# 3 Case-control analysis of 565 known and candidate intellectual disability-associated genes

## 3.1 Introduction

### 3.1.1 The impact of intellectual disability

Intellectual disability (ID) is diagnosed in patients who have an intelligence quotient (IQ) of below 70, along with problems with adaptive functioning (such as problems communicating or caring for themselves), where these symptoms began before the age of 18 (135). ID is typically classified as mild (IQ 50-70) or severe (IQ below 50) although other categories can be used. It is phenotypically heterogeneous; in addition to variable IQ and different manifestations of problems with adaptive functioning, it often occurs in conjunction with other abnormalities, such as seizures, behavioural difficulties, dysmorphic facial features, or other developmental disorders such as congenital heart disease (CHD). A particularly common comorbidity of ID is autism spectrum disorder (ASD), with 28% of people with ID also suffering from ASD (136). ID with additional comorbidities is often classified as ‘syndromic’, and cases with no additional symptoms are ‘non-syndromic’ (137). However, recent opinion in the ID research community has shifted away this dichotomous categorisation, in favour of considering ID as a spectrum, with variable additional phenotypes. This is partly because subtle comorbidities or specific intellectual disabilities shared among groups of patients are often not obvious until they are retrospectively grouped according to aetiology (135).

Collectively, ID is a very common developmental disorder, with a prevalence of around 1-2%, but estimates of prevalence vary widely depending on factors including the definition of ID, the population studied, and age group (138, 139). Importantly, the prevalence also depends on sex, with males accounting for ~57% of ID cases (140). The majority of patients with ID require extensive medical, financial and personal

support throughout their lives, causing ID to be one of the most costly diseases in high-income countries (141). Because ID is so prevalent, it therefore has a profound impact not only on patients and their families, but also on healthcare providers and society as a whole.

The causes of ID are wide ranging and include environmental and genetic factors. Environmental factors that are associated with increased risk of ID include malnutrition during infancy, prenatal exposure to alcohol or the rubella virus, childhood exposure to lead, brain injury during birth, and low birth weight (142-146). Fetal alcohol syndrome affects 0.1-0.7% of births, and is the most common preventable cause of ID in high-income countries. With this exception, environmental factors disproportionately affect people in low-income countries, and explain the increased prevalence of ID in such countries (139).

Genetic causes of ID have been recognised for many decades. Lionel Sharples Penrose was the first to conduct a large study on the subject, which was published in 1938 (147). He assembled and investigated a cohort of 1280 cases of ID. His pioneering observations included the sex bias in prevalence of ID, and the fact that related patients often have similar phenotypes. Historical studies such as this draw attention to two relevant ethical issues. First, ID, possibly more than any other medical condition, uses terminology that has evolved. In Penrose's study, for example, patients are classified according to whether they are "dull", "simpletons", "imbeciles" or "idiots". By 1960, these offensive terms had been replaced in the medical and research communities by the term mental retardation. Gradually, this term too attracted derogatory connotations, and in 2009 a law (known as Rosa's law) was passed in the USA officially replacing it with the term intellectual disability. Second, early studies of the genetics of ID are tainted by their unpleasant association with the eugenics movement. For example, in "The Eugenics Review", Eliot Slater describes aspects of Penrose's study to be "of profound eugenic significance" (148). J. B. S. Haldane, commenting on Penrose's study in Nature, took a moderate approach, emphasising the complexity of the aetiology of ID, and calling the claims that it could be largely eliminated by sterilisation of patients to be "extravagant" (149).

From these beginnings, research into the genetics of ID and intelligence has flourished. Intelligence is a quantitative trait, and is highly heritable (150). Mild, non-syndromic ID represents the bottom of the normal distribution of IQ, and these cases are likely to be influenced by multiple genetic and environmental factors each with a small effect size,

as for any quantitative trait. To start to understand the genetic architecture of these cases will require genome-wide association studies with extremely large sample sizes (151). However, moderate to severe ID is thought to be usually caused by a single pathogenic variant with a large effect. Identification of these variants and understanding how they cause ID is of great importance.

### 3.1.2 Discovery of intellectual disability-associated genes

Cytogenetically visible chromosomal aberrations comprising aneuploidies, rearrangements, and large copy number variants (CNVs) cause around 15% of ID cases (152). Trisomy 21 (also known as Down syndrome) is the most common genetic cause of ID, and was the first ID-associated variant to be discovered. It accounts for ~10% of ID cases (152), and the molecular defect was first identified in 1959, although the syndrome had been recognised since 1866 (153).

The introduction of chromosomal microarrays increased the resolution at which variants could be identified to the submicroscopic level. Submicroscopic CNVs are a frequent cause of ID. For example, heterozygous *de novo* 17q21.31 microdeletions (500-650 kb) can cause a syndrome comprising ID, motor and speech delay, dysmorphic facial features and hypotonia (154). Another study used array comparative genomic hybridisation (aCGH) on a large cohort to demonstrate that submicroscopic CNVs (with a median size of 213 kb) account for ~14% of ID cases (155). Interestingly, they also showed that CNVs disproportionately cause syndromic rather than non-syndromic ID, especially where the additional abnormalities are structural (such as cardiovascular or craniofacial defects). Investigation of the critical region of CNVs often leads to discovery of novel ID-associated genes such as *MBD5* and *KANSL1* (156, 157). There is also evidence that some cases of ID are caused by a 'two-hit' model, where two different CNVs are required for manifestation of disease (158). This finding blurs the dichotomy between monogenic and polygenic models of disease.

Historically, discovery of single gene causes of ID was largely limited to families with a typical pattern of X-linked inheritance. *FMR1* was the first X-linked ID-associated gene to be identified, by positional mapping of yeast artificial chromosome clones followed by Sanger sequencing (159). Triplet expansion repeats within *FMR1* cause fragile X syndrome, which is the most common single gene cause of ID, accounting for ~0.5% of cases (160). Another important example of an X-linked ID-associated gene is *MECP2*,

pathogenic variants in which were originally found to be the cause of Rett syndrome in females. Pathogenic *MECP2* variants have since been implicated in a variety of forms of ID in both males and females, although they are a much more common cause in females (161). Single gene, X-linked ID accounts for ~10% of ID cases overall (162). A study published in 2009 illustrated the importance of large-scale sequencing in the discovery of ID-associated genes (13). The authors recruited 208 families with X-linked ID, and sequenced 65% of all the coding regions of the X chromosome by Sanger sequencing. This was the largest systematic screen for pathogenic variants at the time, and discovered nine novel X-linked ID-associated genes including *CASK*.

