genetic variants can additively and independently affect the phenotype of non-clinically ascertained individuals. Our results go some way to explaining the puzzling observation of apparently healthy carriers of monogenic disease-causing variants in the general population, as well as instances of incomplete penetrance and variable expressivity in families affected by rare diseases. Further research is needed to investigate other modifiers, such as rare non-coding variants and gene-environment interactions, and to understand the mechanisms by which genetic modifiers act. Ultimately, incorporating the additive effects of both rare and common variants will improve our understanding of disease.

5. Chapter five: Genetic modifiers of an incompletely penetrant gene: *KDM5B*

5.1 Introduction

5.11 Introduction

We previously investigated how the penetrance or expressivity of a heterogenous group of genetic variants can be modified through additional rare variant burden, and through the accumulation of common genetic variation as polygenic scores. However, there are an additional number of specific ways in which penetrance or expressivity could be modulated. Other potential genetic modifiers include specific non-coding variants that regulate the expression of particular genes, however as the variants that we are investigating are already very rare within large cohort studies, it can be increasingly difficult to identify further variants that affect their expression. Due to their likely rarity, our aim was to identify a gene in which looking for specific genetic modifiers might be possible within a large population cohort such as UKB.

Rare variants in *KDM5B* have been shown in several different cohort studies to be a cause of developmental disorders[423] with incomplete penetrance. *KDM5B* is also a very unconstrained gene. The high prevalence of individuals with plausibly pathogenic *KDM5B* variants in healthy population cohorts, along with its previously demonstrated incomplete penetrance makes it an excellent exemplar gene for studying potential genetic modifiers of penetrance or expressivity, or to test new ways of exploring the effects of specific genetic variants on an individual gene. Here we outline initial work to evaluate the phenotypic effect of regulatory variants or groups of regulatory variants near *KDM5B.*

5.12 *KDM5B*

Loss of function variants in *KDM5B* have previously been shown to be an incompletely penetrant cause of monogenic developmental disorders, in both a recessive and dominant fashion[424–427]. While homozygous and compound heterozygous LoF variants in *KDM5B* cause a recognizable syndrome which

involves developmental delay and facial dysmorphism[427], dominant LoF variants have also been shown to cause developmental disorders[424]. *KDM5B* differs from many other genes in which haploinsufficiency is known to cause developmental disorders, in that it has a pLI of zero, suggesting that LoF variants are well tolerated in this gene[428]. Most genes in which LoF variants are known to cause DD have a pLI closer to 1[428]. pLI is a measure of constraint derived from the gnomAD database, which can be used to identify genes in which protein truncating variants (PTVs) or other deleterious LoF variants are absent or present at a very low frequency in large population samples. Genes with a high pLI, or a pLI of 1 appear 'intolerant to mutation'[429].

KDM5B encodes for a protein that falls into a subcategory of histone lysine methyltransferases, and demethylates H3K4me3, and is involved in transcriptional initiation and elongation[430]. In general, histone lysine enzymes are produced by large group of 51 protein coding genes[431], and they underpin gene regulation[427]. The epigenetic regulation of chromatin by such enzymes plays a critical role in controlling embryonic stem cell self-renewal and pluripotency, and these gene products are strongly expressed during prenatal development[430]. Overall, the function of *KMD5B* and similar genes is important for managing gene expression networks that control self-renewal or differentiation[430]. Histone modifying enzymes are involved in the posttranslational modification of histones and the epigenetic control of gene expression. *De novo* variants in histone modifier genes have been shown to be the cause of a spectrum of different genetic diseases, including congenital heart disease[432] and developmental disorders[427]. There are four paralogs to *KDM5B*, two of which have also been previously linked to developmental disorders: *KDM5A* to ASD[433,434], and *KDM5C* to X-linked intellectual disability and ASD[435,436].

As *KDM5B* has been shown to be a relatively unconstrained gene, we were interested in why this may be. Variants within the gene are equally distributed, there are no PTV hotspots that may be the cause of the low pLI value, similarly, variants are not located all within one transcript, making the variants present in large population cohorts plausibly pathogenic (**Figure 5.1**).

Figure 5.1: gnomAD information showing where LoF variants are located in *KDM5B*

5.13 Regulatory elements as genetic modifiers

While the functional effects of LoF variants within protein coding regions can be more easily investigated, the remaining ~98% of the non-coding genome contains many regulatory regions that are also functionally important[437], and variants within these regions may explain a large fraction of the heritability of some genetic conditions[438]. Variants in non-coding regions can potentially alleviate or exacerbate clinical conditions caused by a primary PTV[3], and therefore can be an important form of genetic modifier. Previous GWAS results have suggested that more than 88% of trait-associated variants identified are in non-coding regions[439], making them a potentially great source of genetic modifiers to examine when looking at rare disease caused by deleterious rare variants, with variants that have an effect on the resulting phenotype being considered putative regulatory variants[440].

Regulatory regions include promoters, enhancers, boundary elements such as UTRs, and transcription factor binding sites, and all are important in understanding how genes are expressed, and how variation can affect the transcription or translation of genes[440]. Variants in these regions could result in up- or down-regulation of gene expression, potentially ameliorating or exacerbating the effect of a heterozygous LoF on the other haplotype[142].

5.14 Identifying potential genetic modifiers

The identification of the functional effects of non-coding variants is a major challenge within human genetics[370]. Understanding how regulatory functions are defined within genomic sequences is difficult and makes characterizing how genomic variation links to phenotypic traits difficult, even among diseases where the significance of specific regulatory variants has already been shown[441,442]. Increasing our understanding of how variation in the human genome can affect phenotype depends on having a comprehensive and detailed knowledge of both the underlying genetic sequence and the phenotype associated with it, both within coding regions and the rest of the genome[27].

While large population-based databases and population cohorts are increasingly important for investigating the penetrance and expressivity of rare genetic variants that have previously only been identified in clinical cohorts, there is still a significant amount of genetic data that has not yet been explored – much of it within non-coding or regulatory regions of the genome. Because the genome is so vast, the use of machine learning to try and interpret different regions based on underlying patterns is a promising idea. Many common variants that are identified by GWAS are located in non-coding regions of the genome[443], but identifying disease-associated rare variants located in promoters, enhancers, and other regulatory elements is more challenging. Although aggregate burden tests have been developed for performing association tests of rare, functionally similar coding variants, it is unclear how to group regulatory variants as they could have opposing functional effects.

Many machine learning models have been produced with the aim of identifying or prioritizing non-coding variants that may affect or cause disease[444–448]. Machine learning models provide an opportunity to assist in the prioritization within the prediction of variants that may result in functional effects[449], predict gene expression levels[450], or identify novel trait-associated variants[451,452]. For example, the sei machine learning model is a convolutional neural network, using 4kb samples taken from across the genome, with chromosomes 8 and 9 left for testing, and chromosome 10 for validation. The developers of Sei attempted to provide a “comprehensive, chromatin-level sequence model”, using genomic sequence features to predict which regions of the genome were

likely to contain functional or regulatory elements. To calculate these predictions, they used data from the Cistrome[453], ENCODE[454], and Roadmap Epigenomics[455] projects. Data from these consortiums was used to train the model on identifying epigenetic features, so that commonalities could be identified and expanded to predictions within the remainder of the genome[448]. Sei can be adapted for predicting whether variants located within non-coding regions are likely to be found within a number of regulatory regions, which are grouped together based on their predicted function and the type of tissue they have been predicted to have an effect within.

In addition to evaluating the effect of LoF variants in *KDM5B* in UKB, and single variant associations near *KDM5B* that might explain instances of incomplete penetrance, we also utilized a previously published machine learning model to try and identify and predict whether specific genetic variants or groups of functionally similar variants near our gene of interest cause a potential change in gene regulation[448]. The model allowed us to group non-coding variants predicted to have the same direction of effect on gene regulation, and hence perform aggregate burden tests of regulatory variants predicted to be functionally similar.

5.2 Materials and Methods

5.21 Identifying variants in of interest in UK Biobank Cohort

The UKB cohort has been described in previous chapters. Whole genome sequence data relating to 200,000 individuals was released in November 2021, with average coverage of 32.5X, by Illumina Novaseq[27]. Using the genomic data from 200,000 UKB participants, and exome data from 450,000 UKB participants, we attempted to use different methods to identify putative genetic modifiers that could affect the penetrance or expressivity of LoF *KDM5B* variants.

We identified anyone who carried a rare (<5 occurrences in UKB) putatively deleterious LoF variant in *KDM5B* in 450k individuals who have WES, and 200k individuals who have WGS. We only included those with variants in the canonical transcript, variants that fell outside of the final exon, and were

predicted high confidence as being LoF by LOFTEE[401]. To investigate the potential effects of non-coding variants found proximal to *KDM5B*, we identified variants located either up to 1Mb upstream or 1Mb downstream of the gene that were present in the genomes of individuals in UKB (**Figure 5.2**).

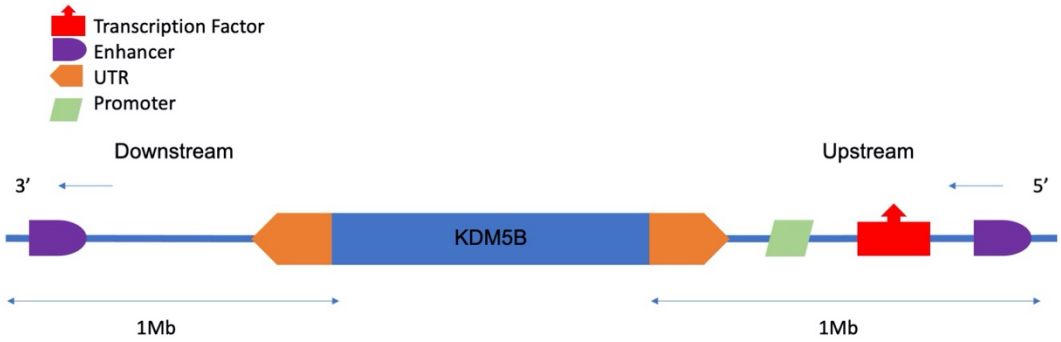


Figure 5.2: Potential regions of interest to identify within 1Mb proximity to *KDM5B*

5.22 Sei Machine Learning Model

We evaluated the Sei machine learning model for identifying rare non-coding variants in the regulatory regions of *KDM5B* (e.g., promoters, enhancer, and up- and down-stream UTRs surrounding the gene) that may be associated with changes in the phenotype in rare variant carriers, either as individual variants than confer an overall effect, or as overall rare variant burden, as both have previously been suggested to affect the clinical presentation of other neurodevelopmental disorders[456]. We used the Sei machine learning model to annotate variants 1Mb either side of *KDM5B* in UKB. Using the predictions from the machine learning model, we identified those predicted to have the biggest effect on gene expression – either negative or positive – which included variants in promoters, enhancers, and transcription factor binding sites.

5.23 Statistical analysis

To evaluate the effect of variants or groups of variants on various neurocognitive phenotypes of interest (see previous chapters), we used logistic regression for binary traits and linear regression for continuous traits, using STATA (Version 16.0). We controlled for age, sex, centre, and 40 principal components. Tested traits were described in Chapter 4.

For the single variant analysis we used REGENIE[457] on 150k individuals in UKB, to test for associations between single variants and fluid intelligence scores.

5.3 Results

5.31 Phenotypic changes in *KDM5B* LoF variant carriers in UKB

We identified 199 individuals in UKB who carried a predicted deleterious LoF variant in *KDM5B,* 76 of whom also have a whole genome sequence (WGS) available. LoF variants were distributed throughout the gene. We performed aggregate gene burden tests on these carriers of *KDM5B* LoF variants across the entire UKB cohort, and those who also had a WGS available for several traits related to the previous cognitive-related phenotypes we identified. Individuals with a *KDM5B* LoF variant showed a decrease in fluid intelligence, fewer years in education, and lower numeric memory scores than the non-carrier group (**Figure 5.3**). They were also more likely to be unable to work, and less likely to have a degree (**Figure 5.4**). Individuals with a *KDM5B* LoF variant were also less likely to have a recorded fluid intelligence score compared to the rest of UKB; among the entire 450k cohort, 40.7% of carriers had taken the fluid intelligence test, compared to 48.2% of the non-carrier cohort (difference: 7.5%, ttest p = 0.01, 95%CI: 0.5-14). Among those who also had WGS, 32/76 *KDM5B* variant carriers had taken the test (42.1%), compared to the 52.4% of non-carriers among the WGS group (difference: 10.3%, p = 0.036, 95%CI: 0.9-21).

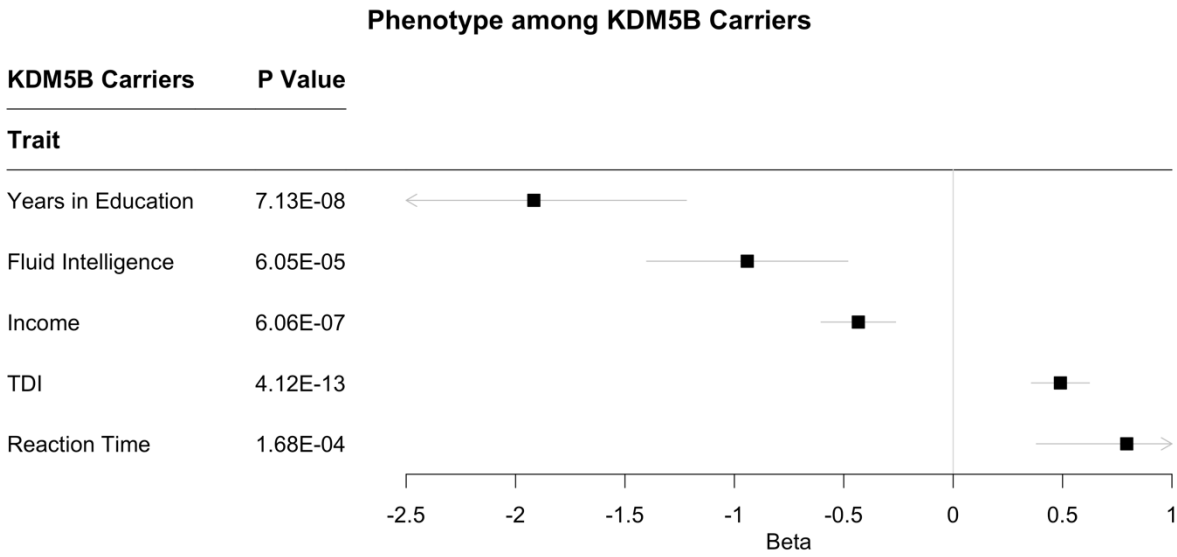


Figure 5.3: The change in phenotype of *KDM5B* LoF variant carriers compared to the remainder of non-carrier UKB for some tested continuous traits. Years in education beta is measured in years, fluid intelligence is measured from 0-13. All tests reached Bonferroni corrected statistically significant levels (0.003), horizontal lines represent 95% confidence intervals.

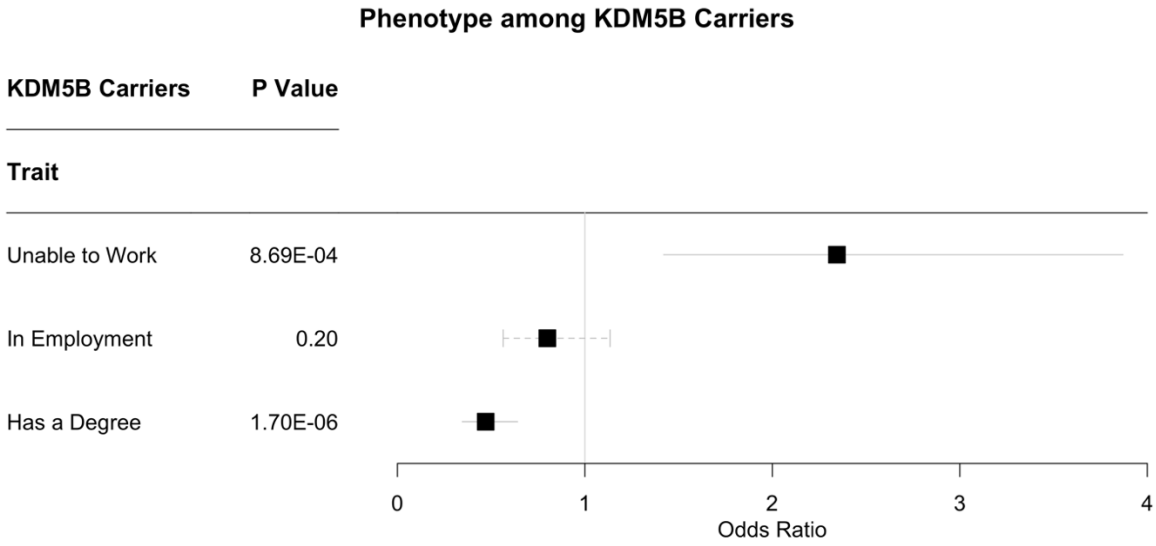


Figure 5.4: The change in phenotype of *KDM5B* LoF variant carriers compared to the remainder of non-carrier UKB for three binary traits. Horizontal lines show 95% confidence intervals, unbroken lines represent statistically significant results (Bonferroni corrected p value is 0.003). Dashed lines represent results that did not reach the threshold for statistical significance.

None of the *KDM5B* LoF variant carriers had an additional coding variant in the gene that could explain the variable penetrance so next we investigated *cis* regulatory variants.

5.32 Non-coding variants in *KDM5B* variant carriers

We first investigated whether any individual common variants within 1Mb of *KDM5B* were associated with fluid intelligence, but found that there were no single variants that were significantly associated with the trait (**Figure 5.5**). Therefore, we chose to use the Sei model to try and identify potential rare variants that could have an effect on fluid intelligence scores, and further group the variants together to potentially increase statistical power and ability to identify such variants.

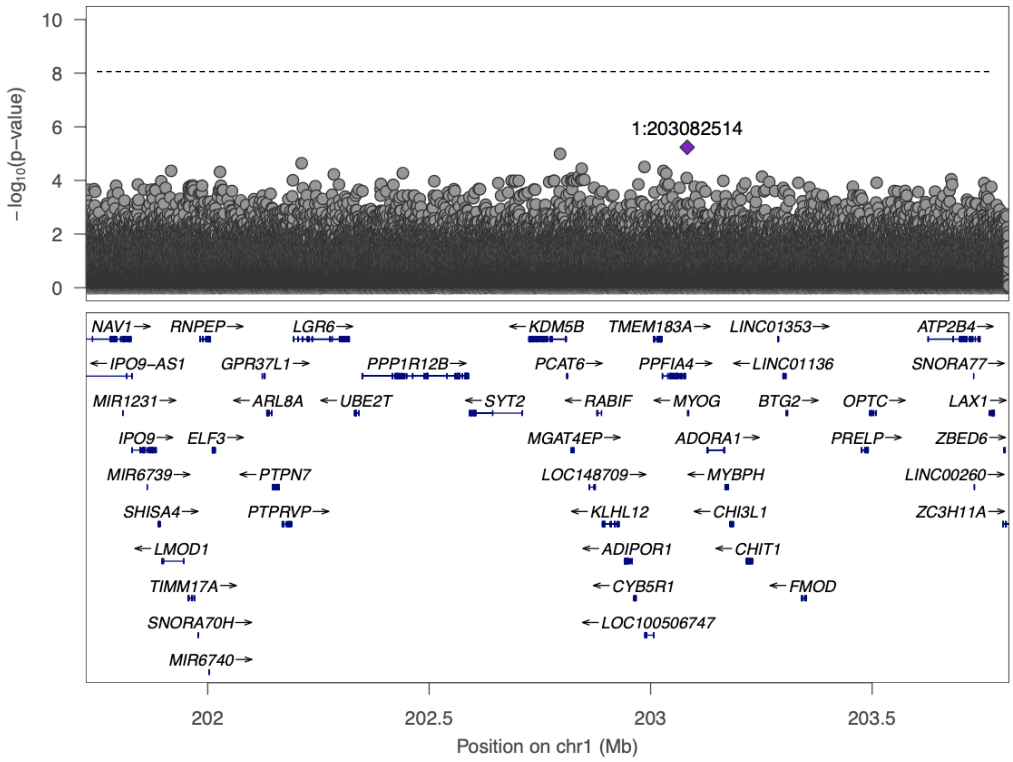


Figure 5.5: Locuszoom plot showing SNP associations with fluid intelligence within a 1Mb region either side of *KDM5B.* Dotted line shows the Bonferroni-corrected p value threshold, of which no individual SNP reached.

Using the Sei model, we identified any variant that was found in at least one individual in UKB and was predicted to have either a >1.1-fold positive or <-1.1-fold negative effect on the predicted genetic functional element. There were 362

variants upstream of *KDM5B* that were predicted to have some functional effect on a regulatory region, and 344 downstream, with a range of minor allele frequencies (**Appendix 7.5.1** and **Appendix 7.5.2**). As there were a large number of individuals who had at least one identified variant, we also took a further subset of individuals who carried a variant that was predicted to have a larger effect on the predicted regulatory region. As nearly every individual who was WGS was predicted to have a non-coding variant that had a positive transcriptional effect on a related predicted regulatory feature (n=165,723), we chose only to examine those that had a variant that had a predicted negative effect (n=13,270).

Among individuals who had WGS data, 76 had an identified *KDM5B* LoF variant, and this subset of the *KDM5B* carrier group showed similar association results to that of the larger group with WES (**Table 5.1**). Among these 76, 13 individuals also carried a non-coding variant that was predicted to have a negative effect on the predicted regulatory region, with 8 having a variant that was upstream of the gene, and 5 having a variant that was downstream of *KDM5B*. Due to having such few carriers, we grouped them into two groups for burden testing – those who had a negative-predicted variant that was upstream, and those who had a negative-predicted variant that was downstream.

Trait	Beta	Standard Error	P Value	95% CI I	95% CI II
Fluid Intelligence	-0.807	0.369	2.880E-02	-1.531	-0.084
Years in Education	-1.847	0.568	1.161E-03	-2.961	-0.732
Income	-0.376	0.139	6.934E-03	-0.649	-0.103
TDI	1.057	0.334	1.554E-03	0.402	1.712

Table 5.1: Association test results for *KDM5B* LoF variant carriers with WGS

We then tested associations for these additional variant carriers and our related traits, both in comparison to the other *KDM5B* LoF variant carriers (**Table 5.2**) and to the remainder of individuals who were WGS in UKB (**Table 5.3**). However, due to the small number of carriers within the *KDM5B* LoF variant carrier group we lacked overall statistical power to identify whether these non-coding variants could be having an overall effect on resulting phenotypes.

Trait	Beta	Standard Error	P Value	95% CI I	95% CI II
Fluid Intelligence					
Negative Downstream Variant (n=5)	2.653	1.551	0.096	-0.496	5.802
Negative Upstream Variant (n=8)	2.800	2.196	0.211	-1.659	7.258
Years in Education					
Negative Downstream Variant	-0.989	1.947	0.612	-4.835	2.858
Negative Upstream Variant	1.310	2.413	0.588	-3.457	6.077
Income					
Negative Downstream Variant	0.512	0.458	0.266	-0.394	1.418
Negative Upstream Variant	-0.291	0.598	0.627	-1.475	0.893
TDI					
Negative Downstream Variant	1.255	1.299	0.336	-1.312	3.822
Negative Upstream Variant	0.356	1.617	0.826	-2.839	3.552

Table 5.2: Association results for *KDM5B* LoF carriers with additional non-coding variants compared to *KDM5B* LoF carriers without additional non-coding variants

Trait	Beta	Standard Error	P Value	95% CI I	95% CI II
Fluid Intelligence					
Negative Downstream Variant (n=5)	0.914	0.934	0.328	-0.917	2.746
Negative Upstream Variant (n=8)	-2.004	1.477	0.175	-4.899	0.892
Years in Education					
Negative Downstream Variant	-1.266	1.740	0.467	-4.677	2.145
Negative Upstream Variant	-1.353	2.201	0.539	-5.668	2.961
Income					
Negative Downstream Variant	-0.045	0.418	0.915	-0.863	0.774
Negative Upstream Variant	-0.583	0.552	0.291	-1.665	0.500
TDI					
Negative Downstream Variant	1.930	1.030	0.061	-0.088	3.948
Negative Upstream Variant	1.593	1.302	0.221	-0.960	4.146

Table 5.3: Association results for *KDM5B* LoF variant carriers with additional non-coding variants compared to the remainder of UKB.

Further to this, we performed linear regression burden tests with the predicted quantitative effects on underlying regulatory regions given by the Sei model for the variant effect among the *KDM5B* LoF variant carriers, but saw the same issue with small numbers and lack of statistical power to identify anything (**Appendix 7.5.3**).