The widespread availability of next generation sequencing (NGS) that flourished very shortly after the publication of the study just described, opened up possibilities of ID-associated gene discovery on a whole new scale. For the first time, autosomal single nucleotide variants (SNVs) and insertion deletions (indels) that cause ID could be identified systematically. To achieve this, one study performed exome sequencing in 136 consanguineous families affected by autosomal recessive ID (163). Homozygosity mapping in each family allowed the analysis to be restricted to loci likely to contain the causative variant. As well as identifying pathogenic variants in known ID-associated genes, 50 possible novel ID-associated genes were identified, including some that have subsequently been confirmed, including *KIF7*, *MAN1B1*, and *TAF2*.

Exome sequencing using a trio study design has repeatedly shown that *de novo* mutations are a very important cause of ID, and account for a large proportion of cases (57, 84, 164). The *de novo* paradigm of ID along with the reduced reproductive fitness of ID patients probably explains the long known observation that many forms of ID occur sporadically. Also, it explains the apparent paradox between the relatively high prevalence of ID, and the fact that it significantly reduces reproductive fitness.

Whole genome sequencing can identify coding pathogenic variants that were missed by exome sequencing (62). Admittedly, the exome sequencing in the original study may have called variants with lower sensitivity than subsequent studies, as demonstrated by the low *de novo* exome mutation rate of 0.53 per patient (57). Nevertheless, whole genome sequencing has fewer biases in variant calling, and greater uniformity of coverage than whole exome sequencing, suggesting that an eventual move away from exome sequencing towards whole genome sequencing is likely (63).

### 3.1.3 Biology of intellectual disability-associated genes

Over 500 single genes in which pathogenic variants may cause ID have been identified thus far, with many more unconfirmed candidates (62). ID is so genetically heterogeneous that it is appropriate to consider the term to be a hypernym describing many individual syndromes and non-syndromic forms (165). ID-associated genes may be classified and understood according to their function, or the pathway in which they act (Table 3-1) (135). This is helpful because it facilitates identification of further candidate genes, and helps with prognosis, and because pathogenic variants in different genes in the same pathway often cause similar phenotypes. Some functional classes affect universal cellular processes, whereas others are very specific to neurological processes.

| Functional class of ID-associated gene | Examples | References |
|---|---|---|
| Presynaptic vesicle release and recycling | *STXBP1; CASK; IL1RAPL1* | (166-168) |
| Neurotransmitter receptors | *GRIA3; GRIN2A; GRIN2B* | (101, 169) |
| Components of the post-synaptic density | *SYNGAP1; SHANK2* | (170, 171) |
| Regulators of gene expression | *MECP2; EHMT1; ARID1B; FMR1* | (161, 172-174) |
| Metabolism | *PAH; PMM2* | (175, 176) |

**Table 3 - 1: Functional classes of ID-associated genes.**

At a typical synapse, the presynaptic terminal contains vesicles filled with neurotransmitter. The primary excitatory neurotransmitter is glutamate, and the primary inhibitory neurotransmitter is γ-aminobutyric acid (GABA). When stimulated, the vesicles fuse with the presynaptic membrane and exocytose their contents into the synaptic cleft, whereupon the cell recycles the vesicles. The release and recycling of pre-synaptic vesicles are complex biological processes involving many proteins. Pathogenic variants in genes that encode some of these proteins can cause ID. For example, *de novo* mutations in *STXBP1* can cause Ohtahara syndrome (166). *STXBP1* encodes Munc18-1, a protein required for fusion of the vesicles with the presynaptic membrane. CASK is also involved in exocytosis (167). IL1RAPL1, on the other hand, *inhibits* neurotransmitter release; pathogenic variants in *IL1RAPL1* can cause non-syndromic X-linked ID, ASD or schizophrenia (168).

In the synaptic cleft, neurotransmitters bind to receptors on the postsynaptic membrane on dendritic spines of the neuron receiving the signal. For excitatory synapses, the two main types of glutamate receptors are NMDA and AMPA receptor. Pathogenic variants in genes that encode subunits of these receptors can cause ID. For example, variants in *GRIA3*, which encodes a subunit of the AMPA receptor, can cause moderate X-linked ID (169). Similarly, *de novo* mutations in *GRIN2A* or *GRIN2B*, which encode subunits of the NMDA receptor, can cause ID and seizures (101).

Neurotransmitter receptors are part of an extensive protein complex called the postsynaptic density (PSD). Proteins in this complex perform many functions from regulating and propagating the signal, to providing structural support to the receptors. Integrity of the PSD is required for various cognitive processes including learning and memory. It is therefore unsurprising that mutations in PSD proteins other than the receptors themselves (such as the regulatory protein SYNGAP1 or the scaffolding protein SHANK2) can cause ID and other neurodevelopmental disorders (170, 171).

People with ID may or may not have structural brain abnormalities apparent on imaging such as magnetic resonance imaging (MRI). Regardless of this, histology on post-mortem brain samples often shows characteristic changes to the structure of dendrites and dendritic spines compared to healthy people, although it is unclear whether this is a cause or a consequence of the cognitive defect (177). Plasticity of dendritic spine morphology is important for cognitive functioning. Rapid changes to dendritic spine morphology are achieved by remodelling of actin filaments and microtubules. Pathogenic variants in genes that encode proteins that regulate this remodelling process can therefore cause ID (including *OPHN1* and *FGD1*) (178, 179).

Glutamate binding to NMDA or AMPA receptors activates signaling cascades such as the RAS-MAPK pathway in the postsynaptic neuron. Pathogenic variants in members of this pathway cause a family of diseases that are becoming known as RASopathies, one common feature of which is ID. For example, *de novo* mutations in *HRAS* can cause Costello syndrome (180). Typical features of Costello syndrome are ID, short stature, excess skin and dysmorphic craniofacial features. Interestingly, RASopathies may potentially be one class of ID that could benefit from pharmaceutical intervention (181).

Another important class of ID-associated genes is regulators of gene expression. Appropriate transcription and translation of downstream genes is necessary for cognitive function. This is demonstrated by the fact that pharmaceutical inhibition of

protein synthesis using an agent such as anisomycin inhibits the formation of memories (182). Several functional classes of genes regulate gene expression. Transcription factors do so by directly binding to DNA response elements, histone modifiers catalyse the addition or removal of groups (e.g. acetyl or methyl groups) to or from histone proteins, and DNA methyltransferases catalyse the transfer of methyl groups onto DNA itself. Transcription regulators are increasingly recognised as an important cause of ID. The problem with understanding how they do so is that usually the downstream genes whose expression is altered are not known. MECP2 is a transcription regulator that binds to the methylated DNA response element of a downstream gene and initiates formation of a complex that silences the gene. Euchromatic histone methyltransferase 1, encoded by *EHMT1*, catalyses the transfer of methyl groups onto lysine residues of histone proteins and is particularly enriched in brown adipose tissue (183). Disruption of *EHMT1* can cause Kleefstra syndrome, where patients have ID, hypotonia, brachycephaly, dysmorphic facial features, and CHD (172). Similarly, heterozygous *de novo* mutations in the SWI/SNF chromatin remodelling complex component *ARID1B* are a more frequent cause of ID, accounting for ~1% of previously undiagnosed cases (173). FMRP, which is encoded by *FMR1*, is an RNA-binding protein that regulates the expression of other proteins including components of the PSD (174).

Finally, pathogenic variants in metabolic genes can cause inborn errors of metabolism (IEM), a common feature of which is ID. For example, recessive variants in *PAH*, which encodes phenylalanine hydroxylase, cause phenylketonuria (PKU). Patients with untreated PKU have ID, seizures, microcephaly and hypopigmentation (175). Most developed countries have implemented screening programs for PKU, and treat patients from infancy with dietary changes and medication. Another example of an IEM where ID is a feature is congenital disorder of glycosylation (CDG). Here, pathogenic recessive variants in genes such as *PMM2*, which are involved in glycosylation of downstream proteins, cause phenotypes including ID, cardiomyopathy, frequent infections, central nervous system and eye defects (176).

Interestingly, the functional class of a gene can affect aspects of the associated disease, such as the mode of inheritance. Genes associated with IEMs have recessive inheritance, whereas genes encoding chromatin modifiers are usually haploinsufficient, so pathogenic variants cause disease with dominant inheritance (172, 173, 175, 176). Intuitively, it seems likely that pathogenic variants in ID-associated genes that are very specific to neurological processes might, on average, cause ID that is largely non-

syndromic, whereas pathogenic variants in ID-associated genes that affect universal cellular processes might cause a more syndromic phenotype, because more systems will be affected. While some of the examples I have given in this section support this hypothesis, others do not. More large-scale, unbiased studies of ID are required to establish whether this pattern exists.

The current diagnostic yield for ID patients is up to 50-65% (62, 135, 152). There are several reasons why 35-50% of ID patients still do not receive a genetic diagnosis, including the possibility that the causative variant could be in a non-coding region, it could be in a gene not known to be ID-associated, or the disease could be caused by several variants acting in an oligogenic manner. Nevertheless, it is clear that more ID-associated genes remain to be identified.

### 3.1.4 Case-control enrichment analysis of rare variants

Rare disease-associated genes are usually identified by means of a classical, case-only diagnostic approach, where they are identified because they contain rare, coding variants which segregate with disease in multiple families, for example. Case-control enrichment analysis is a supplementary method that can yield additional insights into the aetiology of rare disease. Typically, a cohort of cases is assembled, along with a cohort of controls. Rare variants are identified in both cohorts (for example by exome sequencing), and then a statistical test is applied to test the hypothesis that the cases have an excess of a defined category of variants compared to controls. Case-control enrichment analysis can yield insights into the genetic architecture of a rare disease without necessarily assigning causality to individual variants. It can be used with a range of study designs, whereas classical approaches often require very specific study designs. For example, to identify *de novo* mutations DNA samples from both biological parents are required, which are not always available. Perhaps most importantly, case-control enrichment analysis makes fewer assumptions about causative variants than classical approaches, and therefore takes into account non-classical contributors to disease such as variants with incomplete penetrance, and variants that contribute to a phenotype in an oligogenic manner.

Several different statistical tests have been developed for use in case-control enrichment analysis (recently reviewed in (184)). Three of the most commonly used are the cohort allelic sums test (CAST), the weighted sum method, and the sequence

kernel association test (SKAT). CAST is a burden test, whereby information about a variant category of interest is collapsed into whether each individual has any variant of that category, or whether they do not. A statistical test (usually Fisher's exact test) is then applied to this count data to assess the degree and significance of any difference between the cohorts. CAST was first formally described in 2007 (185), although it had been used prior to this (186). It is a very flexible test, in that it can be used to test for association between an individual gene and a phenotype, association between a group of genes and a phenotype, or even a genome wide burden of variants. A disadvantage of the CAST test is that it assumes that the direction and size of effect of all variants are the same. If this assumption is not true, power is lost. CAST also assumes that a fairly large proportion of variants are causal. Also, by collapsing information, power is lost. For example, an individual with ten rare variants of interest is treated with the same weight as an individual with only one such variant, whereas it may be more appropriate for the individual with ten variants to be given a higher weight in the test. However, where the assumptions are true, CAST is a robust and powerful test (185).

Because of the assumption of CAST that a relatively high proportion of variants are pathogenic, prior to performing CAST, filtering based on minor allele frequency should be performed (a typical cutoff is 0.01). However, then a unique variant is still treated with the same weight as a variant with a frequency of 0.01, whereas it may be more appropriate for the unique variant to be given a higher weight in the test. The weighted sum method is very similar to CAST, but variants of all frequencies are included, and collapsed into a single average number of rare alleles per case, weighted according to variant frequency in controls (187). Therefore, the weighted sum method has greater power than CAST if one wants to simultaneously test variants of different frequencies. The weighted sum method makes the same assumptions as CAST about direction and size of effect.

SKAT is a variance-component test, which uses a regression framework to evaluate differences in the distribution of various scores between variants in cases and controls without collapsing the information into a single statistic (188). It is flexible, computationally efficient, can account for covariates, and makes no assumptions about direction and size of genetic effect. Where a phenotype is influenced by variants with different directions of effect, SKAT is much more powerful than CAST or the weighted sum method. However, where the effects are in the same direction, and most variants are pathogenic, CAST is more powerful (184).

Regardless of the statistical test selected, case-control association or enrichment analyses are potentially subject to spurious findings if there are systematic differences between cases and controls. These can be technical differences, if, for example, the cases and controls were sequenced in different batches. Population stratification between cases and controls can lead to differences in allele frequency that can falsely appear as a disease association (189). One commonly used method by which to detect, and if necessary adjust for, population stratification is principal component analysis (PCA).