5.33 Non-coding variants in non *KDM5B* LoF variant carriers

Due to the small number of *KDM5B* LoF variant carriers, and even smaller number of additional non-coding variant carriers within this group, we expanded the association tests to include anyone who had either a negative upstream or a downstream predicted non-coding in one of the regulatory regions proximal to *KDM5B.* We hypothesized that variants that had a negative effect on the transcription may potentially have a similar, albeit smaller, effect to LoF variants. In total, 414 individuals carried a non-coding variant with a predicted negative effect on a regulatory region proximal to *KDM5B*. We therefore repeated the association testing between individuals who carried one of these non-coding variants, for the same cognitive related traits as previously (**Table 5.4**). However, we saw no significant results or trend, possibly because some of these variants were predicted to have a very small effect.

Trait	Beta	Standard Error	P Value	95% CI I	95% CI II
Fluid Intelligence					
Negative Downstream Variant (n=112)	-0.004	0.026	0.869	-0.055	0.046
Negative Upstream Variant (n=302)	0.039	0.030	0.203	-0.021	0.098
Years in Education					
Negative Downstream Variant	0.116	0.044	0.009	0.029	0.202
Negative Upstream Variant	0.055	0.052	0.290	-0.047	0.158
Income					
Negative Downstream Variant	-0.001	0.011	0.889	-0.022	0.019
Negative Upstream Variant	0.009	0.013	0.472	-0.016	0.034
TDI					
Negative Downstream Variant	0.012	0.026	0.632	-0.038	0.063
Negative Upstream Variant	-0.013	0.031	0.664	-0.074	0.047

Table 5.4: Non-coding variant association test results for all *KDM5B* proximal predicted negative non-coding variant carriers

We then repeated the analysis with a smaller group of individuals who were predicted to have a higher impact non-coding variant in a regulatory region proximal to *KDM5B* (**Table 5.5**). We defined a higher impact non-coding variant as one that was predicted to have a predicted negative effect that caused a -5-fold or below relative difference. In total, 414 individuals carried one of these predicted variants, and none of them also carried a *KDM5B* LoF variant.

However, we lacked statistical power to show an effect due to the number of individuals carrying one of these variants being small.

Trait	Beta	Standard Error	P Value	95% CI I	95% CI II
Fluid Intelligence					
Negative Downstream Variant	-0.350	0.192	0.068	-0.726	0.025
Negative Upstream Variant	-1.738	1.206	0.150	-4.102	0.626
Years in Education					
Negative Downstream Variant	-0.117	0.321	0.715	-0.747	0.512
Negative Upstream Variant	-2.187	2.010	0.276	-6.126	1.752
Income					
Negative Downstream Variant	0.073	0.077	0.341	-0.078	0.224
Negative Upstream Variant	-0.259	0.451	0.566	-1.143	0.625
TDI					
Negative Downstream Variant	-0.164	0.189	0.387	-0.534	0.207
Negative Upstream Variant	0.127	1.189	0.915	-2.204	2.457

Table 5.5: Rare non-coding variant association test results for predicted high impact non-coding variant carriers

5.4 Discussion

Deleterious LoF variants within *KDM5B* have previously been shown to be an incompletely penetrant cause of developmental disorders, and we have shown that LoF variants in this gene have an impact on related sub-clinical phenotypes among individuals in UKB who carry such variants. Individuals who carried a *KDM5B* LoF variant had a significantly lower fluid intelligence score and fewer years in education on average when compared to the remainder of UKB, despite being a gene in which potentially damaging variants can be incompletely penetrant.

We attempted to use a machine learning model to predict whether proximal non-coding variants could have an overall phenotypic effect on individuals in UKB who carried *KDM5B* LoF variants, however the accurate prediction of the functional effects of non-coding variants is a difficult challenge, especially when regulatory regions for many genes have yet to be identified. This is made more challenging by the size of cohorts that would be needed to identify associations between rare variants and non-coding genetic modifiers and the overall phenotypic effect of these. Limitations of our work therefore include the

relatively small number of individuals who carry LoF variants in *KDM5B,* even though there are significantly more carriers of LoF variants in this gene than the majority of monoallelic DD genes. The low number and lack of haplotype information resulted in limited statistical power to identify whether non-coding variants with predicted functional effects within proximal regulatory regions had any overall phenotypic effect on carriers. Similarly, even when we expanded the associations to everyone in UKB who had a predicted high impact negative variant proximal to *KDM5B*, there were only 414 individuals in UKB, suggesting that even larger cohort sizes would be needed for future identification or classification of such non-coding variant effects.

The Sei machine learning model we used is trained on the identification of patterns within the primary genetic sequence of the human genome, and translates these patterns to predictions of regulatory features, based on previously identified regions that follow similar motifs. While this is a good method of expanding our current knowledge to identify putative regulatory features in other areas of the genome, it does mean that our results are not predictions of whether specific non-coding variants are likely to have a functional effect on specific genes, just that they are predicted to have an overall effect on the predicted regulatory feature. While the underlying primary sequence can give suggestions as to where enhancers could be located, not all motifs are well known, or have previously been described[458]. To be able to classify non-coding variants as potentially deleterious or having an overall negative effect on the transcription of a specific gene, we need a more detailed map of where regulatory regions are found for that gene, and the effect of the variant on both proximal *cis*- and potentially distant *trans*- genes. Currently, the identification of regulatory features using this method cannot identify the genes that a regulatory feature has an effect on, beyond that of linear proximity. This information would give us a much greater ability to identify potentially deleterious non-coding variants that could have an effect on resulting phenotypes, and therefore give us the ability to limit association tests to those with truly damaging variants. Overall, the identification of non-coding variants that lie within regions that can potentially have a damaging effect on specific gene transcription, or on the effect of specific coding variants, is still a complex task, and will need to include the curation of variants and regulatory features.

Similarly, even larger population cohorts will be needed to investigate the modification of already rare genetic variants, a substantial barrier to our understanding has been the lack of statistical power when searching for enrichment of variants across and between different regulatory regions or variant classification.

While we attempted to identify some potential non-coding genetic modifiers proximal to *KDM5B,* there are many other genetic causes that could be further investigated. KDM5B interacts with many other proteins in its functional pathway, including FOXG1B, PAX9, MYC, MYCN, and RB1, and has been linked to several more. It is also part of a five member gene family, and previous research has suggested that upregulation of the paralogs within this group can occur when one member produces non-functional proteins[459]. To expand the search for potential genetic modifiers, identification of LoF or predicted gain of function variants in the genes that code for these proteins could be an interesting way of examining whether such genes could be potential genetic modifiers for the relating phenotypic effect associated with *KDM5B* LoF variants.

Similarly, the expansion of the Sei machine learning model to identifying proximal non-coding variants in all of the 599 monoallelic DD genes that we previously looked at could potentially increase the ability to test for associations between non-coding variants and overall phenotypic expression. This would involve careful curation of the results predicted by the model. A future machine learning model that could make accurate predictions of non-coding variants and the regions in which they are located, from underlying sequence to resulting phenotypic effect would be incredibly useful, especially when it comes to personalized clinical care. The integration of functional annotations of the non-coding regions of the genome could help to identify disease-associated pathways and to help prioritize the identification of disease specific regulatory variants[460]. We still face great challenges to accurately predict, interpret, and evaluate the biological functions of non-coding regulatory variants in gene regulation[461]. This difficulty is further increased by the fact that many variants can overlap potentially identifiable regulatory elements within the genome, but may cause no phenotypic change[462]. Similarly, the use of specific regulatory

regions has previously been shown to vary across different tissues and time periods, in particular, the usage of promoters and enhancers has been shown to change through different developmental periods, making it difficult to identify variants that modify genes of interest[111,463,464].

6. Chapter six: Conclusion

6.1 Summary

Overall, we have shown that a large number of rare putatively deleterious variants in genes in which such variants are known to cause monogenic developmental disorders have an overall sub-clinical phenotypic effect in individuals in the healthy general population, with individual carriers as an aggregate showing lower fluid intelligence scores and spending fewer years in education compared to non-variant carriers. These results suggest that variants in such genes can vary in their penetrance and expressivity, even among individuals with no recorded clinical symptoms of disease.

We have also shown that the adverse effects of carrying a single deleterious variant within our subset of genes can be modified by additional rare variants within one of these genes, or by the aggregation of common variants across the genome. We have further shown that the phenotypic effect of rare and common variants is additive for these genetically heterogenous rare diseases. Furthermore, among individuals who carried a putatively deleterious variant, those with a clinical diagnosis had overall a significantly lower related polygenic score than those without a clinical diagnosis. These results suggest a mechanism in which rare deleterious genetic variants can be present in healthy populations without causing the corresponding clinical phenotype, and that the overall burden of both rare and common genetic variants can modify the expressivity of a phenotype.

Finally, we investigated whether non-coding variants that could potentially be genetic modifiers could be identified through machine learning, and whether association tests could identify potential effects caused by these variants using an incompletely penetrant gene as an example. This initial exploration focused on a single gene, *KDM5B*, but was underpowered to show any effect.

6.2 Future perspectives

6.21 Estimating penetrance in diverse cohorts

Participants in population studies are usually investigated in a research-based environment rather than a clinical context, and despite the rigorous phenotypic collection present in some population studies, individuals involved may have subclinical manifestations of disease phenotypes that were unnoticed at the time of recruitment, or were not recorded in their medical histories[71]. Lack of comprehensive phenotypic data can make using population cohorts to calculate the penetrance of genotypes very difficult, but can at least provide a lower boundary of penetrance, with small clinical studies providing the upper boundary[465]. Variant interpretation guidelines suggest that the penetrance of pathogenic variants in general population cohorts should be taken into account when calculating the overall penetrance of such variants[328]; however even within healthy population cohorts there have been individuals identified with the associated phenotype but who have previously been described as unaffected[81], as well as individuals who display symptoms but are below the clinical threshold for classification. This is further complicated by conditions that are late onset. In addition, genetic studies of human disease currently fail to capture the diversity that exists across the world, with most studies involving individuals of European descent[466]. This issue directly affects penetrance estimates, particularly as GWAS results and PRS may not be transferrable across diverse populations due to differing allele frequencies[467]. Many deleterious variants may not be sufficient alone to cause disease, and therefore estimates of penetrance need to consider the presence of other genetic variants as well as potential environmental effects. Calculating the etiological fraction of rare variants in specific conditions may provide a useful way to evaluate the probability that a variant detected in an individual with disease is causative[22,468], and disease-specific variant classifiers may also be of use[469].

6.22 Screening of unselected populations

As WGS becomes more common, individuals at risk of genetic disease will be identified earlier in life, potentially even from birth[470] and often prior to the appearance of relevant phenotypes. This can have a positive impact on overall health, with individuals who have no family history but a previously unknown

high risk of disease being identified, enabling preventative screening or early treatment interventions. However, as seen across a number of population cohort studies, healthy individuals can harbour many potentially deleterious variants without ever developing any clinical symptoms. The effective use of genomic data requires a comprehensive understanding of functional genotype-phenotype correlations, that goes beyond that of Mendelian inheritance patterns. The increase in sequencing of unselected populations, linked with electronic health records or other longitudinal phenotypic data, gives us unprecedented ability to identify and reclassify rare variants and calculate penetrance estimates for a wide range of diseases and genotypes. These large-scale studies are crucial to inform the development of genomic screening programmes[470,471] and the management of incidental or secondary findings. Discovery estimates of secondary findings vary from 1-3% of the population, with the majority of identified variants being those that confer susceptibility to cancer[472,473]. Incidental findings are predicted to be detectable at an appreciable level in individuals in the general population, many of whom never develop the corresponding disease, suggesting that more robust determinations of pathogenicity are needed, including penetrance estimates for those without a family history of the disease[474].

6.3 Conclusion

Incomplete penetrance and variable expressivity are a significant concern for the correct interpretation of genetic variation and of diagnosing genetic disease. Correctly estimating penetrance and expressivity is challenging, with clinical cohorts and population studies both offering a different insight into its quantification. Although many monogenic disease-causing variants are fully penetrant, many are not and furthering our knowledge will involve WGS of population cohorts of increasing size and diversity, as well as functional studies of individual patients with specific clinical phenotypes. Achieving a mechanistic understanding of how incomplete penetrance and variable expressivity occur will help inform diagnostic and prognostic testing, clinical management, and accurate genetic counselling. To improve diagnostics and clinical interpretation of incompletely penetrant genotypes, a more sophisticated approach to disease genetics may be needed that integrates disease mechanism and specific variants with variation in levels of gene and isoform expression as well as other

genetic and non-genetic modifiers. Improving our knowledge of how variants exert their effects on genes, cellular pathways, and overall phenotypes will improve our understanding of disease and facilitate the development of new therapeutic interventions.

References

1. Haendel, M. *et al.* How many rare diseases are there? *Nat. Rev. Drug Discov.* **19**, 77–78 (2020).
2. Boycott, K. M. *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).
3. Rahit, K. M. T. H. & Tarailo-Graovac, M. Genetic Modifiers and Rare Mendelian Disease. *Genes* **11**, 239 (2020).
4. OMIM Gene Map Statistics. https://www.omim.org/statistics/geneMap.
5. Human Gene Mutation Database (HGMD) Professional | QIAGEN. https://digitalinsights.qiagen.com/products-overview/clinical-insights-portfolio/human-gene-mutation-database/?cmpid=QDI_GA_QCI&gclid=Cj0KCQjw3IqSBhCoARIsAMBkTb3ww3dMzY0IBhPPmdiIMhjE1FcmjHL7zkrkBABi6i.
6. Stephanou, C. *et al.* Genetic Modifiers at the Crossroads of Personalised Medicine for Haemoglobinopathies. *J. Clin. Med.* **8**, 1927 (2019).
7. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
8. Schacherer, J. Beyond the simplicity of Mendelian inheritance. *C. R. Biol.* **339**, 284–288 (2016).
9. Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
10. Crawford, H. *et al.* Genetic modifiers in rare disorders: the case of fragile X syndrome. *Eur. J. Hum. Genet.* **29**, 173–183 (2021).

11. Kumar, K. R. *et al.* Whole genome sequencing for the genetic diagnosis of heterogenous dystonia phenotypes. *Parkinsonism Relat. Disord.* **69**, 111–118 (2019).
12. McDermott, J. H., Study, D. D. D. & Clayton-Smith, J. Sibling recurrence of total anomalous pulmonary venous drainage. *Eur. J. Med. Genet.* **60**, 265–267 (2017).
13. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
14. Muse, E. D., Chen, S.-F. & Torkamani, A. Monogenic and Polygenic Models of Coronary Artery Disease. *Curr. Cardiol. Rep.* **23**, 107 (2021).
15. Vuckovic, D. *et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases*. http://medrxiv.org/lookup/doi/10.1101/2020.02.02.20020065 (2020) doi:10.1101/2020.02.02.20020065.
16. Goodrich, J. K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat. Commun.* **12**, 3505 (2021).
17. Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).
18. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Ng, P. C. *et al.* Genetic Variation in an Individual Human Exome. *PLOS Genet.* **4**, e1000160 (2008).
20. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

21. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* (2021) doi:10.1002/humu.24309.
22. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203 (2017).
23. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
24. Ye, J. Z., Delmar, M., Lundby, A. & Olesen, M. S. Reevaluation of genetic variants previously associated with arrhythmogenic right ventricular cardiomyopathy integrating population-based cohorts and proteomics data. *Clin. Genet.* **96**, 506–514 (2019).
25. Cox, N. UK Biobank shares the promise of big data. *Nature* **562**, 194–195 (2018).
26. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
27. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
28. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
29. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
30. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

31. Sheridan, E. *et al.* Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *The Lancet* **382**, 1350–1359 (2013).
32. Ropers, H. H. Genetics of Early Onset Cognitive Impairment. *Annu. Rev. Genomics Hum. Genet.* **11**, 161–187 (2010).
33. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
34. Brunet, T. *et al.* De novo variants in neurodevelopmental disorders—experiences from a tertiary care center. *Clin. Genet.* **100**, 14–28 (2021).
35. Moreno-De-Luca, A. *et al.* Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. *Lancet Neurol.* **12**, 406–414 (2013).
36. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet Lond. Engl.* **385**, 1305–1314 (2015).
37. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
38. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
39. Kvarnung, M. & Nordgren, A. Intellectual Disability & Rare Disorders: A Diagnostic Challenge. in *Rare Diseases Epidemiology: Update and Overview* (eds. Posada de la Paz, M., Taruscio, D. & Groft, S. C.) 39–54 (Springer International Publishing, 2017). doi:10.1007/978-3-319-67144-4_3.

40. Niturad, C. E. *et al.* Rare GABRA3 variants are associated with epileptic seizures, encephalopathy and dysmorphic features. *Brain J. Neurol.* **140**, 2879–2894 (2017).
41. Stokes, B. *et al.* SIX3 deletions and incomplete penetrance in families affected by holoprosencephaly. *Congenit. Anom.* **58**, 29–32 (2018).
42. Edwards, S. D. *et al.* Clinical characterization of individuals with the distal 1q21.1 microdeletion. *Am. J. Med. Genet. A.* **185**, 1388–1398 (2021).
43. Minikel, E. V. *et al.* Quantifying penetrance in a dominant disease gene using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
44. Mahat, U., Ambani, N. M., Rotz, S. J. & Radhakrishnan, K. Heterozygous CTLA4 splice site mutation c.458-1G > C presenting with immunodeficiency and variable degree of immune dysregulation in three generation kindred of Caribbean descent. *Pediatr. Hematol. Oncol.* **38**, 658–662 (2021).
45. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
46. Shawky, R. M. Reduced penetrance in human inherited disease. *Egypt. J. Med. Hum. Genet.* **15**, 103–111 (2014).
47. Ittisoponpisan, S., Alhuzimi, E., Sternberg, M. J. E. & David, A. Landscape of Pleiotropic Proteins Causing Human Disease: Structural and System Biology Insights. *Hum. Mutat.* **38**, 289–296 (2017).
48. Gruber, C. & Bogunovic, D. Incomplete penetrance in primary immunodeficiency: a skeleton in the closet. *Hum. Genet.* **139**, 745–757 (2020).

49. Ibrahim, D. M. *et al.* A homozygous HOXD13 missense mutation causes a severe form of synpolydactyly with metacarpal to carpal transformation. *Am. J. Med. Genet. A.* **170**, 615–621 (2016).

50. Johnston, J. J. *et al.* Individualized Iterative Phenotyping for Genome-wide Analysis of Loss-of-Function Mutations. *Am. J. Hum. Genet.* **96**, 913–925 (2015).

51. Zhang, M. *et al.* Brachydactyly type A3 is caused by a novel 13 bp HOXD13 frameshift deletion in a Chinese family. *Am. J. Med. Genet. A.* **182**, 2432–2436 (2020).

52. Li, Q. *et al.* A novel KCNQ4 gene variant (c.857A>G; p.Tyr286Cys) in an extended family with non-syndromic deafness 2A. *Mol. Med. Rep.* **23**, 420 (2021).

53. Raymond, D., Saunders-Pullman, R. & Ozelius, L. SGCE Myoclonus-Dystonia. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

54. Gerrits, M. C. F., Foncke, E. M. J., Koelman, J. H. T. M. & Tijssen, M. a. J. Pediatric writer's cramp in myoclonus-dystonia: maternal imprinting hides positive family history. *Eur. J. Paediatr. Neurol. EJPN Off. J. Eur. Paediatr. Neurol. Soc.* **13**, 178–180 (2009).

55. Smith, F. J. *et al.* Pachyonychia Congenita. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

56. Li, Y. *et al.* A KRT6A mutation p.Ile462Asn in a Chinese family with pachyonychia congenita, and identification of maternal mosaicism: a case report. *BMC Med. Genomics* **14**, 259 (2021).

57. Schmidt, L. S. & Linehan, W. M. FLCN: The causative gene for Birt-Hogg-Dubé syndrome. *Gene* **640**, 28–42 (2018).

58. Nathan, N., Berdah, L., Delestrain, C., Sileo, C. & Clement, A. Interstitial lung diseases in children. *Presse Medicale Paris Fr. 1983* **49**, 103909 (2020).

59. Somaschini, M., Cavazza, A., Riva, S., Sassi, I. & Carrera, P. [Genetic basis in chronic interstitial familial pneumopathy. Familial study of SFTPC]. *Pediatr. Medica E Chir. Med. Surg. Pediatr.* **27**, 103–107 (2005).

60. Aubart, M. *et al.* Association of modifiers and other genetic factors explain Marfan syndrome clinical variability. *Eur. J. Hum. Genet.* **26**, 1759–1772 (2018).

61. Díaz de Bustamante, A., Ruiz-Casares, E., Darnaude, M. T., Perucho, T. & Martínez-Quesada, G. Phenotypic Variability in Marfan Syndrome in a Family With a Novel Nonsense FBN1 Gene Mutation. *Rev. Esp. Cardiol. Engl. Ed.* **65**, 380–381 (2012).

62. Dietz, H. FBN1-Related Marfan Syndrome. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

63. Kraemer, K. H. & DiGiovanna, J. J. Xeroderma Pigmentosum. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

64. Akiyama, M. FLG mutations in ichthyosis vulgaris and atopic eczema: spectrum of mutations and population genetics. *Br. J. Dermatol.* **162**, 472–477 (2010).

65. Kammenga, J. E. The background puzzle: how identical mutations in the same gene lead to different disease symptoms. *FEBS J.* **284**, 3362–3373 (2017).

66. Neeve, V. C. M. *et al.* What is influencing the phenotype of the common homozygous polymerase-γ mutation p.Ala467Thr? *Brain J. Neurol.* **135**, 3614–3626 (2012).

67. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).

68. Shieh, J. T. C. Expanding Genomic Sequencing and Incomplete Penetrance. *Pediatrics* **143**, S22–S26 (2019).

69. Maroilley, T. & Tarailo-Graovac, M. Uncovering Missing Heritability in Rare Diseases. *Genes* **10**, 275 (2019).

70. Xue, Y. *et al.* Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).

71. van Rooij, J. *et al.* Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time. *Genet. Med.* (2020) doi:10.1038/s41436-020-0900-8.

72. Lacaze, P. *et al.* Medically actionable pathogenic variants in a population of 13,131 healthy elderly individuals. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 1883–1886 (2020).

73. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

74. Downs, B. *et al.* Common genetic variants contribute to incomplete penetrance: evidence from cancer-free BRCA1 mutation carriers. *Eur. J. Cancer* **107**, 68–78 (2019).

75. Flannick, J. *et al.* Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
76. Mirshahi, U. L. *et al.* Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. *Am. J. Hum. Genet.* **109**, 2018–2028 (2022).
77. NHS England » NHS Genomic Medicine Service. https://www.england.nhs.uk/genomics/nhs-genomic-med-service/.
78. Stoeklé, H.-C., Mamzer-Bruneel, M.-F., Vogt, G. & Hervé, C. 23andMe: a new two-sided data-banking market model. *BMC Med. Ethics* **17**, 19 (2016).
79. Tarailo-Graovac, M., Zhu, J. Y. A., Matthews, A., van Karnebeek, C. D. M. & Wasserman, W. W. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med.* **19**, 1300–1308 (2017).
80. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
81. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016).
82. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
83. Speletas, M. *et al.* Hereditary angioedema: Molecular and clinical differences among European populations. *J. Allergy Clin. Immunol.* **135**, 570-573.e10 (2015).

84. Austin, E. D. *et al.* Truncating and missense BMPR2 mutations differentially affect the severity of heritable pulmonary arterial hypertension. *Respir. Res.* **10**, 87 (2009).

85. Lu, Y.-Y. & Krebber, H. Nuclear mRNA Quality Control and Cytoplasmic NMD Are Linked by the Guard Proteins Gbp2 and Hrb1. *Int. J. Mol. Sci.* **22**, 11275 (2021).

86. Haeri, M. & Knox, B. E. Endoplasmic Reticulum Stress and Unfolded Protein Response Pathways: Potential for Treating Age-related Retinal Degeneration. *J. Ophthalmic Vis. Res.* **7**, 45–59 (2012).

87. Nguyen, L. S., Wilkinson, M. F. & Gecz, J. Nonsense-mediated mRNA decay: Inter-individual variability and human disease. *Neurosci. Biobehav. Rev.* **46**, 175–186 (2014).

88. Kennedy, J. *et al.* KAT6A Syndrome: genotype–phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet. Med.* **21**, 850–860 (2019).

89. van Leeuwen, J., Pons, C., Boone, C. & Andrews, B. J. Mechanisms of suppression: The wiring of genetic resilience. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **39**, 10.1002/bies.201700042 (2017).

90. Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103**, 171–187 (2018).

91. Lareau, L. F. & Brenner, S. E. Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. *Mol. Biol. Evol.* **32**, 1072–1079 (2015).

92. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* **23**, 198–199 (1998).

93. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).

94. Miller, J. N. & Pearce, D. A. Nonsense-mediated decay in genetic disease: Friend or foe? *Mutat. Res. Mutat. Res.* **762**, 52–64 (2014).

95. Inoue, K. *et al.* Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nat. Genet.* **36**, 361–369 (2004).

96. Zhang, L. X. *et al.* Further Delineation of the Clinical Spectrum of KAT6B Disorders and Allelic Series of Pathogenic Variants. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 1338–1347 (2020).

97. Taniguchi, Y. *et al.* Impact of pathogenic FBN1 variant types on the development of severe scoliosis in patients with Marfan syndrome. *J. Med. Genet.* (2021) doi:10.1136/jmedgenet-2021-108186.

98. Pipis, M. *et al.* Charcot-Marie-Tooth disease type 2CC due to NEFH variants causes a progressive, non-length-dependent, motor-predominant phenotype. *J. Neurol. Neurosurg. Psychiatry* **93**, 48–56 (2022).

99. Moorsel, C. H. M. van, Vis, J. J. van der & Grutters, J. C. Genetic disorders of the surfactant system: focus on adult disease. *Eur. Respir. Rev.* **30**, (2021).

100. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* **38**, 110207 (2022).

101. Niday, Z. & Tzingounis, A. V. Potassium Channel Gain of Function in Epilepsy: An Unresolved Paradox. *The Neuroscientist* **24**, 368–380 (2018).

102. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* **425**, 3919–3936 (2013).
103. Li, X.-H. & Babu, M. M. Human Diseases from Gain-of-Function Mutations in Disordered Protein Regions. *Cell* **175**, 40–42 (2018).
104. Syrbe, S. *et al.* De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat. Genet.* **47**, 393–399 (2015).
105. Platzer, K. *et al.* GRIN2B encephalopathy: novel findings on phenotype, variant clustering, functional consequences and treatment aspects. *J. Med. Genet.* **54**, 460–470 (2017).
106. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).
107. Liu, X.-R. *et al.* GRIN2A Variants Associated With Idiopathic Generalized Epilepsies. *Front. Mol. Neurosci.* **14**, 720984 (2021).
108. Strehlow, V. *et al.* GRIN2A-related disorders: genotype and functional consequence predict phenotype. *Brain* **142**, 80–92 (2019).
109. Masnada, S. *et al.* Clinical spectrum and genotype-phenotype associations of KCNA2-related encephalopathies. *Brain J. Neurol.* **140**, 2337–2354 (2017).
110. Aksentijevich, I. & Schnappauf, O. Molecular mechanisms of phenotypic variability in monogenic autoinflammatory diseases. *Nat. Rev. Rheumatol.* **17**, 405–425 (2021).
111. Ellingford, J. M. *et al.* Recommendations for clinical interpretation of variants found in non-coding regions of the genome. 2021.12.28.21267792 Preprint at https://doi.org/10.1101/2021.12.28.21267792 (2021).

112. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).
113. Wright, C. F. *et al.* Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *Am. J. Hum. Genet.* **108**, 1083–1094 (2021).
114. Bozhilov, Y. K. *et al.* A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. *Nat. Commun.* **12**, 3806 (2021).
115. Paulson, H. Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
116. Hagerman, R. J. *et al.* Fragile X syndrome. *Nat. Rev. Dis. Primer* **3**, 17065 (2017).
117. Payán-Gómez, C., Ramirez-Cheyne, J. & Saldarriaga, W. Variable Expressivity in Fragile X Syndrome: Towards the Identification of Molecular Characteristics That Modify the Phenotype. *Appl. Clin. Genet.* **14**, 305–312 (2021).
118. Cabal-Herrera, A. M., Tassanakijpanich, N., Salcedo-Arellano, M. J. & Hagerman, R. J. Fragile X-Associated Tremor/Ataxia Syndrome (FXTAS): Pathophysiology and Clinical Implications. *Int. J. Mol. Sci.* **21**, E4391 (2020).
119. Fink, D. A. *et al.* Fragile X Associated Primary Ovarian Insufficiency (FXPOI): Case Report and Literature Review. *Front. Genet.* **9**, (2018).
120. Morales, F. *et al.* A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair* **40**, 57–66 (2016).

121. CCG•CGG interruptions in high-penetrance SCA8 families increase RAN translation and protein toxicity. *EMBO Mol. Med.* **13**, e14095 (2021).
122. Laskaratos, A., Breza, M., Karadima, G. & Koutsis, G. Wide range of reduced penetrance alleles in spinal and bulbar muscular atrophy: a model-based approach. *J. Med. Genet.* **58**, 385–391 (2021).
123. Kay, C. *et al.* Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
124. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron* **72**, 245–256 (2011).
125. Kim, E., Napierala, M. & Dent, S. Y. R. Hyperexpansion of GAA repeats affects post-initiation steps of FXN transcription in Friedreich's ataxia. *Nucleic Acids Res.* **39**, 8366–8377 (2011).
126. Holmans, P. A., Massey, T. H. & Jones, L. Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders. *Hum. Mol. Genet.* **26**, R83–R90 (2017).
127. Arning, L. The search for modifier genes in Huntington disease - Multifactorial aspects of a monogenic disorder. *Mol. Cell. Probes* **30**, 404–409 (2016).
128. Massey, T. H. & Jones, L. The central role of DNA damage and repair in CAG repeat diseases. *Dis. Model. Mech.* **11**, dmm031930 (2018).
129. Goula, A.-V. & Merienne, K. Abnormal Base Excision Repair at Trinucleotide Repeats Associated with Diseases: A Tissue-Selective Mechanism. *Genes* **4**, 375–387 (2013).

130. Flower, M. *et al.* MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain* **142**, 1876–1886 (2019).
131. Bahlo, M. *et al.* Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Research* **7**, F1000 Faculty Rev-736 (2018).
132. Ahluwalia, J. K., Hariharan, M., Bargaje, R., Pillai, B. & Brahmachari, V. Incomplete penetrance and variable expressivity: is there a microRNA connection? *BioEssays* **31**, 981–992 (2009).
133. Jordan, W., Rieder, L. E. & Larschan, E. Diverse genome topologies characterize dosage compensation across species. *Trends Genet. TIG* **35**, 308–315 (2019).
134. Hesson, L. B. *et al.* Lynch syndrome associated with two MLH1 promoter variants and allelic imbalance of MLH1 expression. *Hum. Mutat.* **36**, 622–630 (2015).
135. Glazier, A. A., Thompson, A. & Day, S. M. Allelic imbalance and haploinsufficiency in MYBPC3-linked hypertrophic cardiomyopathy. *Pflugers Arch.* **471**, 781–793 (2019).
136. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
137. Pinter, S. F. *et al.* Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome. *Genetics* **200**, 537–549 (2015).
138. Servetti, M. *et al.* Neurodevelopmental Disorders in Patients With Complex Phenotypes and Potential Complex Genetic Basis Involving Non-Coding Genes, and Double CNVs. *Front. Genet.* **12**, 732002 (2021).
139. Carelle-Calmels, N. *et al.* Genetic Compensation in a Human Genomic Disorder. *N. Engl. J. Med.* **360**, 1211–1216 (2009).

140. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol.* **3**, REVIEWS0004 (2002).
141. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
142. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
143. Chen, R. *et al.* FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biol.* **9**, R170 (2008).
144. Green, D. J., Sallah, S. R., Ellingford, J. M., Lovell, S. C. & Sergouniotis, P. I. Variability in Gene Expression is Associated with Incomplete Penetrance in Inherited Eye Disorders. *Genes* **11**, 179 (2020).
145. Aubart, M. *et al.* The clinical presentation of Marfan syndrome is modulated by expression of wild-type FBN1 allele. *Hum. Mol. Genet.* **24**, 2764–2770 (2015).
146. de, M. A. *et al.* Phenotypic Expression and Outcomes in Individuals With Rare Genetic Variants of Hypertrophic Cardiomyopathy. *J. Am. Coll. Cardiol.* **78**, 1097–1110 (2021).
147. Marian, A. J. & Braunwald, E. Hypertrophic Cardiomyopathy: Genetics, Pathogenesis, Clinical Manifestations, Diagnosis, and Therapy. *Circ. Res.* **121**, 749–770 (2017).
148. Falkenberg, K. D. *et al.* Allelic Expression Imbalance Promoting a Mutant PEX6 Allele Causes Zellweger Spectrum Disorder. *Am. J. Hum. Genet.* **101**, 965–976 (2017).

149. Binder, B. J., Landman, K. A., Newgreen, D. F. & Ross, J. V. Incomplete penetrance: The role of stochasticity in developmental cell colonization. *J. Theor. Biol.* **380**, 309–314 (2015).
150. Eckersley-Maslin, M. A. & Spector, D. L. Random Monoallelic Expression: Regulating gene expression one allele at a time. *Trends Genet. TIG* **30**, 237–244 (2014).
151. Gui, B., Slone, J. & Huang, T. Perspective: Is Random Monoallelic Expression a Contributor to Phenotypic Variability of Autosomal Dominant Disorders? *Front. Genet.* **8**, (2017).
152. Vu, V. *et al.* Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell* **162**, 391–402 (2015).
153. Baranzini, S. E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
154. Akinrinade, O. *et al.* Relevance of Titin Missense and Non-Frameshifting Insertions/Deletions Variants in Dilated Cardiomyopathy. *Sci. Rep.* **9**, 4093 (2019).
155. Begay, R. L. *et al.* Role of Titin Missense Variants in Dilated Cardiomyopathy. *J. Am. Heart Assoc.* **4**, e002645 (2015).
156. Kharbanda, M. *et al.* Partial deletion of TCF4 in three generation family with non-syndromic intellectual disability, without features of Pitt-Hopkins syndrome. *Eur. J. Med. Genet.* **59**, 310–314 (2016).
157. Sirp, A. *et al.* Functional consequences of TCF4 missense substitutions associated with Pitt-Hopkins syndrome, mild intellectual disability, and schizophrenia. *J. Biol. Chem.* **297**, 101381 (2021).

158. Dick, I. E., Joshi-Mukherjee, R., Yang, W. & Yue, D. T. Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca(2+)-dependent inactivation. *Nat. Commun.* **7**, 10370 (2016).

159. Scacheri, C. A. & Scacheri, P. C. Mutations in the non-coding genome. *Curr. Opin. Pediatr.* **27**, 659–664 (2015).

160. van der Lee, R., Correard, S. & Wasserman, W. W. Deregulated Regulators: Disease-Causing cis Variants in Transcription Factor Genes. *Trends Genet.* **36**, 523–539 (2020).

161. Silva, J., Fernandes, R. & Romão, L. Translational Regulation by Upstream Open Reading Frames and Human Diseases. *Adv. Exp. Med. Biol.* **1157**, 99–116 (2019).

162. Young, S. K., Palam, L. R., Wu, C., Sachs, M. S. & Wek, R. C. Ribosome Elongation Stall Directs Gene-specific Translation in the Integrated Stress Response. *J. Biol. Chem.* **291**, 6546–6558 (2016).

163. Lee, D. S. M. *et al.* Disrupting upstream translation in mRNAs is associated with human disease. *Nat. Commun.* **12**, 1515 (2021).

164. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 2523 (2020).

165. Steri, M., Idda, M. L., Whalen, M. B. & Orrù, V. Genetic Variants in mRNA Untranslated Regions. *Wiley Interdiscip. Rev. RNA* **9**, e1474 (2018).

166. Jansen, R. P. mRNA localization: message on the move. *Nat. Rev. Mol. Cell Biol.* **2**, 247–256 (2001).

167. Kishore, S. *et al.* A Non-Coding Disease Modifier of Pancreatic Agenesis Identified by Genetic Correction in a Patient-Derived iPSC Line. *Cell Stem Cell* **27**, 137-146.e6 (2020).

168. Amin, A. S. *et al.* Variants in the 3′ untranslated region of the KCNQ1-encoded Kv7.1 potassium channel modify disease severity in patients with type 1 long QT syndrome in an allele-specific manner. *Eur. Heart J.* **33**, 714–723 (2012).
169. Kolder, I. C. R. M. *et al.* Analysis for Genetic Modifiers of Disease Severity in Patients With Long-QT Syndrome Type 2. *Circ. Cardiovasc. Genet.* **8**, 447–456 (2015).
170. Yokoshi, M., Segawa, K. & Fukaya, T. Visualizing the Role of Boundary Elements in Enhancer-Promoter Communication. *Mol. Cell* **78**, 224-235.e5 (2020).
171. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74 (2012).
172. Sun, J. H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224-238.e15 (2018).
173. Emison, E. S. *et al.* Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* **87**, 60–74 (2010).
174. Lee, W. *et al.* Cis-acting modifiers in the ABCA4 locus contribute to the penetrance of the major disease-causing variant in Stargardt disease. *Hum. Mol. Genet.* **30**, 1293–1304 (2021).
175. Zernant, J. *et al.* Extremely hypomorphic and severe deep intronic variants in the ABCA4 locus result in varying Stargardt disease phenotypes. *Cold Spring Harb. Mol. Case Stud.* **4**, a002733 (2018).
176. Park, J. K., Martin, L. J., Zhang, X., Jegga, A. G. & Benson, D. W. Genetic variants in SCN5A promoter are associated with arrhythmia phenotype

severity in patients with heterozygous loss-of-function mutation. *Heart Rhythm* **9**, 1090–1096 (2012).

177. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
178. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).
179. Galupa, R. & Heard, E. Topologically Associating Domains in Chromosome Architecture and Gene Regulatory Landscapes during Development, Disease, and Evolution. *Cold Spring Harb. Symp. Quant. Biol.* **82**, 267–278 (2017).
180. McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.* **108**, 269–283 (2021).
181. Williamson, I. *et al.* Developmentally regulated Shh expression is robust to TAD perturbations. *Development* **146**, dev179523 (2019).
182. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
183. Boltsis, I., Grosveld, F., Giraud, G. & Kolovos, P. Chromatin Conformation in Development and Disease. *Front. Cell Dev. Biol.* **9**, (2021).
184. Lu, L. *et al.* Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Mol. Cell* **79**, 521-534.e15 (2020).
185. Epstein, D. J. Cis-regulatory mutations in human disease. *Brief. Funct. Genomic. Proteomic.* **8**, 310–316 (2009).

186. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

187. Helmbacher, F., Schneider-Maunoury, S., Topilko, P., Tiret, L. & Charnay, P. Targeting of the EphA4 tyrosine kinase receptor affects dorsal/ventral pathfinding of limb motor axons. *Dev. Camb. Engl.* **127**, 3313–3324 (2000).

188. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).

189. Hervé, B. *et al.* Low-level mosaicism of a de novo derivative chromosome 9 from a t(5;9)(q35.1;q34.3) has a major phenotypic impact. *Eur. J. Med. Genet.* **58**, 346–350 (2015).

190. Domogala, D. D. *et al.* Detection of low-level parental somatic mosaicism for clinically relevant SNVs and indels identified in a large exome sequencing dataset. *Hum. Genomics* **15**, 72 (2021).

191. Chen, D. *et al.* The inconsistency between two major aneuploidy-screening platforms—single-nucleotide polymorphism array and next-generation sequencing—in the detection of embryo mosaicism. *BMC Genomics* **23**, 62 (2022).

192. Bickley, V. M. *et al.* Rare Complete and Partial Monosomy 7 Mosaicism Detected in a Case with FTT and Borderline Motor Delay, Subsequently Diagnosed with Juvenile MDS: An Exposition of this Case and Other Interesting Mosaic Cancer Case Studies. *Cancer Genet.* **207**, 286 (2014).

193. Alswied, A. *et al.* Rare monosomy 7 and deletion 7p at diagnosis of chronic myeloid leukemia in accelerated phase. *Cancer Genet.* **252**, 111–114 (2021).

194. Wallis, G. A., Starman, B. J., Zinn, A. B. & Byers, P. H. Variable expression of osteogenesis imperfecta in a nuclear family is explained by somatic mosaicism for a lethal point mutation in the alpha 1(I) gene (COL1A1) of type I collagen in a parent. *Am. J. Hum. Genet.* **46**, 1034–1040 (1990).

195. Cohen, M. M. Proteus syndrome review: molecular, clinical, and pathologic features. *Clin. Genet.* **85**, 111–119 (2014).

196. Ferreira, J. *et al.* CLOVES Syndrome Diagnosis and Treatment in an Adult Patient. *Ann. Vasc. Surg.* **75**, 533.e5-533.e9 (2021).

197. Leon, E., Jamal, S. M., Zou, Y. S. & Milunsky, J. M. Partial trisomy 8 mosaicism due to a pseudoisodicentric chromosome 8. *Am. J. Med. Genet. A.* **155A**, 1740–1744 (2011).

198. Honda, A. *et al.* Somatic HRAS p.G12S mosaic mutation causes unilaterally distributed epidermal nevi, woolly hair and palmoplantar keratosis. *J. Dermatol.* **44**, e109–e110 (2017).

199. Gripp, K. W. & Rauen, K. A. Costello Syndrome. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

200. Tørring, P. M., Kjeldsen, A. D., Ousager, L. B. & Brusgaard, K. ENG mutational mosaicism in a family with hereditary hemorrhagic telangiectasia. *Mol. Genet. Genomic Med.* **6**, 121–125 (2017).

201. Hopp, K. *et al.* Detection and characterization of mosaicism in autosomal dominant polycystic kidney disease. *Kidney Int.* **97**, 370–382 (2020).

202. Jindal, R., Shirazi, N. & Rawat, K. Hereditary segmental neurofibromatosis: a report of three cases in a family. *BMJ Case Rep. CP* **12**, e228826 (2019).

203. Mensa-Vilaró, A. *et al.* Unexpected relevant role of gene mosaicism in patients with primary immunodeficiency diseases. *J. Allergy Clin. Immunol.* **143**, 359–368 (2019).

204. Tuke, M. A. *et al.* Mosaic Turner syndrome shows reduced penetrance in an adult population study. *Genet. Med.* **21**, 877–886 (2019).
205. Lauritsen, K. F. *et al.* A mild form of Stickler syndrome type II caused by mosaicism of COL11A1. *Eur. J. Med. Genet.* **60**, 275–278 (2017).
206. Mastromoro, G. *et al.* Small 7p22.3 microdeletion: Case report of Snx8 haploinsufficiency and neurological findings. *Eur. J. Med. Genet.* **63**, 103772 (2020).
207. Wright, C. F. *et al.* Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nat. Commun.* **10**, 2985 (2019).
208. Campbell, I. M. *et al.* Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. *Am. J. Hum. Genet.* **95**, 173–182 (2014).
209. Acuna-Hidalgo, R. *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *Am. J. Hum. Genet.* **97**, 67–74 (2015).
210. Jing, H. *et al.* Somatic reversion in DOCK8 immunodeficiency modulates disease phenotype. *J. Allergy Clin. Immunol.* **133**, 1667–1675 (2014).
211. Hou, Y. *et al.* Somatic Reversion of a Novel IL2RG Mutation Resulting in Atypical X-Linked Combined Immunodeficiency. *Genes* **13**, 35 (2021).
212. Miyazawa, H. & Wada, T. Reversion Mosaicism in Primary Immunodeficiency Diseases. *Front. Immunol.* **12**, 783022 (2021).
213. Pasmooij, A. M. G. *et al.* Revertant mosaicism due to a second-site mutation in COL7A1 in a patient with recessive dystrophic epidermolysis bullosa. *J. Invest. Dermatol.* **130**, 2407–2411 (2010).

214. Gross, M. *et al.* Reverse mosaicism in Fanconi anemia: natural gene therapy via molecular self-correction. *Cytogenet. Genome Res.* **98**, 126–135 (2002).

215. Nicoletti, E. *et al.* Mosaicism in Fanconi anemia: concise review and evaluation of published cases with focus on clinical course of blood count normalization. *Ann. Hematol.* **99**, 913–924 (2020).

216. Heusinkveld, L. E. *et al.* Pathogenesis, diagnosis and therapeutic strategies in WHIM syndrome immunodeficiency. *Expert Opin. Orphan Drugs* **5**, 813–825 (2017).

217. McDermott, D. H. *et al.* Chromothriptic Cure of WHIM Syndrome. *Cell* **160**, 686–699 (2015).

218. Weinhold, B. Epigenetics: The Science of Change. *Environ. Health Perspect.* **114**, A160–A167 (2006).

219. Velasco, G. & Francastel, C. Genetics meets DNA methylation in rare diseases. *Clin. Genet.* **95**, 210–220 (2019).

220. Castillo-Fernandez, J. E., Spector, T. D. & Bell, J. T. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Med.* **6**, 60 (2014).

221. Rieley, M. B. *et al.* Variable expression of neurofibromatosis 1 in monozygotic twins. *Am. J. Med. Genet. A.* **155**, 478–485 (2011).

222. Biegstraaten, M. *et al.* A monozygotic twin pair with highly discordant Gaucher phenotypes. *Blood Cells. Mol. Dis.* **46**, 39–41 (2011).

223. Lachmann, R. H., Grant, I. R., Halsall, D. & Cox, T. M. Twin pairs showing discordance of phenotype in adult Gaucher's disease. *QJM Int. J. Med.* **97**, 199–204 (2004).

224. Tolmacheva, E. N. *et al.* Delineation of Clinical Manifestations of the Inherited Xq24 Microdeletion Segregating with sXCI in Mothers: Two Novel Cases with Distinct Phenotypes Ranging from UBE2A Deficiency Syndrome to Recurrent Pregnancy Loss. *Cytogenet. Genome Res.* **160**, 245–254 (2020).
225. Skakkebæk, A., Viuff, M., Nielsen, M. M. & Gravholt, C. H. Epigenetics and genomics in Klinefelter syndrome. *Am. J. Med. Genet. C Semin. Med. Genet.* **184**, 216–225 (2020).
226. Vasilyev, S. A. *et al.* Differential DNA Methylation of the IMMP2L Gene in Families with Maternally Inherited 7q31.1 Microdeletions is Associated with Intellectual Disability and Developmental Delay. *Cytogenet. Genome Res.* **161**, 105–119 (2021).
227. Catalanotto, C., Cogoni, C. & Zardo, G. MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *Int. J. Mol. Sci.* **17**, 1712 (2016).
228. Wallace, D. R. *et al.* Toxic-Metal-Induced Alteration in miRNA Expression Profile as a Proposed Mechanism for Disease Development. *Cells* **9**, E901 (2020).
229. Cammaerts, S., Strazisar, M., De Rijk, P. & Del Favero, J. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. *Front. Genet.* **6**, (2015).
230. Tommasi, C. *et al.* Biological Role and Clinical Implications of microRNAs in BRCA Mutation Carriers. *Front. Oncol.* **11**, (2021).
231. Sun, C. *et al.* miR-9 regulation of BRCA1 and ovarian cancer sensitivity to cisplatin and PARP inhibition. *J. Natl. Cancer Inst.* **105**, 1750–1758 (2013).

232. Moskwa, P. *et al.* miR-182-mediated downregulation of BRCA1 impacts DNA repair and sensitivity to PARP inhibitors. *Mol. Cell* **41**, 210–220 (2011).

233. Chang, S. *et al.* Tumor suppressor BRCA1 epigenetically controls oncogenic microRNA-155. *Nat. Med.* **17**, 1275–1282 (2011).

234. Chen, S. & Parmigiani, G. Meta-Analysis of BRCA1 and BRCA2 Penetrance. *J. Clin. Oncol.* **25**, 1329–1333 (2007).

235. Walsh, R., Tadros, R. & Bezzina, C. R. When genetic burden reaches threshold. *Eur. Heart J.* **41**, 3849–3855 (2020).

236. Pizzo, L. *et al.* Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet. Med.* **21**, 816–825 (2019).

237. Ding, J. *et al.* SCN1A Mutation-Beyond Dravet Syndrome: A Systematic Review and Narrative Synthesis. *Front. Neurol.* **12**, 743726 (2021).

238. Hammer, M. F. *et al.* Rare variants of small effect size in neuronal excitability genes influence clinical outcome in Japanese cases of SCN1A truncation-positive Dravet syndrome. *PLOS ONE* **12**, e0180485 (2017).

239. Bertolini, S. *et al.* Homozygous familial hypercholesterolemia in Italy: Clinical and molecular features. *Atherosclerosis* **312**, 72–78 (2020).

240. Arora, V. *et al.* Co-inheritance of pathogenic variants in PKD1 and PKD2 genes presenting as severe antenatal phenotype of autosomal dominant polycystic kidney disease. *Eur. J. Med. Genet.* **63**, 103734 (2020).

241. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* **376**, 21–31 (2017).

242. Racanelli, A. C., Kikkers, S. A., Choi, A. M. K. & Cloonan, S. M. Autophagy and inflammation in chronic respiratory disease. *Autophagy* **14**, 221–232 (2018).

243. Gu, Y. *et al.* Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease. *Nature* **458**, 1039–1042 (2009).

244. Viel, M. *et al.* DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Clin. Respir. J.* **10**, 777–783 (2016).

245. Emond, M. J. *et al.* Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat. Genet.* **44**, 886–889 (2012).

246. Bae, H. T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood* **120**, 1961–1962 (2012).

247. Allard, P. *et al.* Genetic modifiers of fetal hemoglobin affect the course of sickle cell disease in patients treated with hydroxyurea. *Haematologica* (2021) doi:10.3324/haematol.2021.278952.

248. Chang, A. K., Ginter Summarell, C. C., Birdie, P. T. & Sheehan, V. A. Genetic modifiers of severity in sickle cell disease. *Clin. Hemorheol. Microcirc.* **68**, 147–164 (2018).

249. Steinberg, M. H. & Sebastiani, P. Genetic Modifiers of Sickle Cell Disease. *Am. J. Hematol.* **87**, 795–803 (2012).

250. Wonkam, A. *et al.* Genetic modifiers of long-term survival in sickle cell anemia. *Clin. Transl. Med.* **10**, e152 (2020).

251. Crotti, L. *et al.* NOS1AP is a Genetic Modifier of the Long-QT Syndrome. *Circulation* **120**, 1657–1663 (2009).

252. Fahim, A. T. *et al.* Polymorphic Variation of RPGRIP1L and IQCB1 as Modifiers of X-Linked Retinitis Pigmentosa Caused by Mutations in RPGR. *Adv. Exp. Med. Biol.* **723**, 313–320 (2012).

253. Khanna, H. *et al.* A common allele in RPGRIP1L is a modifier of retinal degeneration in ciliopathies. *Nat. Genet.* **41**, 739–745 (2009).

254. Cardenas-Rodriguez, M. *et al.* Characterization of CCDC28B reveals its role in ciliogenesis and provides insight to understand its modifier effect on Bardet-Biedl syndrome. *Hum. Genet.* **132**, 91–105 (2013).

255. Gana, S., Serpieri, V. & Valente, E. M. Genotype–phenotype correlates in Joubert syndrome: A review. *Am. J. Med. Genet. C Semin. Med. Genet.* **n/a**,.

256. Meyer, K. J. *et al.* Genetic modifiers of Cep290-mediated retinal degeneration. 2022.01.26.477596 Preprint at https://doi.org/10.1101/2022.01.26.477596 (2022).

257. Heller, R. & Bolz, H. J. The challenge of defining pathogenicity: the example of AHI1. *Genet. Med.* **17**, 508–508 (2015).

258. Kroes, H. Y. *et al.* DNA analysis of AHI1, NPHP1 and CYCLIN D1 in Joubert syndrome patients from the Netherlands. *Eur. J. Med. Genet.* **51**, 24–34 (2008).

259. Oprea, G. E. *et al.* Plastin 3 is a protective modifier of autosomal recessive spinal muscular atrophy. *Science* **320**, 524–527 (2008).

260. Calucho, M. *et al.* Correlation between SMA type and SMN2 copy number revisited: An analysis of 625 unrelated Spanish patients and a compilation of 2834 reported cases. *Neuromuscul. Disord. NMD* **28**, 208–215 (2018).

261. Magri, S. *et al.* Digenic inheritance of STUB1 variants and TBP polyglutamine expansions explains the incomplete penetrance of SCA17 and SCA48. *Genet. Med.* **24**, 29–40 (2022).
262. Klaassen, K. *et al.* Untreated PKU patients without intellectual disability: SHANK gene family as a candidate modifier. *Mol. Genet. Metab. Rep.* **29**, 100822 (2021).
263. Hanson, E. *et al.* The cognitive and behavioral phenotype of the 16p11.2 deletion in a clinically ascertained population. *Biol. Psychiatry* **77**, 785–793 (2015).
264. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
265. Um, J. W. & Ko, J. Neural Glycosylphosphatidylinositol-Anchored Proteins in Synaptic Specification. *Trends Cell Biol.* **27**, 931–945 (2017).
266. Maussion, G. *et al.* Implication of LRRC4C and DPP6 in neurodevelopmental disorders. *Am. J. Med. Genet. A.* **173**, 395–406 (2017).
267. Poot, M. Connecting the CNTNAP2 Networks with Neurodevelopmental Disorders. *Mol. Syndromol.* **6**, 7–22 (2015).
268. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
269. Schell, R. *et al.* Genetic basis of a spontaneous mutation's expressivity. *Genetics* **220**, iyac013 (2022).
270. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

271. Chami, N., Preuss, M., Walker, R. W., Moscati, A. & Loos, R. J. F. The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. *PLOS Med.* **17**, e1003196 (2020).

272. Oetjens, M. T., Kelly, M. A., Sturm, A. C., Martin, C. L. & Ledbetter, D. H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat. Commun.* **10**, 4897 (2019).

273. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

274. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

275. Kuchenbaecker, K. B. *et al.* Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *J. Natl. Cancer Inst.* **109**, (2017).

276. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2019).

277. van der Kolk, D. M. *et al.* Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age. *Breast Cancer Res. Treat.* **124**, 643–651 (2010).

278. Petrucelli, N., Daly, M. B. & Pal, T. BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

279. Lee, A. *et al.* BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).

280. Gallagher, S. *et al.* Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes. *JAMA Netw. Open* **3**, e208501 (2020).
281. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
282. Harper, A. R. *et al.* Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. *Nat. Genet.* **53**, 135–142 (2021).
283. Natural variants suppress mutations in hundreds of essential genes. *Mol. Syst. Biol.* **17**, e10138 (2021).
284. Buglo, E. *et al.* Genetic compensation in a stable slc25a46 mutant zebrafish: A case for using F0 CRISPR mutagenesis to study phenotypes caused by inherited disease. *PLoS ONE* **15**, e0230566 (2020).
285. Jordan, D. M. *et al.* Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524**, 225–229 (2015).
286. Payne, J. L. & Wagner, A. Mechanisms of mutational robustness in transcriptional regulation. *Front. Genet.* **6**, (2015).
287. Subaran, R. L., Conte, J. M., Stewart, W. C. L. & Greenberg, D. A. Pathogenic EFHC1 mutations are tolerated in healthy individuals dependent on reported ancestry. *Epilepsia* **56**, 188–194 (2015).
288. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
289. El-Brolosy, M. A. & Stainier, D. Y. R. Genetic compensation: A phenomenon in search of mechanisms. *PLOS Genet.* **13**, e1006780 (2017).

290. Chen, W.-H., Zhao, X.-M., van Noort, V. & Bork, P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput. Biol.* **9**, e1003073 (2013).
291. Understanding redundancy and resilience. *EMBO Rep.* **23**, e54742 (2022).
292. Huang, L. & Wilkinson, M. F. Regulation of nonsense-mediated mRNA decay. *Wiley Interdiscip. Rev. RNA* **3**, 807–828 (2012).
293. Sarri, C. A. *et al.* Netherton Syndrome: A Genotype-Phenotype Review. *Mol. Diagn. Ther.* **21**, 137–152 (2017).
294. Sato, H. & Singer, R. H. Cellular variability of nonsense-mediated mRNA decay. *Nat. Commun.* **12**, 7203 (2021).
295. Sarkar, H. *et al.* Nonsense-mediated mRNA decay efficiency varies in choroideremia providing a target to boost small molecule therapeutics. *Hum. Mol. Genet.* **28**, 1865–1871 (2019).
296. Hildebrand, M. S. *et al.* Severe childhood speech disorder: Gene discovery highlights transcriptional dysregulation. *Neurology* **94**, e2148–e2167 (2020).
297. Nguyen, L. S. *et al.* Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability. *Mol. Psychiatry* **17**, 1103–1115 (2012).
298. Miller, E. E. *et al.* EIF4A3 deficient human iPSCs and mouse models demonstrate neural crest defects that underlie Richieri-Costa-Pereira syndrome. *Hum. Mol. Genet.* **26**, 2177–2191 (2017).
299. Alzahrani, F. *et al.* Recessive, Deleterious Variants in SMG8 Expand the Role of Nonsense-Mediated Decay in Developmental Disorders in Humans. *Am. J. Hum. Genet.* **107**, 1178–1185 (2020).

300. Nguyen, L. S. *et al.* Contribution of copy number variants involving nonsense-mediated mRNA decay pathway genes to neuro-developmental disorders. *Hum. Mol. Genet.* **22**, 1816–1825 (2013).

301. Khajavi, M., Inoue, K. & Lupski, J. R. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur. J. Hum. Genet.* **14**, 1074–1081 (2006).

302. Dyle, M. C., Kolakada, D., Cortazar, M. A. & Jagannathan, S. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated RNA decay. *Wiley Interdiscip. Rev. RNA* **11**, e1560 (2020).

303. Supek, F., Lehner, B. & Lindeboom, R. G. H. To NMD or Not To NMD: Nonsense-Mediated mRNA Decay in Cancer and Other Genetic Diseases. *Trends Genet.* **37**, 657–668 (2021).

304. Torella, A. *et al.* The position of nonsense mutations can predict the phenotype severity: A survey on the DMD gene. *PLOS ONE* **15**, e0237803 (2020).

305. Kerr, T. P., Sewry, C. A., Robb, S. A. & Roberts, R. G. Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? *Hum. Genet.* **109**, 402–407 (2001).

306. Brewer, H. R., Jones, M. E., Schoemaker, M. J., Ashworth, A. & Swerdlow, A. J. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Res. Treat.* **165**, 193–200 (2017).

307. Turner, H. & Jackson, L. Evidence for penetrance in patients without a family history of disease: a systematic review. *Eur. J. Hum. Genet.* **28**, 539–550 (2020).

308. Moreno-De-Luca, A. *et al.* The Role of Parental Cognitive, Behavioral, and Motor Profiles in Clinical Variability in Individuals With Chromosome 16p11.2 Deletions. *JAMA Psychiatry* **72**, 119–126 (2015).
309. Olszewski, A. K., Radoeva, P. D., Fremont, W., Kates, W. R. & Antshel, K. M. 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**, 3582–3590 (2014).
310. De Smedt, B. *et al.* Intellectual abilities in a large sample of children with Velo-Cardio-Facial Syndrome: an update. *J. Intellect. Disabil. Res. JIDR* **51**, 666–670 (2007).
311. Klaassen, P. *et al.* Explaining the variable penetrance of CNVs: Parental intelligence modulates expression of intellectual impairment caused by the 22q11.2 deletion. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **171**, 790–796 (2016).
312. Davies, R. W. *et al.* Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat. Med.* **26**, 1912–1918 (2020).
313. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
314. Fetit, R., Price, D. J., Lawrie, S. M. & Johnstone, M. Understanding the clinical manifestations of 16p11.2 deletion syndrome: a series of developmental case reports in children. *Psychiatr. Genet.* **30**, 136–140 (2020).

315. Rosenfeld, J. A., Coe, B. P., Eichler, E. E., Cuckle, H. & Shaffer, L. G. Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15**, 478–481 (2013).
316. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
317. Polyak, A., Rosenfeld, J. A. & Girirajan, S. An assessment of sex bias in neurodevelopmental disorders. *Genome Med.* **7**, 94 (2015).
318. Evans, D. W. & Uljarević, M. Parental education accounts for variability in the IQs of probands with Down syndrome: A longitudinal study. *Am. J. Med. Genet. A.* **176**, 29–33 (2018).
319. Brookes, E. & Shi, Y. Diverse Epigenetic Mechanisms of Human Disease. *Annu. Rev. Genet.* **48**, 237–268 (2014).
320. Bashkeel, N., Perkins, T. J., Kærn, M. & Lee, J. M. Human gene expression variability and its dependence on methylation and aging. *BMC Genomics* **20**, 941 (2019).
321. Correa, H. Li-Fraumeni Syndrome. *J. Pediatr. Genet.* **5**, 84–88 (2016).
322. Biller, L. H., Syngal, S. & Yurgelun, M. B. Recent advances in Lynch syndrome. *Fam. Cancer* **18**, 211–219 (2019).
323. White, M. C. *et al.* Age and Cancer Risk. *Am. J. Prev. Med.* **46**, S7-15 (2014).
324. Antoniou, A. *et al.* Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *Am. J. Hum. Genet.* **72**, 1117–1130 (2003).

325. Chiti, F. & Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).

326. Rose, A. m. & Bhattacharya, S. s. Variant haploinsufficiency and phenotypic non-penetrance in PRPF31-associated retinitis pigmentosa. *Clin. Genet.* **90**, 118–126 (2016).

327. Magrinelli, F., Balint, B. & Bhatia, K. P. Challenges in Clinicogenetic Correlations: One Gene – Many Phenotypes. *Mov. Disord. Clin. Pract.* **8**, 299–310 (2021).

328. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).

329. Kim, Y. & Connor, J. R. The roles of iron and HFE genotype in neurological diseases. *Mol. Aspects Med.* **75**, 100867 (2020).

330. Pilling, L. C. *et al.* Common conditions associated with hereditary haemochromatosis genetic variants: cohort study in UK Biobank. *BMJ* **364**, k5222 (2019).

331. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).

332. Scott, F. J., Baron-Cohen, S., Bolton, P. & Brayne, C. Brief Report Prevalence of Autism Spectrum Conditions in Children Aged 5-11 Years in Cambridgeshire, UK. *Autism* **6**, 231–237 (2002).

333. Christensen, D. L. *et al.* Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and

Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR Surveill. Summ.* **65**, 1–23 (2016).

334. Jacquemont, S. *et al.* A Higher Mutational Burden in Females Supports a "Female Protective Model" in Neurodevelopmental Disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
335. Ratto, A. B. *et al.* What About the Girls? Sex-Based Differences in Autistic Traits and Adaptive Skills. *J. Autism Dev. Disord.* **48**, 1698–1711 (2018).
336. Russell, G., Steer, C. & Golding, J. Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Soc. Psychiatry Psychiatr. Epidemiol.* **46**, 1283–1293 (2011).
337. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).
338. Dolinoy, D. C., Weidman, J. R. & Jirtle, R. L. Epigenetic gene regulation: linking early developmental environment to adult disease. *Reprod. Toxicol. Elmsford N* **23**, 297–307 (2007).
339. Safi-Stibler, S. & Gabory, A. Epigenetics and the Developmental Origins of Health and Disease: Parental environment signalling to the epigenome, critical time windows and sculpting the adult phenotype. *Semin. Cell Dev. Biol.* **97**, 172–180 (2020).
340. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
341. van Karnebeek, C. D. M. & Stockler, S. Treatable inborn errors of metabolism causing intellectual disability: A systematic literature review. *Mol. Genet. Metab.* **105**, 368–381 (2012).
342. Flydal, M. I. & Martinez, A. Phenylalanine hydroxylase: function, structure, and regulation. *IUBMB Life* **65**, 341–349 (2013).

343. Al Hafid, N. & Christodoulou, J. Phenylketonuria: a review of current and future treatments. *Transl. Pediatr.* **4**, 304–317 (2015).

344. Lee, J., Meyerhardt, J. A., Giovannucci, E. & Jeon, J. Y. Association between Body Mass Index and Prognosis of Colorectal Cancer: A Meta-Analysis of Prospective Cohort Studies. *PLoS ONE* **10**, e0120706 (2015).

345. Newcomb, P. A. & Carbone, P. P. The health consequences of smoking. Cancer. *Med. Clin. North Am.* **76**, 305–331 (1992).

346. Wu, S. *et al.* History of Severe Sunburn and Risk of Skin Cancer Among Women and Men in 2 Prospective Cohort Studies. *Am. J. Epidemiol.* **183**, 824–833 (2016).

347. Wei, J., Rahman, S., Ayaub, E. A., Dickhout, J. G. & Ask, K. Protein Misfolding and Endoplasmic Reticulum Stress in Chronic Lung Disease. *Chest* **143**, 1098–1105 (2013).

348. Tukker, A. M., Royal, C. D., Bowman, A. B. & McAllister, K. A. The Impact of Environmental Factors on Monogenic Mendelian Diseases. *Toxicol. Sci.* **181**, 3–12 (2021).

349. Collaco, J. M., Blackman, S. M., McGready, J., Naughton, K. M. & Cutting, G. R. Quantification of the Relative Contribution of Environmental and Genetic Factors to Variation in Cystic Fibrosis Lung Function. *J. Pediatr.* **157**, 802-807.e3 (2010).

350. Schindler, T., Michel, S. & Wilson, A. W. M. Nutrition Management of Cystic Fibrosis in the 21st Century. *Nutr. Clin. Pract. Off. Publ. Am. Soc. Parenter. Enter. Nutr.* **30**, 488–500 (2015).

351. Collaco, J. M. *et al.* Effect of temperature on cystic fibrosis lung disease and infections: a replicated cohort study. *PloS One* **6**, e27784 (2011).

352. Collaco, J. M. & Cutting, G. R. Update on gene modifiers in cystic fibrosis. *Curr. Opin. Pulm. Med.* **14**, 559–566 (2008).

353. Sacco, K. A. & Milner, J. D. Gene-environment interactions in primary atopic disorders. *Curr. Opin. Immunol.* **60**, 148–155 (2019).

354. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).

355. Mirshahi, U. L. *et al.* The penetrance of age-related monogenic disease depends on ascertainment context. *medRxiv* 2021.06.28.21259641 (2021) doi:10.1101/2021.06.28.21259641.

356. Kingdom, R. *et al.* Rare genetic variants in dominant developmental disorder loci cause milder related phenotypes in the general population. 2021.12.15.21267855 Preprint at https://doi.org/10.1101/2021.12.15.21267855 (2021).

357. Harrison, S. M. & Rehm, H. L. Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. *Genome Med.* **11**, 72 (2019).

358. Hosseini, S. M. *et al.* Reappraisal of Reported Genes for Sudden Arrhythmic Death: Evidence-Based Evaluation of Gene Validity for Brugada Syndrome. *Circulation* **138**, 1195–1205 (2018).

359. Hanany, M. & Sharon, D. Allele frequency analysis of variants reported to cause autosomal dominant inherited retinal diseases question the involvement of 19% of genes and 10% of reported pathogenic variants. *J. Med. Genet.* **56**, 536–542 (2019).

360. Guillot, L. *et al.* Lung disease modifier genes in cystic fibrosis. *Int. J. Biochem. Cell Biol.* **52**, 83–93 (2014).

361. Pereira, S. V.-N., Ribeiro, J. D., Ribeiro, A. F., Bertuzzo, C. S. & Marson, F. A. L. Novel, rare and common pathogenic variants in the CFTR gene

screened by high-throughput sequencing technology and predicted by in silico tools. *Sci. Rep.* **9**, 6234 (2019).

362. Hassanin, E. *et al. Assessing the role of polygenic background on the penetrance of monogenic forms in Parkinson's disease*. http://medrxiv.org/lookup/doi/10.1101/2021.06.06.21253270 (2021) doi:10.1101/2021.06.06.21253270.
363. Peltonen, L., Perola, M., Naukkarinen, J. & Palotie, A. Lessons from studying monogenic disease for common disease. *Hum. Mol. Genet.* **15**, R67–R74 (2006).
364. Freund, M. K. *et al.* Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).
365. Senol-Cosar, O. *et al.* Considerations for clinical curation, classification, and reporting of low-penetrance and low effect size variants associated with disease risk. *Genet. Med.* **21**, 2765–2773 (2019).
366. Wei, W.-H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).
367. Lee, I. *et al.* Predicting genetic modifier loci using functional gene networks. *Genome Res.* **20**, 1143–1153 (2010).
368. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLOS Genet.* **11**, e1005492 (2015).
369. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
370. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).

371. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
372. Maya, I., Sukenik-Halevy, R., Basel-Salmon, L. & Sagi-Dain, L. Ten points to consider when providing genetic counseling for variants of incomplete penetrance and variable expressivity detected in a prenatal setting. *Acta Obstet. Gynecol. Scand.* **99**, 1427–1429 (2020).
373. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
374. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
375. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
376. Qi, H., Dong, C., Chung, W. K., Wang, K. & Shen, Y. Deep genetic connection between cancer and developmental disorders. *Hum. Mutat.* **37**, 1042–1050 (2016).
377. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
378. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
379. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).

380. Kendall, K. M. *et al.* Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br. J. Psychiatry* **214**, 297–304 (2019).
381. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
382. Gardner, E. J. *et al.* Reduced reproductive success is associated with selective constraint on human genes. *Nature* **603**, 858–863 (2022).
383. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
384. Plomin, R. & Deary, I. J. Genetics and intelligence differences: five special findings. *Mol. Psychiatry* **20**, 98–108 (2015).
385. Kingdom, R. & Wright, C. F. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. *Front. Genet.* **13**, (2022).
386. Cable, J. *et al.* Harnessing rare variants in neuropsychiatric and neurodevelopment disorders—a Keystone Symposia report. *Ann. N. Y. Acad. Sci.* **1506**, 5–17 (2021).
387. Kurki, M. I. *et al.* Contribution of rare and common variants to intellectual disability in a sub-isolate of Northern Finland. *Nat. Commun.* **10**, 410 (2019).
388. Bergen, S. E. *et al.* Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia. *Am. J. Psychiatry* **176**, 29–35 (2019).
389. Hong, E. P., Heo, S. G. & Park, J. W. The Liability Threshold Model for Predicting the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes: A Multi-Cohort Study of Korean Adults. *Metabolites* **11**, 6 (2020).

390. Zhou, D. *et al.* Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat. Commun.* **12**, 4418 (2021).
391. Antaki, D. *et al.* A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. *Nat. Genet.* **54**, 1284–1292 (2022).
392. Jukarainen, S. *et al.* Genetic risk factors have a substantial impact on healthy life years. 2022.01.25.22269831 Preprint at https://doi.org/10.1101/2022.01.25.22269831 (2022).
393. Genç, E. *et al.* Polygenic Scores for Cognitive Abilities and Their Association with Different Aspects of General Intelligence—A Deep Phenotyping Approach. *Mol. Neurobiol.* **58**, 4145–4156 (2021).
394. Thompson, D. J. *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. 2022.06.16.22276246 Preprint at https://doi.org/10.1101/2022.06.16.22276246 (2022).
395. Smail, C. *et al.* Integration of rare expression outlier-associated variants improves polygenic risk prediction. *Am. J. Hum. Genet.* **109**, 1055–1064 (2022).
396. Darst, B. F. *et al.* Combined Effect of a Polygenic Risk Score and Rare Genetic Variants on Prostate Cancer Risk. *Eur. Urol.* **80**, 134–138 (2021).
397. Kingdom, R. *et al.* Rare genetic variants in genes and loci linked to dominant monogenic developmental disorders cause milder related phenotypes in the general population. *Am. J. Hum. Genet.* (2022) doi:10.1016/j.ajhg.2022.05.011.

398. Oetjens, M. T., Kelly, M. A., Sturm, A. C., Martin, C. L. & Ledbetter, D. H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat. Commun.* **10**, 4897 (2019).
399. Cameli, C. *et al.* An increased burden of rare exonic variants in NRXN1 microdeletion carriers is likely to enhance the penetrance for autism spectrum disorder. *J. Cell. Mol. Med.* **25**, 2459–2470 (2021).
400. Liu, H. *et al.* Polygenic Resilience Modulates the Penetrance of Parkinson Disease Genetic Risk Factors. *Ann. Neurol.* **92**, 270–278 (2022).
401. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
402. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).
403. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
404. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
405. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
406. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).

407. Beaumont, R. N., Mayne, I. K., Freathy, R. M. & Wright, C. F. Common genetic variants with fetal effects on birth weight are enriched for proximity to genes implicated in rare developmental disorders. *Hum. Mol. Genet.* **30**, 1057–1066 (2021).

408. Valli, I., Fabbri, C. & Young, A. H. Uncovering neurodevelopmental features in bipolar affective disorder. *Br. J. Psychiatry* **215**, 383–385 (2019).

409. Burdick, K. E. *et al.* Empirical evidence for discrete neurocognitive subgroups in bipolar disorder: clinical implications. *Psychol. Med.* **44**, 3083–3096 (2014).

410. Bora, E. Differences in cognitive impairment between schizophrenia and bipolar disorder: Considering the role of heterogeneity. *Psychiatry Clin. Neurosci.* **70**, 424–433 (2016).