Several studies demonstrate the utility of case-control enrichment analysis in understanding the role of variants in rare disease. In an early example, Cohen *et al.* Sanger sequenced the coding regions of three genes in which pathogenic variants can cause Mendelian forms of low high-density lipoprotein cholesterol (HDL-C), in individuals from the general population with low HDL-C levels, compared to controls with high HDL-C levels (186). They used the CAST test to demonstrate that individuals with low HDL-C levels had a significant burden of rare non-synonymous variants in the candidate genes compared to the high HDL-C controls, suggesting some shared aetiology between Mendelian forms of low HDL-C, and low HDL-C in the general population. In another important example, Cooper *et al.* identified an enrichment of rare, large (>400 kb) CNVs in children with ID compared to controls (155).

In 2013, Liu *et al.* whole exome sequenced a cohort of over 1000 ASD patients, along with 870 controls (190). The authors used both the weighted sum test and the SKAT test in an attempt to identify novel ASD-associated genes, with an excess of rare, coding variants in cases. They did not find any genes, known or novel, with an exome-wide significant burden, demonstrating that much larger sample sizes are required for gene discovery using this method.

Purcell *et al.* recently took a slightly different approach, in order to investigate the genetic aetiology of schizophrenia (14). Instead of looking for a burden in individual genes, the authors took a ‘top-down’ approach, and focused on groups and subgroups of candidate genes. This method increased their power to detect an enrichment of variants, and simultaneously reduced the burden of multiple testing, which proved to be successful. Using a combination of the CAST and SKAT tests on exome sequencing data, the authors identified an enrichment of rare coding variants in candidate schizophrenia genes in patients with schizophrenia compared to controls. A particularly large enrichment was identified in components of the postsynaptic activity-regulated

cytoskeleton-associated scaffold complex, emphasising the importance of this complex in the aetiology of schizophrenia.

### 3.1.5 Aims, context, and colleagues

The overall aims of this project were threefold. The first aim was to identify pathogenic loss of function (LOF) and missense variants in known ID-associated genes in ID patients, the second was to identify novel ID-associated genes, and the third was to determine whether there is a significant enrichment of variants in ID-associated genes in ID patients compared to controls. These aims were addressed by means of a targeted resequencing study of rare diseases that was carried out as part of the UK10K project. This project was a large collaborative effort. In this chapter, I have included a few instances of work done by other people, where it is necessary to put my own work into context. I have made it clear who did the work at the point I describe it, and I also summarise it here.

The UK10K rare disease consortium, chaired by Dr Matthew Hurles and Dr David Fitzpatrick designed and implemented the study. Dr Lucy Raymond led the ID cohort, and along with Dr Detelina Grozeva and Dr Olivera Spasic-Boskovic assembled and prepared samples, selected ID-associated genes to be sequenced, did the case-only diagnostic analysis, novel gene identification, and validations. An international collaborative team of clinicians and researchers including Dr Michael Parker, Dr Hayley Archer, Dr Helen Firth, Dr Soo-Mi Park, Dr Natalie Canham, Dr Susan Holder, Dr Meredith Wilson, Dr Anna Hackett, and Dr Michael Field contributed samples to the ID cohort. Professor Shoumo Bhattacharya, Dr Jamie Bentham, and Dr Catherine Cosgrove assembled the CHD cohort. Dr James Floyd designed the custom sequencing pull-down experiment and performed quality control analysis on the data. The high-throughput sequencing team at the Wellcome Trust Sanger Institute (WTSI) did the DNA amplification, pull-down and sequencing. Dr Shane McCarthy led the initial bioinformatics including read mapping and variant calling. Dr Saeed Al Turki wrote some Python scripts that I used during this project.

The parts of the project for which I was responsible are as follows: annotating variants, designing and implementing a filtering pipeline to identify possibly causative variants, assisting with interpretation of data to identify causative variants and novel genes, and designing and performing an extensive series of burden tests to investigate the extent

to which variants in ID-associated genes are enriched in ID patients (including PCA and CAST). I carried out this work under the supervision and guidance of Dr Matthew Hurles.

Some parts of this chapter have been published ((191) and manuscript in preparation). Unless otherwise stated, where material in this chapter is taken from those publications, I declare that those sections were originally my own work.

## 3.2 Methods

### 3.2.1 Samples, sequencing, and quality control

Genomic DNA of 2812 individuals with one of seven rare diseases was whole genome amplified using 1µl of 10ng/µl template DNA using GenomiPhi kit (GE Healthcare). Dr James Floyd designed custom targeted Agilent SureSelect pull-down baits using the SureDesign program (Agilent Technologies, Santa Clara, CA, USA). This targets 3.35 Mb of sequence from the coding exons (GRCh37) of 1189 genes. These genes consist of candidates for each of the seven rare diseases. The 565 sequenced ID-associated genes were selected by Dr Lucy Raymond, and an international collaborative team of clinical geneticists assembled the ID samples. Target enrichment was done using a custom SureSelect library (Agilent Technologies), according to the manufacturer's instructions. The Illumina HiSeq 2000 platform (Illumina, Inc. San Diego, USA) was used to perform the sequencing. The high-throughput sequencing team at WTSI did the DNA amplification, pull-down and sequencing.

Dr Shane McCarthy at WTSI led the work described in this paragraph. Each read was aligned to the reference genome (GRCh37) using the Burrows-Wheeler Alignment tool, and SNVs and indels were identified using both SAMtools mpileup and the GATK UnifiedGenotyper, and these variants calls were merged, prioritising GATK calls at sites where there was a discrepancy (76, 192). Variants were stored in variant call format (VCF) files both as single-sample and multi-sample calls. Functional annotations were added using the Ensembl Variant Effect Predictor v2.8 against Ensembl 70 (193). Additionally, some basic filters were applied to the variants; including removal of very low coverage calls, and calls where the reference base is unknown. Dr James Floyd generated and analysed quality control metrics.

### 3.2.2 Annotation and filtering pipeline

I used a python script written by Dr Saeed Al Turki to add minor allele frequency data to each variant from the following sources: 1000 genomes database, UK10K twins cohort, exome sequencing project (ESP) 6500, and a cohort of 2172 control individuals exome sequenced at WTSI. I wrote an R script to calculate and annotate the internal

variant frequency (the frequency with which each variant appeared in the UK10K replication study, including all phenotypes).

I designed and implemented a filtering pipeline using R, to generate a list of rare, possibly causative variants from the merged and annotated VCF files. I only considered variants that had minor allele frequency < 0.01 in all four databases described, internal frequency < 0.01, quality score > 40, and mapping quality score > 50. I selected the quality score and mapping quality score cutoffs by visually inspecting the original sequencing data of a subset of variants using The Integrative Genomics Viewer (IGV) (82). I also removed heterozygous calls on the X chromosome in males. I selected the most severe consequence of each variant, and considered only variants with two categories of consequence: functional (coding sequence variants, in-frame deletions, in-frame insertions, initiator codon variants, missense variants, and variants resulting in loss of a stop codon) or predicted LOF (nonsense, frameshift or essential splice site variants). Finally, I considered only variants in the sequenced ID-associated genes.

To determine whether there was an excess of *de novo* LOF mutations in a particular gene, I calculated the number expected to occur by chance using the known exome mutation rate, the proportion of mutations that are expected to be LOF, and taking into account the length of the coding sequence of the gene (83, 84). I compared this to the observed number of *de novo* LOF mutations, assuming a Poisson distribution to calculate a p-value, which I corrected for testing of multiple genes using the Bonferroni correction.

### 3.2.3 Principal component analysis

PCA was done using the R package SNPRelate, which is a convenient and computationally efficient tool (194). I converted VCF files of multi-sample calls for each of the ID and CHD cohorts to GDS format using the snpgdsVCF2GDS function of the SNPRelate package. I used the snpgdsLDpruning function of the SNPRelate package to identify a list of 2291 high-quality, biallelic and polymorphic SNVs with minor allele frequency ≥0.05, that are not in linkage disequilibrium (LD) with each other, in the UK10K samples. Next, I performed PCA on the UK10K samples along with a subset of unrelated HapMap3.3 samples, using the snpgdsPCA function of the SNPrelate package and the 2291 SNVs identified (195).

### 3.2.4 Cohort allelic sums test

I wrote an R script that reads in a file of variants in sequenced ID-associated genes in the ID cohort and CHD controls. The script identifies and removes samples with excessive numbers of variants. Additional filters can then be applied to the variants, for example to remove those which have an IGV plot suggestive of a false positive call, or to apply more stringent internal frequency cutoffs. Next, the variants are subcategorised. I classified the 565 genes into known (n=204; Table 3-2) and candidate (n=361; Table 3-3) according to whether they are present in a stringent, manually curated list of known ID-associated genes in a recently published study (62). The script counts the number of variants in each sample, and generates a 2x2 contingency table where each row is one of the two cohorts, and the two columns respectively show the number of samples who have at least one variant, and the number who do not. Finally, a one-tailed Fisher's exact test is performed on the contingency tables.

| | | | | | |
|---|---|---|---|---|---|
| *ABCD1* | *ADCK3* | *ADSL* | *AFF2* | *AGA* | *AGTR2* |
| *ALDH18A1* | *ALDH5A1* | *ALG1* | *ALG12* | *ALG3* | *ALG6* |
| *ANK3* | *AP1S2* | *ARFGEF2* | *ARHGEF9* | *ARID1A* | *ARID1B* |
| *ARX* | *ASXL1* | *ATP7A* | *ATRX* | *AUH* | *BCOR* |
| *BRAF* | *CASK* | *CC2D2A* | *CCDC22* | *CDH15* | *CDKL5* |
| *CEP41* | *CHD2* | *CHD7* | *CNTNAP2* | *CREBBP* | *CTNNB1* |
| *CUL4B* | *DCX* | *DHCR7* | *DKC1* | *DLG3* | *DMD* |
| *DNMT3B* | *DYNC1H1* | *DYRK1A* | *EHMT1* | *EP300* | *ERCC6* |
| *EXOSC3* | *FGD1* | *FKRP* | *FKTN* | *FLNA* | *FMR1* |
| *FOXG1* | *FOXP1* | *FTSJ1* | *GCH1* | *GDI1* | *GJC2* |
| *GK* | *GPC3* | *GPR56* | *GRIA3* | *GRIN2A* | *GRIN2B* |
| *HCCS* | *HCFC1* | *HDAC4* | *HDAC8* | *HPRT1* | *HRAS* |
| *HSD17B10* | *HUWE1* | *IDS* | *IDUA* | *IKBKG* | *IL1RAPL1* |
| *INPP5E* | *IQSEC2* | *KANK1* | *KANSL1* | *KAT6B* | *KCNQ3* |
| *KDM5C* | *KIF7* | *KIRREL3* | *KRAS* | *L1CAM* | *LAMP2* |
| *LRP1* | *LRP2* | *MAP2K1* | *MAP2K2* | *MBD5* | *MECP2* |
| *MED12* | *MEF2C* | *MID1* | *MLH1* | *MLL2* | *MLL3* |
| *MLYCD* | *MMAA* | *MMADHC* | *MYT1L* | *NDE1* | *NDP* |
| *NEU1* | *NF1* | *NFIX* | *NHS* | *NLGN4X* | *NRXN1* |
| *NSD1* | *NSDHL* | *NSUN2* | *OCRL* | *OFD1* | *OPHN1* |
| *OTC* | *PAFAH1B1* | *PAK3* | *PARP1* | *PAX6* | *PC* |
| *PCDH19* | *PCNT* | *PDHA1* | *PEPD* | *PGK1* | *PHF6* |
| *PHF8* | *PLP1* | *PNKP* | *POLR3A* | *POLR3B* | *PORCN* |
| *PRPS1* | *PTCHD1* | *PTEN* | *PTPN11* | *RAB3GAP1* | *RAF1* |
| *RAI1* | *RPS6KA3* | *SATB2* | *SCN2A* | *SCN8A* | *SETBP1* |
| *SETD5* | *SHANK2* | *SHANK3* | *SHOC2* | *SHOX* | *SHROOM4* |
| *SLC12A6* | *SLC16A2* | *SLC26A9* | *SLC2A1* | *SLC6A8* | *SLC9A6* |
| *SMARCA2* | *SMARCA4* | *SMARCB1* | *SMARCE1* | *SMC1A* | *SMS* |
| *SOS1* | *SOX3* | *SOX5* | *SPRED1* | *SPTAN1* | *SRGAP3* |
| *STXBP1* | *SYN1* | *SYNE1* | *SYNGAP1* | *SYP* | *TAT* |
| *TBC1D24* | *TCF4* | *TIMM8A* | *TRAPPC9* | *TSC1* | *TSC2* |
| *TSPAN7* | *TUBA1A* | *TUBB2B* | *TUSC3* | *UBE2A* | *UBE3A* |
| *UBR1* | *UPF3B* | *VLDLR* | *VPS13B* | *WDR11* | *WDR62* |
| *ZDHHC9* | *ZEB2* | *ZFHX4* | *ZFYVE26* | *ZNF41* | *ZNF674* |