411. Bora, E. & Özerdem, A. Meta-analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia. *Psychol. Med.* **47**, 2753–2766 (2017).

412. Legge, S. E. *et al.* Association of Genetic Liability to Psychotic Experiences With Neuropsychotic Disorders and Traits. *JAMA Psychiatry* **76**, 1256–1265 (2019).

413. Lam, M. *et al.* Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. *Am. J. Hum. Genet.* **105**, 334–350 (2019).

414. Lu, T., Forgetta, V., Richards, J. B. & Greenwood, C. M. T. Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. *Genet. Med.* (2022) doi:10.1016/j.gim.2022.03.022.

415. Lu, T. *et al.* Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. *Genet. Med.* **23**, 508–515 (2021).

416. May, T., Adesina, I., McGillivray, J. & Rinehart, N. J. Sex differences in neurodevelopmental disorders. *Curr. Opin. Neurol.* **32**, 622 (2019).

417. Morris-Rosendahl, D. J. & Crocq, M.-A. Neurodevelopmental disorders—the history and future of a diagnostic concept. *Dialogues Clin. Neurosci.* **22**, 65–72 (2020).

418. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).

419. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).

420. Männik, K. *et al.* Copy Number Variations and Cognitive Phenotypes in Unselected Populations. *JAMA* **313**, 2044–2054 (2015).

421. Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Modification of Heritability for Educational Attainment and Fluid Intelligence by Socioeconomic Deprivation in the UK Biobank. *Am. J. Psychiatry* **178**, 625–634 (2021).

422. Genes influence complex traits through environments that vary between geographic regions. *Nat. Genet.* **54**, 1265–1266 (2022).

423. Chia-Yen Chen *et al.* The impact of rare protein coding genetic variation on adult cognitive function. *medRxiv* 2022.06.24.22276728 (2022) doi:10.1101/2022.06.24.22276728.

424. Lebrun, N. *et al.* Novel KDM5B splice variants identified in patients with developmental disorders: Functional consequences. *Gene* **679**, 305–313 (2018).
425. Lebon, S. *et al.* Agenesis of the Corpus Callosum with Facial Dysmorphism and Intellectual Disability in Sibs Associated with Compound Heterozygous KDM5B Variants. *Genes* **12**, 1397 (2021).
426. Mangano, G. D. *et al.* A complex epileptic and dysmorphic phenotype associated with a novel frameshift KDM5B variant and deletion of SCN gene cluster. *Seizure* **97**, 20–22 (2022).
427. Faundes, V. *et al.* Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *Am. J. Hum. Genet.* **102**, 175–187 (2018).
428. KDM5B | gnomAD. https://gnomad.broadinstitute.org/gene/ENSG00000117139?dataset=gnomad_r2_1.
429. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
430. He, R. & Kidder, B. L. H3K4 demethylase KDM5B regulates global dynamics of transcription elongation and alternative splicing in embryonic stem cells. *Nucleic Acids Res.* **45**, 6427–6441 (2017).
431. Chromatin modifying enzymes | HUGO Gene Nomenclature Committee. https://www.genenames.org/data/genegroup/#!/group/484.
432. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).

433. El Hayek, L. *et al.* KDM5A mutations identified in autism spectrum disorder using forward genetics. *eLife* **9**, e56883 (2020).
434. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478**, 57–63 (2011).
435. Adegbola, A., Gao, H., Sommer, S. & Browning, M. A novel mutation in JARID1C/SMCX in a patient with autism spectrum disorder (ASD). *Am. J. Med. Genet. A.* **146A**, 505–511 (2008).
436. Jensen, L. R. *et al.* Mutations in the JARID1C Gene, Which Is Involved in Transcriptional Regulation and Chromatin Remodeling, Cause X-Linked Mental Retardation. *Am. J. Hum. Genet.* **76**, 227–236 (2005).
437. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
438. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
439. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
440. Rojano, E., Seoane, P., Ranea, J. A. G. & Perkins, J. R. Regulatory variants: from detection to predicting impact. *Brief. Bioinform.* **20**, 1639–1654 (2019).
441. VandenBosch, L. S. *et al.* Machine Learning Prediction of Non-Coding Variant Impact in Human Retinal cis-Regulatory Elements. *Transl. Vis. Sci. Technol.* **11**, 16 (2022).

442. Moyon, L., Berthelot, C., Louis, A., Nguyen, N. T. T. & Roest Crollius, H. Classification of non-coding variants with high pathogenic impact. *PLoS Genet.* **18**, e1010191 (2022).

443. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102-110 (2015).

444. Dong, S. & Boyle, A. P. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.* **50**, e6 (2022).

445. Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R. & Magni, P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci. Rep.* **12**, 2517 (2022).

446. Schubach, M., Re, M., Robinson, P. N. & Valentini, G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci. Rep.* **7**, 2959 (2017).

447. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).

448. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).

449. MacEachern, S. J. & Forkert, N. D. Machine learning for precision medicine. *Genome* **64**, 416–425 (2021).

450. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).

451. Alharbi, W. S. & Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum. Genomics* **16**, 26 (2022).

452. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
453. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
454. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
455. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
456. Zhou, J. *et al.* Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
457. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
458. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
459. Jensen, L. R. *et al.* A distinctive gene expression fingerprint in mentally retarded male patients reflects disease-causing defects in the histone demethylase KDM5C. *PathoGenetics* **3**, 2 (2010).
460. Ohnmacht, J., May, P., Sinkkonen, L. & Krüger, R. Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. *J. Neural Transm. Vienna Austria 1996* **127**, 729–748 (2020).
461. Chen, L. & Li, M. J. Editorial: Deciphering Non-Coding Regulatory Variants: Computational and Functional Validation. *Front. Bioeng. Biotechnol.* **9**, (2021).
462. Castellanos-Rubio, A. & Ghosh, S. Disease-Associated SNPs in Inflammation-Related lncRNAs. *Front. Immunol.* **10**, 420 (2019).

463. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
464. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
465. Elfatih, A., Mohammed, I., Abdelrahman, D. & Mifsud, B. Frequency and management of medically actionable incidental findings from genome and exome sequencing data: a systematic review. *Physiol. Genomics* **53**, 373–384 (2021).
466. Lawson, D. J. *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* **139**, 23–41 (2020).
467. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
468. Walsh, R. *et al.* Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: the case of hypertrophic cardiomyopathy. *Genome Med.* **11**, 5 (2019).
469. Zhang, X. *et al.* Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genet. Med.* **23**, 69–79 (2021).
470. Holm, I. A. *et al.* The BabySeq project: implementing genomic sequencing in newborns. *BMC Pediatr.* **18**, 225 (2018).
471. Wojcik, M. H. *et al.* Discordant results between conventional newborn screening and genomic sequencing in the BabySeq Project. *Genet. Med.* **23**, 1372–1375 (2021).
472. Gordon, A. S. *et al.* Frequency of genomic secondary findings among 21,915 eMERGE network participants. *Genet. Med.* **22**, 1470–1477 (2020).

473. Hart, M. R. *et al.* Secondary findings from clinical genomic sequencing: Prevalence, patient perspectives, family history assessment, and healthcare costs from a multi-site study. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **21**, 1100–1110 (2019).
474. Johnston, J. J. *et al.* Secondary Variants in Individuals Undergoing Exome Sequencing: Screening of 572 Individuals Identifies High-Penetrance Mutations in Cancer-Susceptibility Genes. *Am. J. Hum. Genet.* **91**, 97–108 (2012).

Appendix

Tables and Figures for Chapter Three

Appendix table 7.3.1: List of genes included in each subset

Gene Name	599 Gene Set	325 Gene Set	125 Gene Set	25 Gene Set
ABCC9	x			
ABL1	x			
ACAN	x	x		
ACTB	x	x		
ACTG1	x			
ACVR1	x			
ADAR	x	x		
ADCY5	x			
ADNP	x	x	x	x
AFF3	x			
AFF4	x			
AGO1	x			
AHDC1	x	x	x	x
ALDH18A1	x			
ALX4	x	x		
ANKH	x	x		
ANKRD11	x	x	x	x
ANKRD26	x			
AP2M1	x			
AP2S1	x			
ARCN1	x	x		
ARF1	x			
ARHGAP31	x	x		
ARHGAP35	x	x		
ARID1A	x	x	x	
ARID1B	x	x	x	x
ARID2	x	x	x	
ASH1L	x			
ASXL1	x	x	x	
ASXL2	x	x		
ASXL3	x	x	x	x
ATAD3A	x			
ATN1	x			
ATP1A1	x			
ATP1A2	x	x		
ATP6V0A1	x			
ATP6V1B2	x			

ATPA2	x				
AUTS2	x	x	x		
BCL11A	x	x	x		
BCL11B	x	x	x		
BFSP2	x				
BHLHA9	x				
BICD2	x				
BMP2	x	x			
BMP4	x	x			
BNC2	x	x			
BPTF	x	x			
BRAF	x				
BRD4	x	x			
BRPF1	x	x	x		
BRSK2	x	x			
CACNA1A	x				
CACNA1C	x				
CACNA1D	x				
CACNA1E	x				
CACNA1G	x	x			
CAMK2A	x	x			
CAMK2B	x	x			
CAMTA1	x	x	x		
CBL	x				
CCDC78	x	x			
CCND2	x				
CD96	x	x			
CDH2	x				
CDK13	x				
CDK8	x				
CDON	x				
CERT1	x				
CFC1	x	x			
CHAMP1	x	x	x		
CHD1	x	x			
CHD2	x	x	x		
CHD3	x		x		
CHD4	x	x			
CHD7	x	x	x	x	
CHD8	x	x	x		
CHRNA4	x				
CHRNB2	x				
CIC	x	x			
CLTC	x	x	x		
CNOT1	x				

CNOT3	x	x	x		
COG4	x				
COL10A1	x				
COL11A1	x				
COL11A2	x				
COL1A1	x				
COL2A1	x	x			
COL4A3	x				
COL6A1	x				
COL9A1	x				
COL9A2	x				
COL9A3	x				
COMP	x				
CREBBP	x	x	x		
CRELD1	x				
CRX	x	x			
CRYAA	x				
CRYBA1	x	x			
CRYBA4	x				
CRYBB1	x	x			
CRYBB2	x	x			
CRYGC	x	x			
CRYGD	x				
CSNK2A1	x				
CSNK2B	x	x	x		
CTBP1	x	x			
CTCF	x	x	x		
CTNNB1	x	x	x	x	
CTNND1	x	x			
CTNND2	x	x			
CUL3	x				
CUX2	x				
DDX23	x				
DDX6	x				
DEAF1	x				
DEPDC5	x	x			
DHDDS	x				
DHX30	x				
DLG4	x	x	x		
DLL4	x	x			
DMPK	x				
DNM1	x				
DNM1L	x				
DNMT3A	x	x	x		
DPF2	x				

DPYSL5	x			
DSPP	x	x		
DSTYK	x	x		
DVL1	x			
DVL3	x			
DYNC1H1	x			
DYRK1A	x	x	x	x
EBF3	x	x		
EDAR	x		x	
EDN1	x	x		
EDNRA	x			
EDNRB	x	x		
EED	x			
EEF1A2	x			
EEF2	x			
EFTUD2	x	x	x	x
EHMT1	x	x	x	x
EIF5A	x			
ELN	x	x		
EP300	x	x	x	x
ERF	x	x		
EXT1	x	x		
EXT2	x	x		
EYA1	x	x		
EZH2	x			
FAM111A	x			
FBN1	x	x	x	
FBN2	x			
FBXO11	x	x	x	
FBXW11	x			
FBXW7	x			
FGF10	x	x		
FGF12	x			
FGF14	x	x		
FGF9	x	x		
FGFR1	x	x		
FGFR2	x			
FGFR3	x			
FLNB	x			
FLT4	x			
FN1	x			
FOXC1	x	x		
FOXC2	x	x		
FOXE3	x	x		
FOXF1	x	x		

FOXG1	x	x	x		
FOXJ1	x	x			
FOXL2	x	x			
FOXP1	x	x	x		
FOXP2	x	x	x		
FTL	x	x			
FZD5	x				
GABBR2	x				
GABRA1	x	x			
GABRB2	x				
GABRB3	x				
GABRG2	x	x			
GATA2	x	x			
GATA3	x	x	x		
GATA4	x	x			
GATA6	x	x			
GATAD2B	x	x	x	x	
GCH1	x				
GDF5	x	x			
GDF6	x				
GFAP	x				
GIGYF1	x	x	x		
GJA1	x				
GJA3	x				
GJA8	x				
GJC2	x				
GLI2	x	x			
GLI3	x	x			
GLMN	x	x			
GLUD1	x				
GMNN	x				
GNAI1	x	x			
GNAI3	x				
GNAO1	x	x			
GNAS	x	x			
GNB1	x				
GNB2	x				
GREB1L	x	x			
GRHL3	x	x			
GRIA2	x	x			
GRIN1	x				
GRIN2A	x	x			
GRIN2B	x	x	x		
GRIN2D	x				
GUCY2C	x				

H1-4	x	x			
HCN1	x				
HDAC4	x	x			
HECW2	x				
HESX1	x	x			
HIST1H1E	x		x		
HIST1H2AC	x				
HIST1H4C	x				
HIVEP2	x	x	x		
HK1	x	x			
HNF1B	x	x			
HNF4A	x	x			
HNRNPD	x	x	x		
HNRNPH1	x	x			
HNRNPK	x	x	x		
HNRNPR	x	x			
HNRNPU	x	x	x		
HOXA13	x	x			
HOXD13	x				
HPD	x				
HRAS	x				
HSF4	x				
IFIH1	x				
IFITM5	x				
IGF1R	x	x			
IHH	x				
IRF2BPL	x	x	x		
IRF6	x	x			
ITPR1	x	x			
JAG1	x	x			
KANSL1	x	x	x		
KAT6A	x	x	x	x	
KAT6B	x	x	x	x	
KBTBD13	x				
KCNA2	x	x			
KCNB1	x		x		
KCNC1	x				
KCNC3	x				
KCND3	x				
KCNH1	x				
KCNJ11	x				
KCNJ6	x				
KCNJ8	x				
KCNK3	x				
KCNK4	x				

KCNN3	x				
KCNQ2	x	x			
KCNQ3	x				
KCNQ5	x	x			
KCNT1	x				
KCTD1	x				
KDM1A	x				
KDM3B	x	x			
KDM5B	x	x	x		
KDM6B	x	x	x		
KIDINS220	x	x	x		
KIF11	x	x	x		
KIF1A	x				
KIF22	x				
KIF2A	x				
KIF5C	x				
KLF1	x				
KMT2A	x	x	x	x	
KMT2B	x	x	x		
KMT2C	x	x	x		
KMT2D	x	x	x	x	
KMT2E	x	x	x		
KMT5B	x	x	x		
KRAS	x				
KRT74	x				
LEMD2	x				
LEMD3	x	x			
LHX4	x	x			
LMX1B	x	x			
LRP5	x	x			
LZTR1	x				
MAB21L2	x				
MACF1	x				
MAF	x				
MAFB	x	x			
MAP2K1	x				
MAP2K2	x				
MAP3K1	x	x			
MAP3K7	x				
MAPK8IP3	x	x			
MAST1	x				
MATN3	x				
MBD5	x	x	x		
MECOM	x				
MED13	x	x			

MED13L	x	x	x	x
MEF2C	x	x	x	
MEIS2	x	x		
MIB1	x	x	x	
MIR17HG	x	x		
MITF	x	x		
MMP13	x			
MN1	x	x	x	
MNX1	x	x		
MSL2	x	x	x	
MSX1	x	x		
MSX2	x	x		
MTOR	x			
MYCN	x	x	x	
MYH3	x			
MYH9	x			
MYRF	x	x		
MYT1L	x	x	x	
NAA15	x	x	x	
NACC1	x			
NALCN	x			
NBEA	x	x		
NDNF	x	x		
NEDD4L	x			
NF1	x	x		
NFIA	x	x	x	
NFIB	x	x		
NFIX	x	x	x	
NIPBL	x	x		
NKX2-1	x	x		
NKX2-5	x	x		
NODAL	x	x		
NOG	x	x		
NOTCH1	x	x		
NOTCH2	x			
NOVA2	x	x		
NPM1	x			
NR2F1	x	x		
NR2F2	x	x		
NR4A2	x	x		
NRAS	x			
NRXN1	x	x		
NRXN2	x	x		
NSD1	x	x	x	
NSD2	x	x	x	

NTRK2	x				
NUS1	x	x			
ODC1	x	x	x		
OTX2	x	x			
P4HB	x				
PACS1	x				
PACS2	x				
PAFAH1B1	x	x			
PAK1	x				
PAX2	x	x			
PAX3	x	x			
PAX6	x	x			
PAX8	x	x			
PAX9	x	x			
PBX1	x	x			
PCBP2	x				
PCGF2	x				
PDE10A	x				
PDE4D	x				
PDGFRB	x				
PHACTR1	x				
PHF12	x				
PHF21A	x	x	x		
PHIP	x	x	x		
PHOX2B	x				
PIEZO2	x	x			
PIGU	x				
PIK3R1	x				
PIK3R2	x				
PITX1	x	x			
PITX2	x	x			
PITX3	x	x			
PLCB4	x				
POGZ	x	x	x	x	
POLR1A	x	x			
POLR1D	x	x			
POLR2A	x				
POU3F3	x		x		
PPM1D	x	x	x		
PPP1CB	x				
PPP1R12A	x	x			
PPP2CA	x	x			
PPP2R1A	x				
PPP2R5D	x				
PPP3CA	x				

PRKAR1A	x				
PRKD1	x				
PRPF8	x				
PRR12	x	x	x		
PRRT2	x	x			
PSMC5	x				
PTCH1	x	x	x		
PTDSS1	x				
PTEN	x	x	x		
PTH1R	x	x			
PTHLH	x	x			
PTPN11	x				
PUF60	x	x	x		
PURA	x	x	x	x	
QRICH1	x	x	x		
RAB11A	x				
RAB11B	x				
RAB14	x				
RAC1	x				
RAC3	x				
RAD21	x	x			
RAF1	x				
RAI1	x	x	x		
RARB	x				
RBPJ	x				
RERE	x	x			
RHOBTB2	x				
RIT1	x				
RNF13	x				
ROBO4	x	x			
ROR2	x	x			
RORA	x	x			
RPL11	x	x			
RPS19	x	x			
RPS23	x				
RPS26	x	x			
RRAS	x				
RRAS2	x				
RUNX2	x	x			
SALL1	x	x			
SALL4	x	x			
SAMD9	x				
SATB1	x	x			
SATB2	x	x	x	x	
SCAF4	x	x			

SCN11A	x				
SCN1A	x	x	x		
SCN1B	x	x			
SCN2A	x	x	x		
SCN3A	x				
SCN4A	x				
SCN8A	x	x			
SET	x	x	x		
SETBP1	x	x	x		
SETD1A	x	x	x		
SETD1B	x	x			
SETD2	x	x	x		
SETD5	x	x	x	x	
SF3B4	x	x			
SHANK1	x	x			
SHANK2	x	x			
SHANK3	x	x	x	x	
SHH	x	x			
SHOC2	x				
SHROOM3	x	x			
SIK1	x				
SIM1	x	x			
SIN3A	x	x	x		
SIX1	x	x			
SIX3	x	x			
SIX5	x				
SKI	x		x		
SLC1A2	x				
SLC25A24	x				
SLC25A4	x				
SLC2A1	x	x	x		
SLC6A1	x	x	x		
SMAD3	x	x			
SMAD4	x				
SMARCA2	x				
SMARCA4	x	x			
SMARCB1	x	x			
SMARCC2	x	x			
SMARCD1	x				
SMARCE1	x				
SMC3	x				
SNAP25	x				
SNRPB	x	x			
SNRPE	x				
SON	x	x	x		

SOS1	x				
SOX10	x	x	x		
SOX11	x		x		
SOX17	x				
SOX2	x	x			
SOX4	x				
SOX5	x	x	x		
SOX6	x	x			
SOX9	x	x			
SPAST	x				
SPECC1L	x				
SPEN	x	x	x		
SPRED1	x	x			
SPTAN1	x				
SPTBN1	x				
SPTBN2	x				
SRCAP	x		x		
SRP54	x				
SRRM2	x	x	x		
SRSF1	x				
STAG1	x	x			
STX1B	x	x			
STXBP1	x	x	x	x	
SUZ12	x	x			
SYNCRIP	x				
SYNGAP1	x	x	x	x	
SYT1	x				
TAB2	x		x		
TAOK1	x	x	x		
TBL1XR1	x	x	x		
TBR1	x	x			
TBX1	x	x			
TBX18	x	x			
TBX20	x	x			
TBX3	x	x			
TBX4	x	x			
TBX5	x	x			
TCF12	x	x			
TCF20	x	x	x		
TCF4	x	x	x	x	
TCF7L2	x	x	x		
TCOF1	x	x			
TEK	x				
TET3	x	x			
TFAP2A	x				

TFAP2B	x				
TGFB1	x				
TGFB3	x	x			
TGFBR1	x				
TGFBR2	x	x			
TGIF1	x	x			
THRA	x	x			
TINF2	x	x			
TLK2	x	x	x		
TMEM63A	x				
TNRC6B	x	x			
TP63	x	x			
TPM2	x				
TRAF7	x				
TRIM8	x	x			
TRIO	x				
TRIP12	x	x	x		
TRPM3	x				
TRPS1	x	x			
TRPV3	x				
TRPV4	x				
TRRAP	x				
TSC1	x	x			
TSC2	x	x			
TSHR	x				
TUBA1A	x	x			
TUBB	x				
TUBB2A	x				
TUBB2B	x				
TUBB3	x				
TUBB4A	x				
TUBG1	x				
TWIST1	x	x			
TWIST2	x				
U2AF2	x				
UBTF	x				
UPF1	x	x			
USP7	x	x			
VAMP2	x				
VCP	x				
WAC	x	x	x		
WASF1	x	x			
WDFY3	x	x			
WDR11	x				
WDR26	x	x	x		

WDR37	x			
WNT4	x			
WNT5A	x			
WT1	x			
YAP1	x	x		
YWHAG	x			
YY1	x		x	
ZBTB18	x	x	x	
ZBTB20	x			
ZEB2	x	x	x	
ZFHX3	x			
ZFHX4	x	x	x	
ZIC1	x			
ZIC2	x	x		
ZMIZ1	x	x		
ZMYM2	x	x		
ZMYND11	x	x	x	
ZNF148	x	x	x	
ZNF292	x	x	x	
ZNF462	x	x		
ZNF750	x	x		
ZSWIM6	x			

Appendix 7.3.2: List of included CNVs

Chromosome	Start	End	Size	Name	Number of Individuals in UKB	Number of Genes Overlapping
CNV Deletions:						
1	145806438	146149533	0.3	TAR syndrome	71	17
1	147101794	147921262	0.8	1q21.1 deletion	92	7
2	57519361	61509361	4.0	2p15-16.1 microdeletion syndrome	1	11
2	96060525	97010536	1.0	2q11.2 duplication	26	20
2	110625954	112335952	1.7	2q13 duplication	54	9
3	87188160	87508160	0.3	3p11.2 (CHMP2B to POU1F1)	33	2
3	191799517	193299517	1.5	3q28-29 (FGF12)	1	4
3	195988732	197628732	1.6	3q29 duplication	8	24
7	73328061	74727726	1.4	Williams syndrome duplication	1	24
7	75332889	77032747	1.7	Wms-distal deletion	1	21
10	48181951	50630234	2.4	10q11 duplication	52	28
10	79930264	87180263	7.2	10q23.1 deletion	3	30
12	64679953	68249953	3.6	12q14 microdeletion syndrome	2	20
15	22784508	23074431	0.3	15q11.2 deletion	1345	4

15	24573760	28181259	3.6	Prader-Willi/Angelman	1	11
15	30840505	32190507	1.4	15q13.3 deletion	39	6
15	73720606	77840603	4.1	15q24 B to E deletion	1	56
15	84595765	85155765	0.6	15q25.2 deletion	11	8
16	15408642	16198642	0.8	16p13.11 deletion	101	9
16	21931178	22451178	0.5	16p12.1 duplication	243	8
16	28761178	29101178	0.3	16p11.2 distal deletion	48	10
16	29641178	30191178	0.6	NF1 microduplication syndrome	104	31
17	14165958	15595961	1.4	HNPP	220	10
17	30838856	31888868	1.1	16p11.2 duplication	11	12
17	36460073	37846263	1.4	17q12 deletion (HNF1B)	8	15
22	21555711	23307813	1.8	16p11.2-p12.2 microduplication syndrome	6	19
CNV Duplications:						
1	145806438	146149533	0.3	TAR syndrome	320	17
1	147101794	147921262	0.8	1q21.1 duplication	130	7
2	96060525	97010536	1.0	2q11.2 deletion	16	20
2	110625954	112335952	1.7	2q13 deletion	51	9
3	195988732	197628732	1.6	3q29 deletion	2	24
5	10000	11726888	11.7	Cri du Chat syndrome	1	50
7	73328061	74727726	1.4	15q24 A to D duplication	11	24
10	48181951	50630234	2.4	10q11 deletion	27	28
10	79930264	87180263	7.2	10q23.1 duplication	6	30
15	22784508	23074431	0.3	15q11.2 duplication	1616	4
15	24573760	28181259	3.6	Prader-Willi/Angelman	14	11
15	30840505	32190507	1.4	15q13.3 duplication	219	6