**Table 3 - 2: List of 204 sequenced intellectual disability-associated genes that are known.** Genes were classified as known if they are in a stringent, manually curated list of known ID-associated genes from a recently published study (62). *SETD5* was included in this list for the purposes of the case-control enrichment analyses, as a result of findings described in this chapter.

| | | | | | |
|---|---|---|---|---|---|
| *ACBD6* | *ACE2* | *ACIN1* | *ACOT9* | *ACSL4* | *ACTL6A* |
| *ACTL6B* | *ACY1* | *ADK* | *ADRA2B* | *AIMP1* | *AKAP17A* |
| *AKAP4* | *ALDH4A1* | *ALG13* | *ALG8* | *AP4B1* | *AP4E1* |
| *AP4M1* | *AP4S1* | *ARG1* | *ARHGAP36* | *ARHGAP6* | *ARHGEF4* |
| *ARHGEF6* | *ARID2* | *ARIH1* | *ARL14EP* | *ARSF* | *ASB12* |
| *ASCC3* | *ASCL1* | *ASH1L* | *ASMT* | *ASMTL* | *ATM* |
| *ATP2B3* | *ATXN3L* | *AVPR2* | *AWAT2* | *BCORL1* | *BDP1* |
| *BMP15* | *BRWD3* | *BTK* | *C12orf57* | *CA8* | *CACNA1F* |
| *CACNA1G* | *CAMK2A* | *CAMK2G* | *CAP1* | *CAPN10* | *CASP2* |
| *CC2D1A* | *CCDC23* | *CCNA2* | *CCNB3* | *CD99* | *CDK16* |
| *CDK8* | *CFP* | *CHL1* | *CLCN4* | *CLCN5* | *CLIC2* |
| *CMC4* | *CNKSR1* | *CNKSR2* | *COL4A3BP* | *COL4A6* | *COQ5* |
| *COX10* | *CPXCR1* | *CRLF2* | *CSF2RA* | *CSTF2* | *CTPS2* |
| *CTSD* | *CTTNBP2* | *CUX2* | *CXORF22* | *CXORF58* | *CYP7B1* |
| *DCHS2* | *DDOST* | *DDX26B* | *DDX3X* | *DDX53* | *DEAF1* |
| *DGKH* | *DGKK* | *DHRSX* | *DHX30* | *DIAPH2* | *DLG1* |
| *DLG2* | *DLG4* | *DOCK11* | *DPF1* | *DPF2* | *DPF3* |
| *EEF1A2* | *EEF1B2* | *EIF2C1* | *EIF2S3* | *ELK1* | *ELP2* |
| *ENOX2* | *ENTHD2* | *ENTPD1* | *EPPK1* | *ERLIN2* | *ESX1* |
| *FAAH2* | *FAM120C* | *FAM47B* | *FAM58A* | *FASN* | *FKBPL* |
| *FRMPD4* | *FRY* | *FTL* | *GAB3* | *GABRQ* | *GAD1* |
| *GATAD2B* | *GCDH* | *GLB1* | *GLRA2* | *GM2A* | *GON4L* |
| *GPR112* | *GPRASP1* | *GRB14* | *GRIA1* | *GRIA2* | *GRIK2* |
| *GSPT2* | *GTPBP8* | *HAUS7* | *HDHD1* | *HEXA* | *HEXB* |
| *HGSNAT* | *HIST1H4B* | *HIST3H3* | *HIVEP2* | *HS6ST2* | *HSPD1* |
| *IFNAR2* | *IGSF1* | *IL3RA* | *INPP4A* | *ITGA4* | *ITIH6* |
| *KCNC3* | *KCND1* | *KCNH1* | *KCNK12* | *KDM1A* | *KDM5A* |
| *KDM6B* | *KIAA2022* | *KIF1A* | *KIF26B* | *KIF4A* | *KIF5C* |
| *KLHL15* | *KLHL21* | *KLHL34* | *KLHL4* | *LAMA1* | *LARP7* |
| *LAS1L* | *LHFPL3* | *LIMK1* | *LINS* | *LRRK1* | *MAGEA11* |
| *MAGEB1* | *MAGEB10* | *MAGEB2* | *MAGEC1* | *MAGEC3* | *MAGED1* |
| *MAGEE2* | *MAGIX* | *MAGT1* | *MAN1B1* | *MAOA* | *MAOB* |
| *MAP3K15* | *MAP7D3* | *MBNL3* | *MED17* | *MED23* | *MGAT5B* |
| *MIB1* | *MLC1* | *MMAB* | *MORC4* | *MSL3* | *MTF1* |
| *MTMR1* | *MTMR8* | *MXRA5* | *MYO1D* | *MYO1G* | *NAA10* |
| *NDST1* | *NDUFA1* | *NECAB2* | *NKAP* | *NLGN3* | *NR1I3* |
| *NRK* | *NRXN2* | *NTM* | *NXF4* | *NXF5* | *ODF2L* |
| *OGT* | *OR5M1* | *OXCT1* | *P2RY4* | *P2RY8* | *PABPC5* |
| *PAH* | *PASD1* | *PBRM1* | *PCDH10* | *PECR* | *PGRMC1* |
| *PHACTR1* | *PHF10* | *PHIP* | *PHKA1* | *PIGN* | *PIK3C3* |
| *PIN4* | *PJA1* | *PLA2G6* | *PLCXD1* | *PLXNB3* | *POLA1* |
| *PPP2R5D* | *PPT1* | *PQBP1* | *PRDX4* | *PRICKLE3* | *PRMT10* |
| *PROX2* | *PRRG1* | *PRRG3* | *PRRT2* | *PRSS12* | *PSMA7* |
| *PSMD10* | *PTPN21* | *RAB39B* | *RAB40AL* | *RABL6* | *RALGDS* |
| *RAPGEF1* | *RBM10* | *RENBP* | *RGAG1* | *RGN* | *RGS7* |
| *RLIM* | *RNASET2* | *RPGR* | *SCAPER* | *SETDB2* | *SGSH* |
| *SHANK1* | *SHROOM2* | *SLC25A22* | *SLC25A53* | *SLC25A6* | *SLC31A1* |
| *SLC6A1* | *SLC6A17* | *SMARCC1* | *SMARCC2* | *SMARCD1* | *SMARCD2* |
| *SMARCD3* | *SNTG1* | *SPG11* | *SPRY3* | *SPTLC2* | *SREBF2* |
| *SRPX2* | *ST3GAL3* | *STAB2* | *STAG1* | *STARD8* | *SYNCRIP* |
| *SYT1* | *SYTL4* | *SYTL5* | *TAF1* | *TAF2* | *TAF7L* |
| *TANC2* | *TBC1D8B* | *TCEAL3* | *TCP10L2* | *TENM1* | *THAP1* |
| *THOC2* | *ThumpD1* | *TKTL1* | *TLR8* | *TM4SF2* | *TMEM132E* |
| *TMEM135* | *TMLHE* | *TNKS2* | *TNPO2* | *TREX2* | *TRIO* |
| *TRMT1* | *TSC22D3* | *TSEN2* | *TSEN34* | *TSEN54* | *TTI2* |
| *TUBA8* | *TUBAL3* | *UBR7* | *UBTF* | *USP27X* | *USP9X* |
| *UTP14A* | *VAMP7* | *VRK1* | *WAC* | *WDR13* | *WDR45L* |
| *WNK3* | *WWC3* | *XIAP* | *XKRX* | *YY1* | *ZBTB40* |
| *ZC3H14* | *ZCCHC12* | *ZCCHC8* | *ZDHHC15* | *ZFX* | *ZMYM3* |
| *ZMYM6* | *ZMYND12* | *ZNF238* | *ZNF425* | *ZNF526* | *ZNF711* |
| *ZNF81* | | | | | |

**Table 3 - 3: List of 361 sequenced intellectual disability-associated genes that are candidates.**

Genes were allocated as candidate if they are not in a stringent, manually curated list of known ID-associated genes from a recently published study (62).

## 3.3 Results

### 3.3.1 Targeted resequencing of 565 intellectual disability-associated genes in cases and controls

The coding regions of a set of 565 known or candidate ID-associated genes were sequenced in 996 individuals (94% male) with moderate to severe, sporadic ID. This was a subset of a large replication study of seven rare diseases, comprising a total of 2812 individuals, which was carried out within the UK10K study (www.UK10K.org). The phenotypes studied were CHD, ciliopathy, coloboma, ID, neuromuscular disease, severe insulin resistance, and congenital thyroid disease, along with internal technical control samples. Coding regions of a total of 1189 genes (of which 565 are known or candidate ID-associated genes) were selected using a custom pull-down approach, then sequenced on the Illumina HiSeq 2000 platform.

The 565 ID-associated genes included known genes in which pathogenic variants in multiple unrelated individuals have been shown to cause ID, and also candidate genes selected, for example, because a variant has been identified in a single patient with ID, or because the gene is in the same family as known ID-associated genes. Some recently published studies of ID have larger lists of ID-associated genes (62). This is because some new ID-associated genes have been identified since the design of our study, and also because of restrictions on the size of targeted regions imposed by the pull-down method. I classified the 565 genes into known and candidate genes.

### 3.3.2 The sequencing data are of good quality

There are around 1500 coding SNVs and 50 coding indels per sample in this study that pass standard quality control filters (Figure 3-1). The mean depth of variant coverage per sample is 40.55X (Figure 3-1c). This is higher than the minimum 30X estimated to be required for accurate detection of heterozygous variants (87). Dr James Floyd calculated these figures.