15	72670606	75720604	3.0	Williams syndrome deletion	3	49
15	82513967	84070244	1.6	15q25.2 duplication	1	15
15	84595765	85155765	0.6	15q25.2 duplication	11	8
16	15408642	16198642	0.8	16p13.11 duplication	559	9
16	21601178	29031178	7.4	22q11.2 distal deletion	1	69
16	21931178	22451178	0.5	16p12.1 deletion	86	8
16	28761178	29101178	0.3	16p11.2 distal duplication	77	10
16	29641178	30191178	0.6	NF1 microdeletion syndrome	91	31
17	2459956	3019956	0.6	17p13.3 duplication (including PAFAH1B1)	1	4
17	14165958	15595961	1.4	HNPP	111	10
17	16805961	20576095	3.8	Smith-Magenis syndrome duplication	2	52
17	30838856	31888868	1.1	16p11.2 deletion	1	12
17	36460073	37846263	1.4	17q12 duplication (HNF1B)	89	15
22	19032487	20302477	1.3	22q11.2dup/DiGeorge/VCFS duplication	152	28
22	21555711	23307813	1.8	22q11.2 distal duplication	5	19

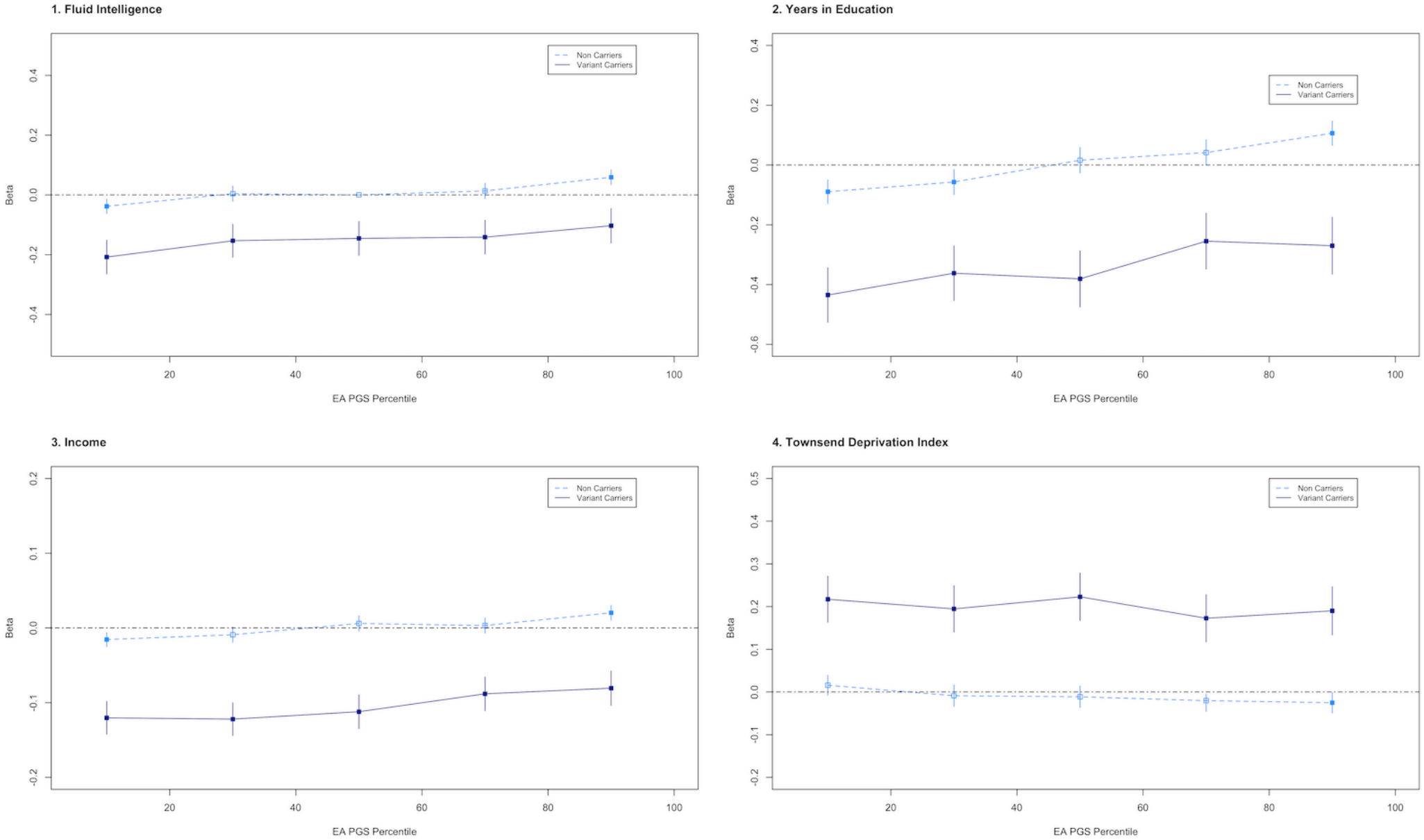
Appendix 7.3.3 (1): Gene panel association tests excluding individuals diagnosed with a childhood developmental disorder

Dataset:	**Individuals with CNV Deletions**				**Individuals with CNV Duplications**				**Individuals with loss of function variants in 599 gene set**			
Phenotype	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**
Binary Traits:	**Odds Ratio:**				**Odds Ratio:**				**Odds Ratio:**			
In employment	0.764	1.908E-08	0.696	0.839	0.822	3.018E-06	0.757	0.892	0.917	2.824E-03	0.867	0.971
Have a degree	0.661	5.170E-25	0.611	0.715	0.748	1.179E-16	0.698	0.801	0.833	1.402E-14	0.795	0.873
Have an epilepsy diagnosis	1.481	0.031	1.037	2.113	1.311	0.113	0.938	1.832	1.407	3.438E-03	1.119	1.770
Diagnosed with Adult DD*	1.386	9.532E-05	1.176	1.633	1.392	8.270E-06	1.204	1.610	1.137	0.022	1.018	1.270
Is unable to work	1.799	6.627E-14	1.543	2.098	1.589	2.201E-10	1.377	1.833	1.329	7.059E-07	1.188	1.487
Continuous Traits:	**Beta:**				**Beta:**				**Beta:**			
Fluid Intelligence	-0.499	7.529E-17	-0.616	-0.382	-0.331	2.931E-10	-0.434	-0.228	-0.158	1.682E-06	-0.223	-0.094
Number of years in education	-0.942	3.403E-23	-1.128	-0.756	-0.745	9.829E-19	-0.911	-0.580	-0.385	1.676E-11	-0.497	-0.273
Income	-0.291	2.143E-36	-0.336	-0.245	-0.218	1.917E-26	-0.258	-0.178	-0.121	1.548E-18	-0.149	-0.094
Reaction time	0.169	5.562E-21	0.134	0.204	0.076	1.789E-06	0.045	0.107	0.041	2.087E-04	0.019	0.062
Pairs test score	0.192	1.919E-03	0.071	0.313	0.306	2.501E-08	0.198	0.413	0.111	3.139E-03	0.037	0.184
Townsend Deprivation Index	0.422	6.151E-14	0.312	0.533	0.468	6.938E-21	0.370	0.566	0.269	9.548E-16	0.203	0.334
Age left education	-0.173	3.355E-04	-0.267	-0.078	-0.211	1.269E-06	-0.296	-0.126	-0.105	5.849E-04	-0.165	-0.045
Height	-1.250	2.627E-25	-1.486	-1.015	-0.621	6.094E-09	-0.831	-0.412	-0.425	5.593E-09	-0.568	-0.282
Reported a mental health issue	0.083	8.682E-05	0.042	0.125	0.029	0.127	-0.008	0.065	0.043	7.923E-04	0.018	0.068
Numeric memory score	-0.210	7.158E-09	-0.281	-0.139	-0.047	0.151	-0.111	0.017	-0.067	1.423E-03	-0.108	-0.026
BMI	0.136	5.683E-13	0.099	0.173	0.110	5.883E-11	0.077	0.143	0.037	1.341E-03	0.014	0.059
Number of children fathered	-0.174	2.067E-07	-0.240	-0.109	-0.095	1.413E-03	-0.154	-0.037	-0.059	5.408E-03	-0.101	-0.018
Number of pregnancies	-0.026	0.543	-0.108	0.057	-0.036	0.337	-0.109	0.037	-0.040	0.102	-0.088	0.008
Number of stillbirths	0.004	0.481	-0.007	0.014	0.008	0.118	-0.002	0.017	0.004	0.209	-0.002	0.010

Appendix 7.3.3 (2): Gene panel association tests excluding individuals diagnosed with a childhood developmental disorder

Dataset:	**Individuals with missense variants in 599 gene set**				**Individuals with synonymous variants in 599 gene set**			
Phenotype	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**	**-**	**P Value**	**Lower 95% CI**	**Upper 95% CI**
Binary Traits:	**Odds Ratio:**				**Odds Ratio:**			
In employment	0.990	0.593	0.956	1.026	1.008	0.514	0.984	1.033
Have a degree	0.926	2.087E-07	0.899	0.953	1.029	5.892E-03	1.008	1.050
Have an epilepsy diagnosis	1.094	0.276	0.931	1.286	0.915	0.142	0.814	1.030
Diagnosed with Adult DD*	1.059	0.119	0.985	1.138	1.013	0.625	0.963	1.066
Is unable to work	1.119	3.954E-03	1.037	1.207	0.989	0.695	0.936	1.045
Continuous Traits:	**Beta:**				**Beta:**			
Fluid Intelligence	-0.086	2.739E-05	-0.127	-0.046	0.003	0.823	-0.025	0.031
Number of years in education	-0.179	6.733E-07	-0.250	-0.109	0.061	0.014	0.013	0.110
Income	-0.057	4.350E-11	-0.074	-0.040	0.011	0.058	0.000	0.023
Reaction time	0.014	0.036	0.001	0.028	-0.006	0.238	-0.015	0.004
Pairs test score	0.057	1.556E-02	0.011	0.103	-0.018	0.260	-0.050	0.014
Townsend Deprivation Index	0.085	6.315E-05	0.043	0.127	0.019	0.205	-0.010	0.047
Age left education	-0.043	0.027	-0.082	-0.005	0.002	0.891	-0.025	0.029
Height	-0.232	4.343E-07	-0.322	-0.142	0.048	0.131	-0.014	0.110
Reported a mental health issue	0.018	0.027	0.002	0.034	0.000	0.962	-0.011	0.011
Numeric memory score	-0.025	0.053	-0.051	0.000	-0.004	0.668	-0.021	0.014
BMI	0.016	0.023	0.002	0.030	-0.002	0.618	-0.012	0.007
Number of children fathered	-0.012	0.389	-0.038	0.015	-0.009	0.312	-0.027	0.009
Number of pregnancies	-0.020	0.191	-0.051	0.010	0.006	0.604	-0.015	0.027
Number of stillbirths	0.004	0.064	-2.126E-04	0.008	0.001	0.373	-0.001	0.004

Appendix Figure 7.4.1: Results from 74 SNP EA-PGS sensitivity analysis



Appendix Table 7.4.2: Results from 74 SNP EA-PGS sensitivity analysis

Association Test Results for EA-PGS Sensitivity Analysis					
Trait	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Fluid Intelligence					
Non-Carriers					
EA-PGS Quintile 1	-0.038	0.012	1.824E-03	-0.062	-0.014
EA-PGS Quintile 2	0.004	0.013	7.573E-01	-0.021	0.030
EA-PGS Quintile 3	.	.	.	.	.
EA-PGS Quintile 4	0.013	0.013	3.124E-01	-0.012	0.039
EA-PGS Quintile 5	0.059	0.013	2.581E-06	0.034	0.084
Variant Carriers					
EA-PGS Quintile 1	-0.208	0.029	5.283E-13	-0.264	-0.151
EA-PGS Quintile 2	-0.153	0.028	6.444E-08	-0.209	-0.098
EA-PGS Quintile 3	-0.145	0.029	5.409E-07	-0.202	-0.089
EA-PGS Quintile 4	-0.141	0.029	9.735E-07	-0.198	-0.085
EA-PGS Quintile 5	-0.103	0.030	4.772E-04	-0.161	-0.045
Years in Education					
Non-Carriers					
EA-PGS Quintile 1	-0.089	0.020	9.177E-06	-0.129	-0.050
EA-PGS Quintile 2	-0.057	0.022	8.334E-03	-0.099	-0.015
EA-PGS Quintile 3	.	.	.	.	.
EA-PGS Quintile 4	0.042	0.022	5.527E-02	-0.001	0.084
EA-PGS Quintile 5	0.107	0.021	3.139E-07	0.066	0.147
Variant Carriers					
EA-PGS Quintile 1	-0.435	0.047	1.046E-20	-0.526	-0.344
EA-PGS Quintile 2	-0.362	0.047	9.606E-15	-0.454	-0.270
EA-PGS Quintile 3	-0.381	0.048	1.898E-15	-0.475	-0.287
EA-PGS Quintile 4	-0.255	0.048	1.027E-07	-0.348	-0.161
EA-PGS Quintile 5	-0.270	0.049	3.116E-08	-0.365	-0.174
Income					
Non-Carriers					
EA-PGS Quintile 1	-0.015	0.005	1.359E-03	-0.025	-0.006
EA-PGS Quintile 2	-0.009	0.005	7.777E-02	-0.019	0.001
EA-PGS Quintile 3	.	.	.	.	.
EA-PGS Quintile 4	0.003	0.005	5.344E-01	-0.007	0.013
EA-PGS Quintile 5	0.020	0.005	4.793E-05	0.011	0.030
Variant Carriers					
EA-PGS Quintile 1	-0.120	0.011	8.374E-27	-0.142	-0.098
EA-PGS Quintile 2	-0.122	0.011	1.436E-27	-0.144	-0.100
EA-PGS Quintile 3	-0.112	0.012	3.653E-22	-0.135	-0.089
EA-PGS Quintile 4	-0.088	0.012	1.841E-14	-0.111	-0.066
EA-PGS Quintile 5	-0.081	0.012	7.389E-12	-0.104	-0.058
TDI					

Non-Carriers					
EA-PGS Quintile 1	0.016	0.012	1.844E-01	-0.008	0.039
EA-PGS Quintile 2	-0.009	0.013	4.915E-01	-0.034	0.016
EA-PGS Quintile 3	.	.	.	.	.
EA-PGS Quintile 4	-0.020	0.013	1.150E-01	-0.045	0.005
EA-PGS Quintile 5	-0.025	0.012	4.142E-02	-0.049	-0.001
Variant Carriers					
EA-PGS Quintile 1	0.217	0.028	3.341E-15	0.163	0.271
EA-PGS Quintile 2	0.195	0.028	2.074E-12	0.140	0.249
EA-PGS Quintile 3	0.223	0.028	3.916E-15	0.167	0.279
EA-PGS Quintile 4	0.173	0.028	1.036E-09	0.117	0.228
EA-PGS Quintile 5	0.190	0.029	4.391E-11	0.134	0.247

Appendix Table 7.4.3: ICD9 and ICD10 codes used to identify related HES information to categorize related clinical diagnoses

Child DD ICD-10 Codes	Description	Adult Neuropsychiatric ICD-10 Codes	Description	Mental Health ICD-10 Codes	Description
F70	Mild mental retardation	F20	Schizophrenia	F40	Phobic anxiety disorders
F71	Moderate mental retardation	F21	Schizotypal	F41	Other anxiety disorders
F72	Severe mental retardation	F22	Persistant Delusional Disorders	F42	OCD
F73	Profound mental retardation	F23	Acute and transient psychotic disorders	F43	Reaction to severe stress
G403		F24	Induced delusional disorder	F44	Dissociative disorders
F80	Developmental disorders of speech and language	F25	Schizoaffective disorders	F45	Somatoform disorders
F81	Developmental disorders of scholastic skills	F26		F48	Other neurotic behaviours
F82	Specific developmental disorder of motor function	F27		F50	Eating Disorders
F83	Mixed specific developmental disorders	F28	Other nonorganic psychotic disorders	F51	Nonorganic sleep disorders
F84	Pervasive developmental disorders	F29	Unspecified nonorganic disorders	F53	Mental and behavioural disorders associated w/ the puerperium
Q00-99	Congenital malformations	F30	Manic Episode	F54	Psychological and behavioural factors associated with disorders elsewhere
F78	Other mental retardation				
F79	Unspecified mental retardation	F31	Bipolar affective disorder	F99	Mental disorder, not otherwise specified
F84	Autism	F32	Depressive episode	G47	Sleep Disorders
F88	Other disorders of psychological development	F33	Recurrent depressive disorder	R45	Emotional State
F89	Unspecified disorder of psychological development	F34	Persistent mood disorder	**Mental Health ICD-9 Codes**	
F90	Hyperkinetic Disorders	F35		300	Neurotic disorders
F91	Conduct Disorders	F36		307	anorexia, etc
F92	Mixed disorders of conduct and emotions	F37		308	stress reactions
F93	Emotional disorders, childhood onset	F38	Other mood affective disorders	309	adjustment issues
F94	Social functioning, childhood onset	F39	Unspecified mood affective disorders	311	depressive disorder
F95	Tic disorders	G40	Epilepsy	780.5	sleep disturbance

F98	Other behavioural and emotional disorders, childhood onset
R62	Lack of expected normal physiological development
R48	Dyslexia
Z55	Problems related to education and literacy
F90	ADHD

Child DD ICD-9 Codes	
299	Childhood psychoses
317	mild mental retardation
318	Other mental retardation
319	Unspecified mental retardation
740-759	Congenital malformations
312	disturbance of conduct
313	disturbance of emotions in childhood
314	Hyperkinetic Disorders
315	specific delays in development
314	hyperkinetic
299.0	infantile autism

Adult Neuropsychiatric ICD-9 Codes	
295	schizophrenia
296	Manic depressive
1289	
345	epilepsy

Appendix 7.4.4: Continuous association results for rare variant carriers including and excluding missense variants: Rare variant burden association tests for individuals grouped by number of rare variants identified in their exome sequencing, for both individuals with any type of variant in these genes, and limited to those with a CNV deletion, duplication or LoF variant in any of these genes.

Rare Variant Burden: Association Test Results for individuals with any variant in the DDG2P gene set (n= 54,445): Continuous Trait Results					
Trait	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Fluid Intelligence					
One Variant Only	-0.133	0.014	2.377E-20	-0.161	-0.104
Two Variants	-0.229	0.048	2.291E-06	-0.324	-0.134
Three + Variants	-0.541	0.214	1.142E-02	-0.959	-0.122
Age Left Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.088	0.013	2.156E-12	-0.113	-0.064
Two Variants	-0.145	0.043	8.177E-04	-0.230	-0.060
Three + Variants	-0.581	0.183	1.479E-03	-0.939	-0.223
Years in Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.307	0.024	1.170E-38	-0.354	-0.261
Two Variants	-0.409	0.081	4.205E-07	-0.567	-0.250
Three + Variants	-0.377	0.336	2.612E-01	-1.036	0.281
Income	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.091	0.006	1.213E-57	-0.102	-0.080
Two Variants	-0.153	0.020	5.627E-15	-0.191	-0.115
Three + Variants	-0.207	0.083	1.244E-02	-0.368	-0.045
Townsend Deprivation Index	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.172	0.014	1.151E-34	0.145	0.199
Two Variants	0.423	0.048	7.407E-19	0.329	0.516
Three + Variants	0.359	0.198	7.010E-02	-0.029	0.747

Rare Variant Burden: Association Test Results for individuals with any LoF variant or CNV in the DDG2P gene set (n= 16,934): Continuous Trait Results					
Trait	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Fluid Intelligence					
One Variant Only	-0.216	0.024	4.230E-19	-0.263	-0.168
Two Variants	-0.151	0.138	2.742E-01	-0.423	0.120
Three Variants	-2.943	0.851	5.448E-04	-4.611	-1.275
Age Left Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.127	0.021	7.773E-10	-0.167	-0.086
Two Variants	-0.043	0.119	7.172E-01	-0.277	0.191
Three + Variants	-0.890	0.728	2.214E-01	-2.316	0.536
Years in Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.506	0.039	6.742E-38	-0.584	-0.429
Two Variants	-0.374	0.224	9.511E-02	-0.812	0.065
Three Variants	0.278	1.320	8.333E-01	-2.310	2.866
Income	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.152	0.009	2.049E-57	-0.170	-0.133
Two Variants	-0.199	0.055	2.727E-04	-0.306	-0.092
Three Variants	0.030	0.350	9.324E-01	-0.655	0.715
Townsend Deprivation Index	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.316	0.023	6.733E-42	0.271	0.362
Two Variants	0.625	0.133	2.383E-06	0.366	0.885
Three Variants	0.788	0.784	3.152E-01	-0.750	2.325

Numeric Memory	Beta	Standard Error	P Value	95% CI I	95% CI II
One Variant Only	-0.035	0.009	5.396E-05	-0.052	-0.018
Two Variants	-0.039	0.029	1.809E-01	-0.097	0.018
Three + Variants	-0.247	0.125	4.825E-02	-0.493	-0.002
Reaction Time	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.036	0.004	1.118E-15	0.027	0.045
Two Variants	0.073	0.015	1.617E-06	0.043	0.103
Three + Variants	0.026	0.063	6.839E-01	-0.098	0.150
Time taken on Pairs Test	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.106	0.015	7.060E-12	0.076	0.136
Two Variants	0.170	0.053	1.178E-03	0.067	0.273
Three + Variants	0.538	0.218	1.374E-02	0.110	0.966
Height	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.325	0.030	2.361E-27	-0.384	-0.267
Two Variants	-0.520	0.102	3.784E-07	-0.720	-0.319
Three + Variants	-1.167	0.425	5.993E-03	-1.999	-0.335

Numeric Memory	Beta	Standard Error	P Value	95% CI I	95% CI II
One Variant Only	-0.056	0.015	1.363E-04	-0.085	-0.027
Two Variants	-0.104	0.084	2.150E-01	-0.268	0.060
Three Variants	-0.015	0.569	9.783E-01	-1.131	1.100
Reaction Time	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.063	0.007	4.050E-17	0.048	0.077
Two Variants	0.132	0.042	1.791E-03	0.049	0.215
Three Variants	-0.043	0.250	8.634E-01	-0.532	0.446
Time taken on Pairs Test	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.183	0.026	9.366E-13	0.133	0.233
Two Variants	0.476	0.146	1.081E-03	0.191	0.762
Three Variants	0.088	0.863	9.188E-01	-1.604	1.780
Height	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	-0.532	0.050	1.436E-26	-0.630	-0.435
Two Variants	-0.810	0.283	4.286E-03	-1.365	-0.254
Three Variants	-3.414	1.679	4.203E-02	-6.705	-0.123

Appendix table 7.4.5: Binary association results for rare variant carriers including and excluding missense variants

Rare Variant Burden: Association Test Results for individuals with any variant in the DDG2P gene set (n= 54,445): Binary Trait Results					
Trait	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Unable To Work					
One Variant Only	1.221	0.029	1.687E-17	1.166	1.278
Two Variants	1.299	0.101	7.526E-04	1.116	1.512
Three + Variants	1.414	0.440	2.651E-01	0.769	2.602
In Employment	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.959	0.011	4.363E-04	0.937	0.982
Two Variants	0.851	0.034	6.280E-05	0.786	0.921
Three + Variants	0.856	0.146	3.626E-01	0.613	1.196
Has a Degree	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.890	0.009	3.373E-33	0.873	0.907
Two Variants	0.841	0.028	2.133E-07	0.788	0.898
Three + Variants	0.878	0.121	3.451E-01	0.669	1.151
Has a Child DD Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.277	0.042	1.084E-13	1.197	1.362
Two Variants	1.662	0.163	2.250E-07	1.371	2.014
Three + Variants	2.139	0.771	3.495E-02	1.055	4.336
Has an Adult DD Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.125	0.026	2.824E-07	1.075	1.177
Two Variants	1.299	0.095	3.297E-04	1.126	1.499
Three + Variants	1.725	0.463	4.202E-02	1.020	2.917
Has a Mental Health Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.107	0.021	5.860E-08	1.067	1.149
Two Variants	1.149	0.072	2.712E-02	1.016	1.300

Rare Variant Burden: Association Test Results for individuals with any LoF variant or CNV in the DDG2P gene set (n= 16,934): Binary Trait Results					
Trait	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Unable To Work					
One Variant Only	1.413	0.051	1.123E-21	1.316	1.517
Two Variants	1.161	0.260	5.063E-01	0.748	1.800
Three Variants	1.000	.	.	.	.
In Employment	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.896	0.018	2.729E-08	0.862	0.932
Two Variants	0.827	0.093	9.216E-02	0.663	1.032
Three Variants	3.931	2.621	4.001E-02	1.064	14.520
Has a Degree	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	0.811	0.013	5.463E-38	0.786	0.837
Two Variants	0.845	0.078	6.676E-02	0.705	1.012
Three Variants	0.935	0.511	9.026E-01	0.320	2.731
Has a Child DD Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.664	0.080	3.218E-26	1.514	1.828
Two Variants	2.138	0.514	1.584E-03	1.334	3.425
Three Variants	4.344	4.509	1.571E-01	0.568	33.224
Has an Adult DD Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.218	0.045	7.212E-08	1.134	1.309
Two Variants	1.248	0.257	2.809E-01	0.834	1.868
Three Variants	3.774	2.889	8.271E-02	0.842	16.917
Has a Mental Health Related Diagnosis	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.170	0.036	2.221E-07	1.103	1.242
Two Variants	1.387	0.224	4.300E-02	1.010	1.904

Three + Variants	1.255	0.317	3.696E-01	0.764	2.060
Never a Parent	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.107	0.014	2.000E-16	1.080	1.134
Two Variants	1.319	0.053	3.915E-12	1.220	1.427
Three + Variants	1.452	0.234	2.088E-02	1.058	1.992
Never Pregnant	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.071	0.019	1.423E-04	1.034	1.110
Two Variants	1.286	0.075	1.463E-05	1.148	1.441
Three + Variants	1.436	0.333	1.182E-01	0.912	2.262
Never a Father	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.139	0.019	1.469E-14	1.102	1.178
Two Variants	1.351	0.075	5.121E-08	1.212	1.505
Three + Variants	1.455	0.328	9.660E-02	0.935	2.264

Three Variants	2.319	1.774	2.714E-01	0.518	10.386
Never a Parent	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.183	0.024	7.203E-17	1.137	1.230
Two Variants	1.196	0.136	1.149E-01	0.957	1.493
Three Variants	3.165	1.850	4.873E-02	1.006	9.951
Never Pregnant	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.128	0.034	5.245E-05	1.064	1.196
Two Variants	1.064	0.181	7.146E-01	0.763	1.484
Three Variants	4.119	3.215	6.972E-02	0.892	19.020
Never a Father	**Odds Ratio**	**SE**	**P Value**	**95% CI I**	**95% CI II**
One Variant Only	1.232	0.034	2.447E-14	1.168	1.300
Two Variants	1.328	0.204	6.466E-02	0.983	1.796
Three Variants	2.210	1.976	3.752E-01	0.383	12.749

Appendix table 7.4.6: Numbers of individuals in each rare variant carrier group excluding missense variants

Overall: LoF and CNV Carriers	Number of Individuals
One Variant:	**16,429**
EA-PGS Quintile 1	3461
EA-PGS Quintile 2	3318
EA-PGS Quintile 3	3227
EA-PGS Quintile 4	3122
EA-PGS Quintile 5	3301
Two Variants:	**491**
EA-PGS Quintile 1	100
EA-PGS Quintile 2	97
EA-PGS Quintile 3	93
EA-PGS Quintile 4	105
EA-PGS Quintile 5	96
Three Variants:	**14**
EA-PGS Quintile 1	4
EA-PGS Quintile 2	2
EA-PGS Quintile 3	0
EA-PGS Quintile 4	4
EA-PGS Quintile 5	4

Appendix table 7.4.7 (1): EA-PGS and rare variant association results across quintiles excluding missense variants: Continuous Results

Association Test Results for individuals with LoF variants and CNVs only (n= 16,934): Continuous Trait Results					
Trait	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Fluid Intelligence					
One Variant					
EA-PGS Quintile 1	-0.846	0.055	6.868E-54	-0.953	-0.739
EA-PGS Quintile 2	-0.382	0.055	2.683E-12	-0.489	-0.275
EA-PGS Quintile 3	-0.220	0.054	4.032E-05	-0.325	-0.115
EA-PGS Quintile 4	-0.017	0.054	7.583E-01	-0.123	0.090
EA-PGS Quintile 5	0.377	0.051	1.285E-13	0.277	0.477
Two Variants					
EA-PGS Quintile 1	-0.734	0.352	3.711E-02	-1.424	-0.044
EA-PGS Quintile 2	-0.160	0.293	5.867E-01	-0.735	0.416
EA-PGS Quintile 3	-0.203	0.300	4.980E-01	-0.790	0.384
EA-PGS Quintile 4	0.131	0.288	6.485E-01	-0.433	0.695
EA-PGS Quintile 5	0.053	0.303	8.618E-01	-0.541	0.646
Three Variants					
EA-PGS Quintile 1	-4.238	2.052	3.893E-02	-8.261	-0.215
EA-PGS Quintile 2	0.000	.	.	.	.
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	-1.402	1.185	2.366E-01	-3.724	0.920
EA-PGS Quintile 5	-4.639	1.451	1.392E-03	-7.484	-1.795
Age Left Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	-0.543	0.042	7.641E-39	-0.625	-0.461
EA-PGS Quintile 2	-0.249	0.044	2.034E-08	-0.336	-0.162
EA-PGS Quintile 3	-0.150	0.046	1.237E-03	-0.241	-0.059
EA-PGS Quintile 4	0.140	0.049	4.671E-03	0.043	0.236
EA-PGS Quintile 5	0.232	0.053	1.137E-05	0.128	0.335
Two Variants					
EA-PGS Quintile 1	-0.432	0.267	1.058E-01	-0.955	0.091
EA-PGS Quintile 2	-0.359	0.246	1.444E-01	-0.840	0.123
EA-PGS Quintile 3	-0.065	0.263	8.050E-01	-0.580	0.450
EA-PGS Quintile 4	0.484	0.271	7.417E-02	-0.047	1.015
EA-PGS Quintile 5	0.081	0.297	7.846E-01	-0.501	0.664
Three Variants					
EA-PGS Quintile 1	-1.292	1.544	4.025E-01	-4.318	1.733
EA-PGS Quintile 2	-1.266	1.544	4.123E-01	-4.291	1.760
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	-0.578	1.260	6.463E-01	-3.049	1.892
EA-PGS Quintile 5	-0.727	1.544	6.379E-01	-3.753	2.300
Years in Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					

EA-PGS Quintile 1	-2.084	0.086	4.550E-130	-2.252	-1.916
EA-PGS Quintile 2	-1.088	0.087	9.398E-36	-1.258	-0.917
EA-PGS Quintile 3	-0.542	0.088	8.539E-10	-0.715	-0.369
EA-PGS Quintile 4	0.152	0.089	9.012E-02	-0.024	0.327
EA-PGS Quintile 5	1.245	0.087	2.411E-46	1.074	1.415
Two Variants					
EA-PGS Quintile 1	-1.160	0.493	1.864E-02	-2.126	-0.194
EA-PGS Quintile 2	-1.766	0.495	3.634E-04	-2.737	-0.795
EA-PGS Quintile 3	-0.649	0.506	1.999E-01	-1.640	0.343
EA-PGS Quintile 4	0.556	0.476	2.431E-01	-0.377	1.489
EA-PGS Quintile 5	1.272	0.503	1.150E-02	0.285	2.258
Three Variants					
EA-PGS Quintile 1	-0.313	2.438	8.977E-01	-5.092	4.465
EA-PGS Quintile 2	-6.537	3.448	5.796E-02	-13.295	0.221
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	0.370	2.438	8.794E-01	-4.408	5.148
EA-PGS Quintile 5	4.231	2.438	8.265E-02	-0.547	9.010
Income	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	-0.403	0.021	3.375E-82	-0.444	-0.362
EA-PGS Quintile 2	-0.234	0.021	1.287E-28	-0.275	-0.192
EA-PGS Quintile 3	-0.146	0.021	6.791E-12	-0.187	-0.104
EA-PGS Quintile 4	-0.050	0.021	1.980E-02	-0.092	-0.008
EA-PGS Quintile 5	0.118	0.021	1.879E-08	0.077	0.159
Two Variants					
EA-PGS Quintile 1	-0.429	0.119	3.164E-04	-0.663	-0.196
EA-PGS Quintile 2	-0.278	0.124	2.565E-02	-0.522	-0.034
EA-PGS Quintile 3	-0.117	0.121	3.346E-01	-0.355	0.121
EA-PGS Quintile 4	-0.175	0.115	1.293E-01	-0.401	0.051
EA-PGS Quintile 5	0.072	0.125	5.651E-01	-0.174	0.318
Three Variants					
EA-PGS Quintile 1	0.127	0.772	8.691E-01	-1.386	1.640
EA-PGS Quintile 2	-1.179	1.092	2.803E-01	-3.319	0.961
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	0.024	0.546	9.646E-01	-1.046	1.094
EA-PGS Quintile 5	0.381	0.630	5.454E-01	-0.854	1.617
Townsend Deprivation Index	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.771	0.051	9.801E-52	0.671	0.871
EA-PGS Quintile 2	0.483	0.052	1.059E-20	0.382	0.585
EA-PGS Quintile 3	0.364	0.052	3.620E-12	0.262	0.467
EA-PGS Quintile 4	0.177	0.053	9.122E-04	0.072	0.281
EA-PGS Quintile 5	0.037	0.052	4.740E-01	-0.064	0.139
Two Variants					
EA-PGS Quintile 1	1.001	0.291	5.708E-04	0.432	1.571
EA-PGS Quintile 2	0.812	0.295	5.913E-03	0.234	1.391

EA-PGS Quintile 3	1.290	0.301	1.880E-05	0.699	1.881
EA-PGS Quintile 4	0.449	0.284	1.134E-01	-0.107	1.005
EA-PGS Quintile 5	-0.081	0.298	7.861E-01	-0.665	0.503
Three Variants					
EA-PGS Quintile 1	0.002	1.452	9.989E-01	-2.844	2.848
EA-PGS Quintile 2	6.233	2.054	2.404E-03	2.208	10.258
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	-0.081	1.452	9.555E-01	-2.927	2.765
EA-PGS Quintile 5	0.017	1.452	9.908E-01	-2.829	2.863
Numeric Memory	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	-0.185	0.035	1.018E-07	-0.253	-0.117
EA-PGS Quintile 2	-0.122	0.034	3.179E-04	-0.188	-0.055
EA-PGS Quintile 3	-0.106	0.033	1.527E-03	-0.171	-0.040
EA-PGS Quintile 4	0.009	0.033	7.885E-01	-0.056	0.073
EA-PGS Quintile 5	0.107	0.030	4.475E-04	0.047	0.167
Two Variants					
EA-PGS Quintile 1	-0.058	0.226	7.966E-01	-0.501	0.384
EA-PGS Quintile 2	-0.223	0.183	2.237E-01	-0.581	0.136
EA-PGS Quintile 3	-0.099	0.189	6.018E-01	-0.470	0.272
EA-PGS Quintile 4	-0.071	0.169	6.752E-01	-0.402	0.260
EA-PGS Quintile 5	-0.043	0.180	8.124E-01	-0.395	0.310
Three Variants					
EA-PGS Quintile 1	-0.703	0.984	4.748E-01	-2.632	1.225
EA-PGS Quintile 2	0.000	.	.	.	.
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	-1.933	0.984	4.948E-02	-3.861	-0.004
EA-PGS Quintile 5	2.602	0.984	8.160E-03	0.674	4.531
Reaction Time	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.093	0.016	1.397E-08	0.061	0.125
EA-PGS Quintile 2	0.059	0.017	4.413E-04	0.026	0.091
EA-PGS Quintile 3	0.058	0.017	5.915E-04	0.025	0.091
EA-PGS Quintile 4	0.085	0.017	6.478E-07	0.052	0.119
EA-PGS Quintile 5	0.048	0.017	4.344E-03	0.015	0.080
Two Variants					
EA-PGS Quintile 1	0.327	0.094	4.885E-04	0.143	0.511
EA-PGS Quintile 2	0.145	0.096	1.291E-01	-0.042	0.333
EA-PGS Quintile 3	0.192	0.097	4.705E-02	0.003	0.382
EA-PGS Quintile 4	-0.062	0.091	4.949E-01	-0.241	0.116
EA-PGS Quintile 5	0.087	0.096	3.624E-01	-0.100	0.275
Three Variants					
EA-PGS Quintile 1	0.140	0.466	7.638E-01	-0.774	1.054
EA-PGS Quintile 2	-1.011	0.659	1.251E-01	-2.303	0.281
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	-0.013	0.466	9.783E-01	-0.926	0.901

EA-PGS Quintile 5	0.234	0.466	6.158E-01	-0.680	1.148
Time taken on Pairs Test	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.221	0.057	9.342E-05	0.110	0.332
EA-PGS Quintile 2	0.218	0.058	1.597E-04	0.105	0.331
EA-PGS Quintile 3	0.211	0.058	3.057E-04	0.097	0.326
EA-PGS Quintile 4	0.121	0.059	4.233E-02	0.004	0.237
EA-PGS Quintile 5	0.121	0.058	3.609E-02	0.008	0.235
Two Variants					
EA-PGS Quintile 1	0.226	0.325	4.858E-01	-0.410	0.863
EA-PGS Quintile 2	0.367	0.330	2.660E-01	-0.279	1.013
EA-PGS Quintile 3	1.054	0.337	1.762E-03	0.393	1.714
EA-PGS Quintile 4	0.185	0.317	5.591E-01	-0.436	0.806
EA-PGS Quintile 5	0.596	0.332	7.218E-02	-0.054	1.246
Three Variants					
EA-PGS Quintile 1	1.941	1.622	2.316E-01	-1.239	5.120
EA-PGS Quintile 2	-3.319	2.294	1.480E-01	-7.815	1.178
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	0.697	1.622	6.674E-01	-2.482	3.876
EA-PGS Quintile 5	-0.660	1.622	6.841E-01	-3.840	2.520
Height	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	-1.374	0.110	6.389E-36	-1.589	-1.159
EA-PGS Quintile 2	-0.706	0.112	2.824E-10	-0.925	-0.487
EA-PGS Quintile 3	-0.551	0.113	1.155E-06	-0.773	-0.329
EA-PGS Quintile 4	-0.274	0.115	1.748E-02	-0.499	-0.048
EA-PGS Quintile 5	0.354	0.112	1.581E-03	0.135	0.574
Two Variants					
EA-PGS Quintile 1	-0.928	0.628	1.396E-01	-2.160	0.303
EA-PGS Quintile 2	-0.768	0.638	2.284E-01	-2.018	0.482
EA-PGS Quintile 3	-1.247	0.651	5.565E-02	-2.523	0.030
EA-PGS Quintile 4	-0.357	0.613	5.602E-01	-1.559	0.845
EA-PGS Quintile 5	-0.730	0.641	2.552E-01	-1.987	0.527
Three Variants					
EA-PGS Quintile 1	-2.690	3.139	3.914E-01	-8.843	3.462
EA-PGS Quintile 2	-12.299	4.439	5.595E-03	-21.000	-3.599
EA-PGS Quintile 3	0.000	.	.	.	.
EA-PGS Quintile 4	0.981	3.139	7.546E-01	-5.171	7.133
EA-PGS Quintile 5	-4.076	3.139	1.941E-01	-10.228	2.076

Appendix table 7.4.7 (2): EA-PGS and rare variant association results across quintiles excluding missense variants: Binary Results

Association Test Results for individuals with LoF variants and CNVs only (n= 16,934): Binary Trait Results					
Trait	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
Unable To Work					
One Variant					
EA-PGS Quintile 1	2.535	0.166	1.333E-45	2.229	2.883
EA-PGS Quintile 2	1.487	0.123	1.457E-06	1.265	1.748
EA-PGS Quintile 3	1.370	0.118	2.484E-04	1.158	1.622
EA-PGS Quintile 4	1.357	0.119	5.220E-04	1.142	1.612
EA-PGS Quintile 5	0.901	0.093	3.109E-01	0.736	1.103
Two Variants					
EA-PGS Quintile 1	0.255	0.257	1.746E-01	0.035	1.834
EA-PGS Quintile 2	2.248	0.890	4.076E-02	1.035	4.886
EA-PGS Quintile 3	1.817	0.773	1.601E-01	0.790	4.182
EA-PGS Quintile 4	1.501	0.691	3.777E-01	0.609	3.702
EA-PGS Quintile 5	0.670	0.480	5.768E-01	0.165	2.731
Three Variants					
EA-PGS Quintile 1	1.000	.	.	.	.
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	1.000	.	.	.	.
EA-PGS Quintile 5	1.000	.	.	.	.
In Employment	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.712	0.031	9.074E-15	0.654	0.776
EA-PGS Quintile 2	0.950	0.042	2.487E-01	0.871	1.037
EA-PGS Quintile 3	0.959	0.043	3.550E-01	0.879	1.048
EA-PGS Quintile 4	0.855	0.039	5.701E-04	0.782	0.935
EA-PGS Quintile 5	1.043	0.047	3.451E-01	0.956	1.139
Two Variants					
EA-PGS Quintile 1	0.697	0.176	1.523E-01	0.425	1.143
EA-PGS Quintile 2	0.796	0.201	3.653E-01	0.485	1.305
EA-PGS Quintile 3	0.873	0.236	6.163E-01	0.514	1.484
EA-PGS Quintile 4	0.676	0.163	1.049E-01	0.422	1.085
EA-PGS Quintile 5	1.203	0.303	4.645E-01	0.733	1.972
Three Variants					
EA-PGS Quintile 1	4.299	5.516	2.558E-01	0.348	53.165
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	1.000	.	.	.	.
EA-PGS Quintile 5	5.749	7.127	1.583E-01	0.506	65.298
Has a Degree	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.427	0.017	3.230E-104	0.395	0.461

EA-PGS Quintile 2	0.641	0.024	1.319E-32	0.595	0.690
EA-PGS Quintile 3	0.771	0.029	2.963E-12	0.717	0.830
EA-PGS Quintile 4	1.062	0.039	1.050E-01	0.988	1.142
EA-PGS Quintile 5	1.711	0.063	7.561E-48	1.591	1.839
Two Variants					
EA-PGS Quintile 1	0.549	0.119	5.456E-03	0.359	0.838
EA-PGS Quintile 2	0.728	0.153	1.302E-01	0.483	1.098
EA-PGS Quintile 3	0.681	0.148	7.655E-02	0.446	1.042
EA-PGS Quintile 4	1.114	0.220	5.835E-01	0.757	1.641
EA-PGS Quintile 5	1.531	0.324	4.409E-02	1.011	2.317
Three Variants					
EA-PGS Quintile 1	1.224	1.237	8.417E-01	0.169	8.873
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	0.381	0.446	4.093E-01	0.038	3.775
EA-PGS Quintile 5	3.987	4.650	2.357E-01	0.405	39.210
Has a Child DD Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	2.001	0.199	3.036E-12	1.647	2.432
EA-PGS Quintile 2	1.937	0.200	1.382E-10	1.583	2.370
EA-PGS Quintile 3	1.846	0.197	9.166E-09	1.498	2.276
EA-PGS Quintile 4	1.533	0.179	2.612E-04	1.219	1.929
EA-PGS Quintile 5	1.353	0.163	1.222E-02	1.068	1.713
Two Variants					
EA-PGS Quintile 1	3.151	1.451	1.268E-02	1.278	7.771
EA-PGS Quintile 2	1.238	0.886	7.652E-01	0.305	5.034
EA-PGS Quintile 3	2.573	1.319	6.520E-02	0.942	7.026
EA-PGS Quintile 4	1.695	0.995	3.688E-01	0.536	5.355
EA-PGS Quintile 5	1.852	1.089	2.947E-01	0.585	5.863
Three Variants					
EA-PGS Quintile 1	1.000	.	.	.	.
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	1.000	.	.	.	.
EA-PGS Quintile 5	18.906	21.934	1.129E-02	1.946	183.718
Has an Adult DD Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	1.701	0.122	1.437E-13	1.478	1.958
EA-PGS Quintile 2	1.396	0.111	2.813E-05	1.194	1.631
EA-PGS Quintile 3	1.215	0.104	2.278E-02	1.027	1.436
EA-PGS Quintile 4	1.092	0.100	3.352E-01	0.913	1.305
EA-PGS Quintile 5	1.008	0.092	9.324E-01	0.842	1.206
Two Variants					
EA-PGS Quintile 1	0.732	0.430	5.947E-01	0.231	2.313
EA-PGS Quintile 2	1.044	0.535	9.330E-01	0.383	2.848
EA-PGS Quintile 3	1.674	0.709	2.244E-01	0.729	3.841

EA-PGS Quintile 4	1.527	0.645	3.157E-01	0.668	3.493
EA-PGS Quintile 5	1.683	0.712	2.186E-01	0.734	3.858
Three Variants					
EA-PGS Quintile 1	1.000	.	.	.	.
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	18.433	18.511	3.709E-03	2.575	131.943
EA-PGS Quintile 5	1.000	.	.	.	.
Has a Mental Health Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	1.624	0.096	3.174E-16	1.446	1.825
EA-PGS Quintile 2	1.244	0.083	1.088E-03	1.091	1.419
EA-PGS Quintile 3	1.138	0.080	6.679E-02	0.991	1.307
EA-PGS Quintile 4	0.963	0.074	6.229E-01	0.828	1.120
EA-PGS Quintile 5	0.989	0.073	8.781E-01	0.855	1.144
Two Variants					
EA-PGS Quintile 1	0.635	0.324	3.738E-01	0.233	1.728
EA-PGS Quintile 2	1.948	0.626	3.799E-02	1.038	3.656
EA-PGS Quintile 3	1.385	0.514	3.796E-01	0.670	2.865
EA-PGS Quintile 4	1.454	0.508	2.843E-01	0.733	2.886
EA-PGS Quintile 5	1.771	0.594	8.828E-02	0.918	3.417
Three Variants					
EA-PGS Quintile 1	1.000	.	.	.	.
EA-PGS Quintile 2	13.646	19.329	6.502E-02	0.850	219.114
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	4.574	5.287	1.884E-01	0.475	44.074
EA-PGS Quintile 5	1.000	.	.	.	.
Never a Parent	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	0.996	0.047	9.394E-01	0.909	1.092
EA-PGS Quintile 2	1.106	0.051	2.907E-02	1.010	1.211
EA-PGS Quintile 3	1.280	0.058	4.285E-08	1.172	1.398
EA-PGS Quintile 4	1.257	0.058	6.577E-07	1.149	1.376
EA-PGS Quintile 5	1.400	0.061	1.118E-14	1.286	1.525
Two Variants					
EA-PGS Quintile 1	1.191	0.297	4.846E-01	0.730	1.941
EA-PGS Quintile 2	1.257	0.320	3.699E-01	0.763	2.071
EA-PGS Quintile 3	1.494	0.366	1.008E-01	0.925	2.413
EA-PGS Quintile 4	1.063	0.273	8.114E-01	0.643	1.757
EA-PGS Quintile 5	1.049	0.285	8.610E-01	0.616	1.785
Three Variants					
EA-PGS Quintile 1	9.782	12.150	6.634E-02	0.857	111.596
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	4.435	4.626	1.533E-01	0.574	34.255
EA-PGS Quintile 5	2.036	2.391	5.448E-01	0.204	20.336

Never Pregnant	Odds Ratio	Standard Error	P Value	95% CI I	95% CI II
One Variant					
EA-PGS Quintile 1	0.848	0.063	2.545E-02	0.733	0.980
EA-PGS Quintile 2	1.060	0.073	3.944E-01	0.926	1.214
EA-PGS Quintile 3	1.261	0.084	4.838E-04	1.107	1.437
EA-PGS Quintile 4	1.239	0.083	1.455E-03	1.086	1.414
EA-PGS Quintile 5	1.312	0.082	1.425E-05	1.161	1.483
Two Variants					
EA-PGS Quintile 1	0.453	0.237	1.297E-01	0.162	1.262
EA-PGS Quintile 2	1.128	0.415	7.425E-01	0.549	2.320
EA-PGS Quintile 3	1.959	0.668	4.845E-02	1.005	3.822
EA-PGS Quintile 4	0.778	0.316	5.361E-01	0.351	1.724
EA-PGS Quintile 5	1.386	0.491	3.567E-01	0.692	2.773
Three Variants					
EA-PGS Quintile 1	3.996	5.751	3.357E-01	0.238	67.077
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	7.616	10.782	1.515E-01	0.475	122.108
EA-PGS Quintile 5	3.197	3.997	3.526E-01	0.276	37.060
Never a Father	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
One Variant					
EA-PGS Quintile 1	1.122	0.068	5.841E-02	0.996	1.265
EA-PGS Quintile 2	1.141	0.071	3.570E-02	1.009	1.290
EA-PGS Quintile 3	1.303	0.080	1.635E-05	1.155	1.470
EA-PGS Quintile 4	1.280	0.081	9.529E-05	1.131	1.448
EA-PGS Quintile 5	1.497	0.092	4.028E-11	1.328	1.688
Two Variants					
EA-PGS Quintile 1	2.073	0.638	1.786E-02	1.134	3.788
EA-PGS Quintile 2	1.380	0.494	3.675E-01	0.685	2.782
EA-PGS Quintile 3	1.145	0.401	6.983E-01	0.577	2.274
EA-PGS Quintile 4	1.377	0.466	3.440E-01	0.710	2.672
EA-PGS Quintile 5	0.773	0.324	5.391E-01	0.340	1.759
Three Variants					
EA-PGS Quintile 1	1.000	.	.	.	.
EA-PGS Quintile 2	1.000	.	.	.	.
EA-PGS Quintile 3	1.000	.	.	.	.
EA-PGS Quintile 4	2.431	3.458	5.325E-01	0.150	39.515
EA-PGS Quintile 5	1.000	.	.	.	.

Appendix table 7.4.8 (1): EA-PGS and rare variant association results across quintiles within the 325 gene subset: Continuous Results

Individuals with any LoF variants in the 325 Gene Set: Continuous Trait Results (n = 5776)					
Fluid Intelligence	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.801	0.093	9.055E-18	-0.984	-0.618
EA-PGS Quintile 2	-0.434	0.091	1.973E-06	-0.613	-0.255
EA-PGS Quintile 3	-0.215	0.090	1.648E-02	-0.391	-0.039
EA-PGS Quintile 4	-0.055	0.091	5.443E-01	-0.233	0.123
EA-PGS Quintile 5	0.202	0.085	1.791E-02	0.035	0.368
Age Left Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.575	0.070	1.873E-16	-0.712	-0.438
EA-PGS Quintile 2	-0.258	0.073	3.688E-04	-0.401	-0.116
EA-PGS Quintile 3	-0.159	0.078	4.053E-02	-0.311	-0.007
EA-PGS Quintile 4	0.116	0.081	1.502E-01	-0.042	0.275
EA-PGS Quintile 5	0.190	0.085	2.562E-02	0.023	0.358
Years in Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-2.382	0.144	2.050E-61	-2.664	-2.100
EA-PGS Quintile 2	-1.308	0.144	9.713E-20	-1.590	-1.026
EA-PGS Quintile 3	-0.710	0.148	1.578E-06	-1.000	-0.420
EA-PGS Quintile 4	0.057	0.148	7.023E-01	-0.233	0.347
EA-PGS Quintile 5	1.002	0.144	2.960E-12	0.721	1.283
Income	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.428	0.036	7.127E-33	-0.498	-0.357
EA-PGS Quintile 2	-0.301	0.035	1.410E-17	-0.370	-0.232
EA-PGS Quintile 3	-0.199	0.035	1.936E-08	-0.268	-0.130
EA-PGS Quintile 4	-0.140	0.036	8.658E-05	-0.210	-0.070
EA-PGS Quintile 5	0.061	0.035	7.691E-02	-0.007	0.129
Townsend Deprivation Index	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.833	0.085	1.650E-22	0.666	1.000
EA-PGS Quintile 2	0.622	0.085	3.320E-13	0.455	0.790
EA-PGS Quintile 3	0.598	0.088	1.014E-11	0.426	0.770
EA-PGS Quintile 4	0.419	0.088	1.939E-06	0.247	0.592
EA-PGS Quintile 5	0.138	0.085	1.063E-01	-0.029	0.305
Numeric Memory	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.156	0.059	7.895E-03	-0.272	-0.041
EA-PGS Quintile 2	-0.086	0.056	1.235E-01	-0.195	0.023
EA-PGS Quintile 3	-0.133	0.055	1.648E-02	-0.242	-0.024
EA-PGS Quintile 4	-0.133	0.055	1.632E-02	-0.242	-0.024
EA-PGS Quintile 5	0.094	0.053	7.700E-02	-0.010	0.197
Reaction Time	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.121	0.028	1.304E-05	0.066	0.175
EA-PGS Quintile 2	0.083	0.028	2.431E-03	0.030	0.137
EA-PGS Quintile 3	0.076	0.028	7.081E-03	0.021	0.131
EA-PGS Quintile 4	0.107	0.028	1.604E-04	0.051	0.162
EA-PGS Quintile 5	0.059	0.027	3.062E-02	0.006	0.113

Time taken on Pairs Test	Beta	Standard Error	P Value	95% CI I	95% CI II
EA-PGS Quintile 1	0.159	0.095	9.510E-02	-0.028	0.345
EA-PGS Quintile 2	0.304	0.095	1.415E-03	0.117	0.491
EA-PGS Quintile 3	0.306	0.098	1.794E-03	0.114	0.499
EA-PGS Quintile 4	0.271	0.098	5.825E-03	0.078	0.463
EA-PGS Quintile 5	0.130	0.095	1.722E-01	-0.057	0.317
Height	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-1.480	0.184	1.034E-15	-1.841	-1.118
EA-PGS Quintile 2	-1.151	0.184	4.382E-10	-1.513	-0.790
EA-PGS Quintile 3	-0.666	0.190	4.523E-04	-1.039	-0.294
EA-PGS Quintile 4	-0.497	0.190	9.020E-03	-0.869	-0.124
EA-PGS Quintile 5	-0.135	0.185	4.657E-01	-0.497	0.227

Appendix table 7.4.8 (2): EA-PGS and rare variant association results across quintiles within the 325 gene subset: Binary Results

Individuals with any LoF variants in the 325 Gene Set: Binary Trait Results (n = 5776)					
Unable To Work	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.801	0.293	8.391E-23	2.281	3.439
EA-PGS Quintile 2	1.372	0.194	2.544E-02	1.040	1.810
EA-PGS Quintile 3	1.605	0.217	4.561E-04	1.232	2.092
EA-PGS Quintile 4	1.483	0.207	4.688E-03	1.129	1.950
EA-PGS Quintile 5	1.032	0.166	8.428E-01	0.753	1.415
In Employment	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.682	0.051	2.449E-07	0.589	0.788
EA-PGS Quintile 2	0.848	0.063	2.569E-02	0.734	0.980
EA-PGS Quintile 3	0.966	0.073	6.413E-01	0.833	1.119
EA-PGS Quintile 4	0.806	0.062	4.848E-03	0.694	0.937
EA-PGS Quintile 5	0.913	0.068	2.224E-01	0.790	1.056
Has a Degree	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.435	0.029	1.320E-36	0.382	0.495
EA-PGS Quintile 2	0.608	0.038	1.875E-15	0.538	0.687
EA-PGS Quintile 3	0.755	0.047	7.289E-06	0.668	0.854
EA-PGS Quintile 4	0.996	0.061	9.430E-01	0.883	1.123
EA-PGS Quintile 5	1.484	0.089	5.272E-11	1.319	1.670
Has a Child DD Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.727	0.392	2.983E-12	2.058	3.615
EA-PGS Quintile 2	2.962	0.412	5.894E-15	2.255	3.890
EA-PGS Quintile 3	2.715	0.405	2.149E-11	2.027	3.637
EA-PGS Quintile 4	2.164	0.354	2.451E-06	1.570	2.983
EA-PGS Quintile 5	1.940	0.325	7.798E-05	1.396	2.694
Has an Adult NP Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.093	0.230	1.964E-11	1.687	2.597
EA-PGS Quintile 2	1.230	0.170	1.347E-01	0.938	1.614
EA-PGS Quintile 3	1.310	0.182	5.181E-02	0.998	1.719
EA-PGS Quintile 4	1.303	0.181	5.664E-02	0.993	1.710
EA-PGS Quintile 5	1.357	0.179	2.113E-02	1.047	1.758
Has a Mental Health Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.644	0.162	4.927E-07	1.354	1.995
EA-PGS Quintile 2	1.219	0.135	7.475E-02	0.980	1.514
EA-PGS Quintile 3	1.170	0.136	1.788E-01	0.931	1.470
EA-PGS Quintile 4	0.951	0.122	6.950E-01	0.740	1.222
EA-PGS Quintile 5	1.337	0.144	7.087E-03	1.082	1.651
Never a Parent	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.076	0.083	3.387E-01	0.926	1.252
EA-PGS Quintile 2	1.206	0.091	1.288E-02	1.040	1.397
EA-PGS Quintile 3	1.513	0.110	1.166E-08	1.312	1.744

EA-PGS Quintile 4	1.353	0.101	4.985E-05	1.169	1.565
EA-PGS Quintile 5	1.377	0.099	8.204E-06	1.196	1.585
Never Pregnant	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.893	0.108	3.510E-01	0.704	1.133
EA-PGS Quintile 2	1.131	0.127	2.710E-01	0.908	1.409
EA-PGS Quintile 3	1.409	0.151	1.392E-03	1.142	1.739
EA-PGS Quintile 4	1.241	0.136	4.902E-02	1.001	1.539
EA-PGS Quintile 5	1.251	0.128	2.837E-02	1.024	1.529
Never a Father	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.245	0.126	3.018E-02	1.021	1.517
EA-PGS Quintile 2	1.269	0.129	1.938E-02	1.039	1.549
EA-PGS Quintile 3	1.620	0.161	1.166E-06	1.333	1.967
EA-PGS Quintile 4	1.468	0.150	1.736E-04	1.202	1.794
EA-PGS Quintile 5	1.530	0.155	2.777E-05	1.254	1.866

Appendix 7.4.9 (1): EA-PGS and rare variant association results across quintiles within the 125 gene subset: Continuous Results

Individuals with any LoF variants in the 125 Gene Set: Continuous Trait Results (N = 2407)					
Fluid Intelligence	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.996	0.151	4.135E-11	-1.292	-0.700
EA-PGS Quintile 2	-0.548	0.141	1.027E-04	-0.825	-0.272
EA-PGS Quintile 3	-0.214	0.135	1.129E-01	-0.478	0.051
EA-PGS Quintile 4	-0.267	0.134	4.688E-02	-0.530	-0.004
EA-PGS Quintile 5	0.141	0.136	3.003E-01	-0.126	0.408
Age Left Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.705	0.107	5.236E-11	-0.916	-0.495
EA-PGS Quintile 2	-0.264	0.113	1.961E-02	-0.486	-0.042
EA-PGS Quintile 3	-0.191	0.118	1.049E-01	-0.422	0.040
EA-PGS Quintile 4	0.175	0.121	1.483E-01	-0.062	0.412
EA-PGS Quintile 5	0.069	0.129	5.946E-01	-0.184	0.321
Years in Education	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-2.730	0.224	4.484E-34	-3.170	-2.291
EA-PGS Quintile 2	-1.067	0.224	1.886E-06	-1.505	-0.628
EA-PGS Quintile 3	-0.472	0.223	3.438E-02	-0.909	-0.035
EA-PGS Quintile 4	-0.156	0.224	4.841E-01	-0.595	0.282
EA-PGS Quintile 5	0.549	0.225	1.471E-02	0.108	0.991
Income	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.458	0.056	3.073E-16	-0.568	-0.348
EA-PGS Quintile 2	-0.383	0.055	2.943E-12	-0.491	-0.276
EA-PGS Quintile 3	-0.174	0.054	1.170E-03	-0.279	-0.069
EA-PGS Quintile 4	-0.184	0.054	6.396E-04	-0.290	-0.079
EA-PGS Quintile 5	-0.015	0.054	7.886E-01	-0.121	0.092
Townsend Deprivation Index	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.100	0.133	1.203E-16	0.840	1.360
EA-PGS Quintile 2	0.709	0.133	1.020E-07	0.448	0.971
EA-PGS Quintile 3	0.663	0.132	5.413E-07	0.404	0.922
EA-PGS Quintile 4	0.353	0.133	8.000E-03	0.092	0.613
EA-PGS Quintile 5	0.122	0.134	3.606E-01	-0.140	0.384
Numeric Memory	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-0.279	0.099	4.930E-03	-0.473	-0.084
EA-PGS Quintile 2	-0.073	0.091	4.199E-01	-0.250	0.104
EA-PGS Quintile 3	-0.057	0.083	4.912E-01	-0.221	0.106
EA-PGS Quintile 4	-0.181	0.081	2.613E-02	-0.340	-0.021
EA-PGS Quintile 5	-0.024	0.087	7.877E-01	-0.195	0.148
Reaction Time	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.204	0.043	2.548E-06	0.119	0.288
EA-PGS Quintile 2	0.177	0.043	3.758E-05	0.093	0.262
EA-PGS Quintile 3	0.053	0.043	2.089E-01	-0.030	0.137
EA-PGS Quintile 4	0.161	0.043	1.645E-04	0.077	0.245
EA-PGS Quintile 5	0.123	0.043	4.473E-03	0.038	0.207

Time taken on Pairs Test	Beta	Standard Error	P Value	95% CI I	95% CI II
EA-PGS Quintile 1	0.232	0.148	1.171E-01	-0.058	0.522
EA-PGS Quintile 2	0.224	0.149	1.318E-01	-0.067	0.515
EA-PGS Quintile 3	0.106	0.148	4.742E-01	-0.184	0.395
EA-PGS Quintile 4	0.540	0.148	2.749E-04	0.249	0.831
EA-PGS Quintile 5	0.210	0.150	1.607E-01	-0.083	0.504
Height	**Beta**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	-1.134	0.286	7.557E-05	-1.695	-0.572
EA-PGS Quintile 2	-1.325	0.288	4.311E-06	-1.889	-0.760
EA-PGS Quintile 3	-0.420	0.286	1.421E-01	-0.982	0.141
EA-PGS Quintile 4	0.046	0.288	8.725E-01	-0.518	0.610
EA-PGS Quintile 5	0.369	0.289	2.016E-01	-0.197	0.936

Appendix 7.4.9 (2): EA-PGS and rare variant association results across quintiles within the 125 gene subset: Binary Results

Individuals with any LoF variants in the 125 Gene Set: Binary Trait Results (N = 2407)					
Unable To Work	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	3.126	0.480	1.185E-13	2.313	4.225
EA-PGS Quintile 2	1.913	0.371	8.134E-04	1.309	2.797
EA-PGS Quintile 3	1.924	0.367	6.007E-04	1.324	2.796
EA-PGS Quintile 4	1.325	0.292	2.014E-01	0.860	2.042
EA-PGS Quintile 5	1.112	0.270	6.604E-01	0.692	1.789
In Employment	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.677	0.078	7.354E-04	0.539	0.849
EA-PGS Quintile 2	0.826	0.096	9.799E-02	0.658	1.036
EA-PGS Quintile 3	0.976	0.110	8.302E-01	0.783	1.217
EA-PGS Quintile 4	0.788	0.092	4.134E-02	0.627	0.991
EA-PGS Quintile 5	0.858	0.100	1.878E-01	0.683	1.078
Has a Degree	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.358	0.039	1.380E-21	0.290	0.442
EA-PGS Quintile 2	0.599	0.059	1.632E-07	0.495	0.726
EA-PGS Quintile 3	0.780	0.073	8.360E-03	0.649	0.938
EA-PGS Quintile 4	0.931	0.087	4.461E-01	0.776	1.118
EA-PGS Quintile 5	1.333	0.125	2.155E-03	1.109	1.602
Has a Child DD Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.644	0.596	1.588E-05	1.700	4.112
EA-PGS Quintile 2	2.140	0.533	2.245E-03	1.314	3.485
EA-PGS Quintile 3	2.289	0.555	6.273E-04	1.424	3.680
EA-PGS Quintile 4	1.723	0.471	4.625E-02	1.009	2.943
EA-PGS Quintile 5	1.641	0.464	7.999E-02	0.943	2.858
Has an Adult NP Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.224	0.371	1.654E-06	1.604	3.083
EA-PGS Quintile 2	1.480	0.295	4.931E-02	1.001	2.189
EA-PGS Quintile 3	1.467	0.293	5.451E-02	0.993	2.170
EA-PGS Quintile 4	1.740	0.321	2.671E-03	1.212	2.498
EA-PGS Quintile 5	1.419	0.288	8.504E-02	0.953	2.113
Has a Mental Health Related Diagnosis	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	2.047	0.289	3.842E-07	1.553	2.700
EA-PGS Quintile 2	1.492	0.238	1.194E-02	1.092	2.039
EA-PGS Quintile 3	1.164	0.205	3.888E-01	0.824	1.645
EA-PGS Quintile 4	1.139	0.204	4.667E-01	0.802	1.617
EA-PGS Quintile 5	1.417	0.233	3.412E-02	1.026	1.956
Never a Parent	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.196	0.139	1.238E-01	0.952	1.501
EA-PGS Quintile 2	1.240	0.146	6.639E-02	0.986	1.561
EA-PGS Quintile 3	1.620	0.175	8.233E-06	1.311	2.003

EA-PGS Quintile 4	1.401	0.156	2.489E-03	1.126	1.742
EA-PGS Quintile 5	1.402	0.157	2.557E-03	1.126	1.746
Never Pregnant	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	0.989	0.183	9.512E-01	0.688	1.421
EA-PGS Quintile 2	1.242	0.206	1.918E-01	0.897	1.720
EA-PGS Quintile 3	1.466	0.234	1.665E-02	1.072	2.006
EA-PGS Quintile 4	1.360	0.217	5.408E-02	0.995	1.861
EA-PGS Quintile 5	1.174	0.195	3.344E-01	0.848	1.625
Never a Father	**Odds Ratio**	**Standard Error**	**P Value**	**95% CI I**	**95% CI II**
EA-PGS Quintile 1	1.386	0.210	3.083E-02	1.031	1.864
EA-PGS Quintile 2	1.230	0.204	2.134E-01	0.888	1.703
EA-PGS Quintile 3	1.790	0.266	8.947E-05	1.338	2.396
EA-PGS Quintile 4	1.439	0.224	1.962E-02	1.060	1.954
EA-PGS Quintile 5	1.664	0.257	9.720E-04	1.230	2.253

Tables and Figures for Chapter Five

Appendix Table 7.5.1: Top *KDM5B* upstream variant predictions:

				Predicted effect on				
chromosome	position	ref	alt	Transcribed regions	Enhancer	Promoter	CTCF	Transcription factors
1	202812044	A	G	-1.063	-1.815	-1.175	-1.945	-16.211
1	202803114	T	C	2.692	-5.986	-4.512	-6.710	-14.680
1	202812051	G	C	-0.647	-1.239	-0.757	-1.466	-11.498
1	202810791	A	C	-1.744	-2.399	-1.118	-1.651	-10.947
1	202803120	C	T	4.372	-3.566	-3.193	-4.160	-10.770
1	202810906	G	A	-2.271	-5.540	-10.367	-9.720	-10.348
1	202810592	G	A	-0.452	-3.601	-4.071	-5.252	-7.689
1	202815224	A	G	-0.631	-2.358	-1.207	-1.727	-7.603
1	202810662	CCGG	C	-1.065	-3.206	-3.551	-4.489	-7.240
1	202812017	C	T	-0.681	-1.512	-0.692	-1.082	-7.065
1	202815224	A	AG	-0.769	-2.394	-1.117	-1.587	-6.999
1	202803112	G	C	3.339	-2.654	-2.441	-3.939	-6.768
1	202815222	T	G	-0.410	-1.979	-1.064	-1.565	-6.560
1	202815217	T	TA	-0.517	-1.647	-0.926	-1.355	-6.404
1	202815219	A	G	-0.199	-1.496	-1.144	-1.598	-6.195
1	202810589	T	C	-0.849	-3.240	-3.002	-3.757	-6.035
1	202803116	G	A	2.131	-1.941	-1.965	-2.918	-5.978
1	202810792	C	T	-0.382	-0.386	-0.176	-0.442	-5.927
1	202810786	T	C	-0.229	-0.208	1.048	0.716	-5.548
1	202815226	C	G	-0.441	-1.155	-0.690	-1.048	-5.217
1	202812669	G	A	-0.686	-1.390	-1.805	-2.553	-5.144
1	202810810	T	C	-1.806	-2.416	-1.980	-2.023	-5.140
1	202803339	A	C	5.305	-0.517	-1.701	-2.381	-5.127

Appendix Table 7.5.2: Top *KDM5B* downstream variant predictions:

				Predicted effect on				
chromosome	**position**	**ref**	**alt**	**Transcribed regions**	**Enhancer**	**Promoter**	**CTCF**	**Transcription factors**
1	202713239	A	G	-1.637	-2.946	-2.301	-3.289	-16.933
1	202713400	GCAAA	G	-0.953	-1.317	-1.448	-2.278	-15.367
1	202719777	C	T	-2.806	-12.437	-8.270	-21.106	-14.789
1	202732726	T	A	-0.069	-0.966	-0.237	-0.400	-11.177
1	202713277	T	C	-1.157	-2.548	-2.008	-3.042	-10.466
1	202732731	A	C	0.177	-0.627	-0.156	-0.271	-9.488
1	202732762	T	A	-0.220	-0.739	-0.195	-0.320	-9.326
1	202713271	A	T	-1.126	-2.103	-1.729	-2.539	-8.918
1	202710533	CGTGGGA	C	-1.926	-5.304	-4.995	-5.604	-6.838
1	202732733	CAACTTTAAA	C	0.059	-0.383	-0.136	-0.250	-6.134
1	202713389	C	G	-0.690	-1.364	-0.656	-1.007	-5.775
1	202710522	AGACCAAGCGGCGTGGGAGGGCGGG	A	-1.037	-3.199	-4.881	-5.030	-5.500
1	202710585	GC	G	-1.829	-4.537	-3.620	-4.092	-5.301
1	202710518	G	A	-1.213	-3.502	-4.614	-4.696	-5.296
1	202732727	G	A	0.536	-0.423	-0.158	-0.234	-5.151
1	202722168	A	G	0.131	-2.517	-1.326	-6.560	-2.633
1	202736643	GTTT	G	-7.610	0.630	0.519	0.295	-0.106
1	202726614	G	T	-7.059	-0.479	0.295	0.134	-0.937
1	202728764	T	TG	-5.200	-0.400	0.151	0.002	-0.803
1	202726651	T	C	-5.111	-0.376	0.197	0.091	-0.711

Appendix Table 7.5.3: Negative linear regression results for *KDM5B* association tests

Fluid Intelligence	Beta	Standard Error	P Value	95% CI I	95% CI II
Negative Enhancer Variant	0.081	1.470	0.957	-2.903	3.064
Negative Promoter Variant	1.065	1.972	0.593	-2.939	5.070
Negative Transcribed Region Variant	0.788	3.268	0.811	-5.847	7.423
Negative CTCF Variant	0.850	1.797	0.639	-2.798	4.499
Negative TF Variant	-0.839	0.696	0.236	-2.252	0.574
Years in Education					
Negative Enhancer Variant	-0.260	0.816	0.750	-1.873	1.352
Negative Promoter Variant	-0.229	0.816	0.780	-1.840	1.383
Negative Transcribed Region Variant	-0.515	0.825	0.533	-2.144	1.114
Negative CTCF Variant	-0.155	0.707	0.827	-1.552	1.242
Negative TF Variant	0.079	0.316	0.802	-0.544	0.703
Income					
Negative Enhancer Variant	-0.037	0.186	0.843	-0.405	0.332
Negative Promoter Variant	-0.087	0.185	0.637	-0.453	0.279
Negative Transcribed Region Variant	-0.388	0.182	0.035	-0.748	-0.028
Negative CTCF Variant	-0.026	0.159	0.868	-0.341	0.288
Negative TF Variant	0.041	0.070	0.565	-0.099	0.180
Townsend Deprivation Index					
Negative Enhancer Variant	-0.266	0.547	0.627	-1.346	0.814
Negative Promoter Variant	0.122	0.547	0.824	-0.958	1.202
Negative Transcribed Region Variant	-0.235	0.553	0.672	-1.327	0.857
Negative CTCF Variant	0.056	0.474	0.907	-0.880	0.992
Negative TF Variant	-0.064	0.211	0.762	-0.482	0.353