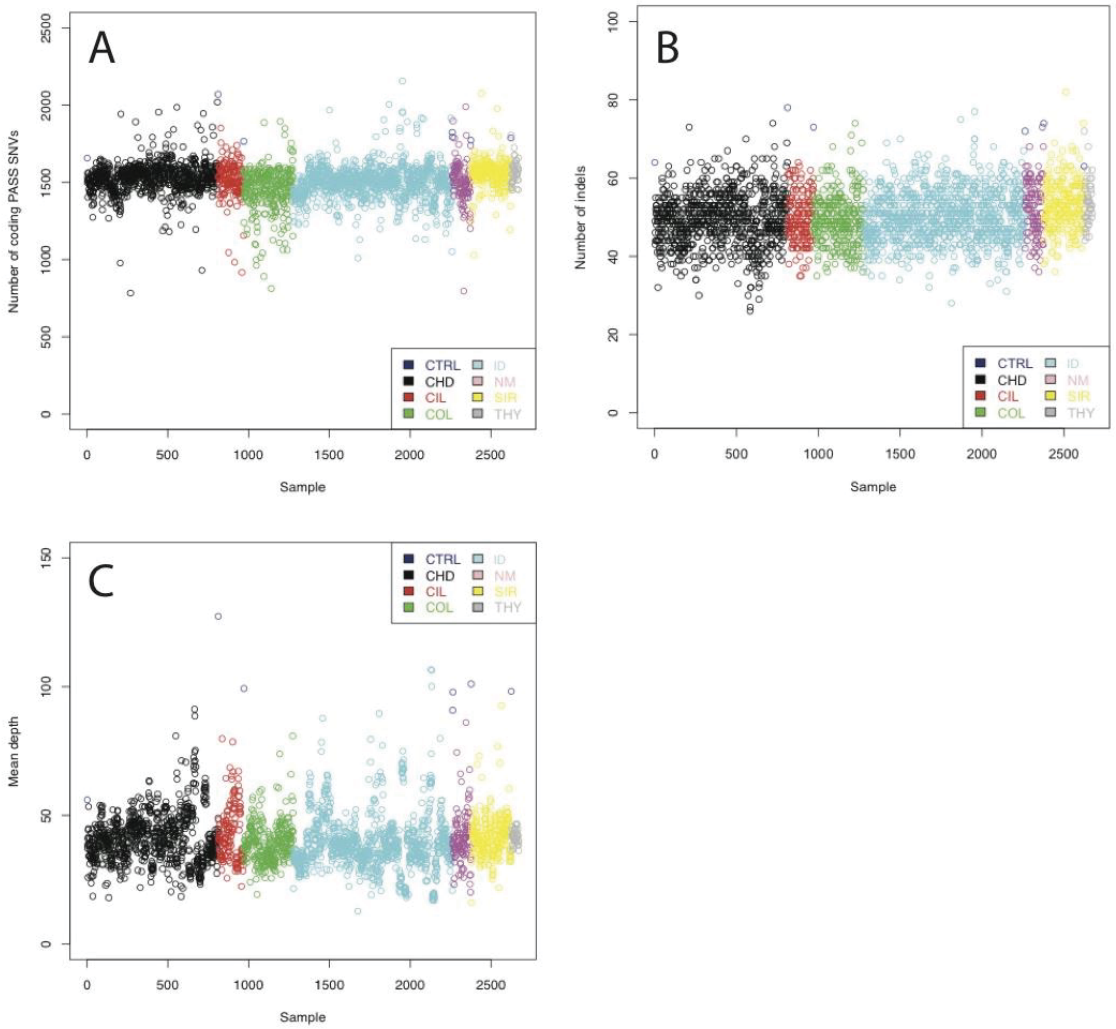


**Figure 3 - 1: Quality control metrics for the UK10K targeted resequencing study.**
**A)** Number of pass, coding SNVs per sample, across 1189 sequenced genes. **B)** Number of coding indels per sample, across 1189 sequenced genes. **C)** Mean depth of variant coverage for each replication sample. CTRL = controls, CHD = congenital heart disease, CIL = ciliopathy, COL = coloboma, ID = intellectual disability, NM = neuromuscular disorders, SIR = severe insulin resistance, THY = thyroid disease. Numbers and plots generated by Dr James Floyd, and included here with permission.

### 3.3.3 There is no substantial difference in population structure between the intellectual disability and congenital heart disease cohorts

Identification of an enrichment of predicted damaging variants in selected disease-associated genes in individuals with the disease of interest, compared to controls, often leads to insights about the disease pathology (14). I hypothesised that there might be an excess of variants in sequenced ID-associated genes in the ID part of the UK10K rare disease cohort, compared to controls. For controls, I selected the CHD cohort. This is an appropriate control because the two cohorts are of similar size, and have minimal overlap in phenotypic spectra. The CHD DNA samples had been treated, stored, amplified, sequenced, and analysed in an identical manner to those of the ID cohort.

However, population stratification between cohorts can lead to spurious findings in case-control analyses (189). The majority of the ID and CHD cohorts reported as being of European ancestry. Nevertheless, to find out whether there was a substantial difference in population structure between the two cohorts I used PCA. This is a widely used method for this purpose (184). I performed PCA on the ID and CHD samples, along with a subset of unrelated HapMap3.3 samples, using the SNPrelate package (195).

The first two principal components were sufficient to cluster the HapMap samples into their four component populations (Figure 3-2). The data points for the UK10K ID cases and CHD controls overlie each other, suggesting that there is no substantial difference in population structure between these cohorts. Additionally, they overlap to a large extent with the data points from the HapMap samples of European ancestry, confirming that the majority of both the ID cases and CHD controls are of European ancestry.

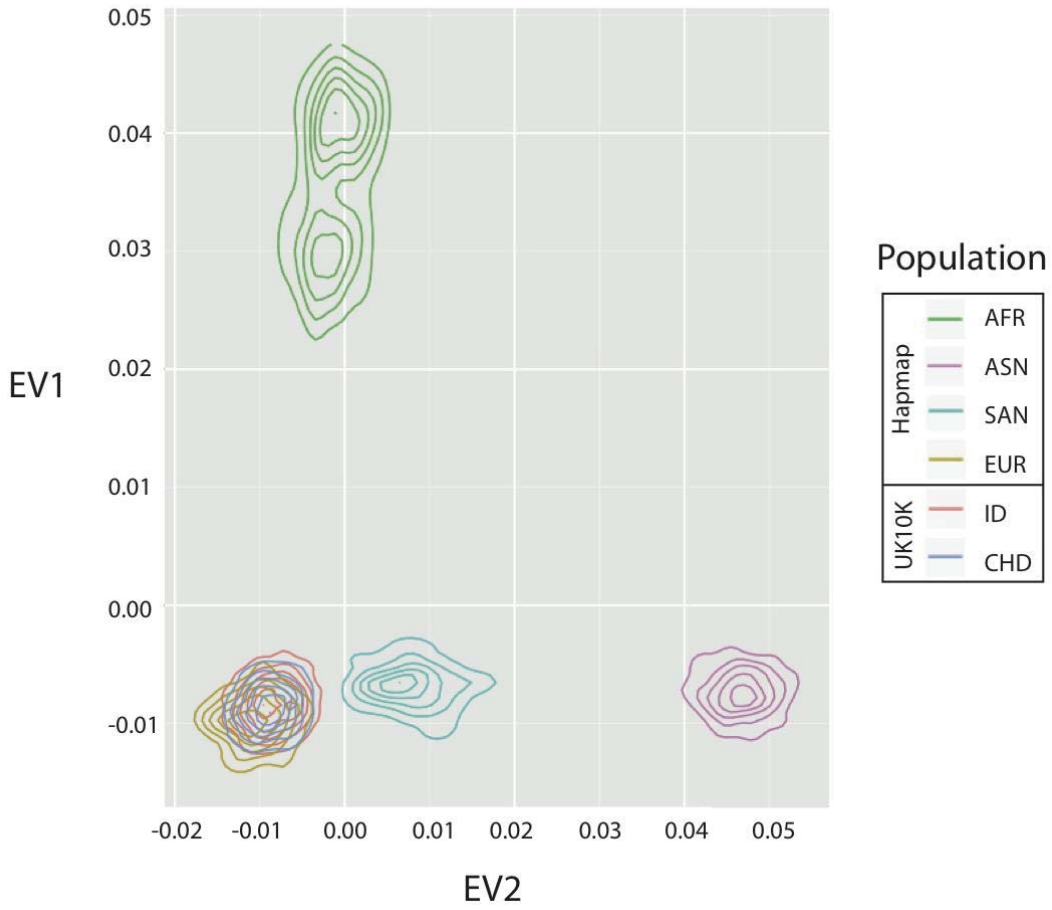


**Figure 3 - 2: Principal component analysis.**
The first two eigenvectors (EVs) cluster the HapMap3.3 samples into their component populations (AFR = individuals of African ancestry; ASN = individuals of East Asian ancestry; SAN = individuals of South Asian ancestry; EUR = individuals of European ancestry) (195). The UK10K ID and CHD samples overlie with each other, and overlap with the European HapMap3.3 samples.

### 3.3.4 14% of intellectual disability patients have a likely causative variant in a sequenced intellectual disability-associated gene

I wrote a set of R scripts to generate a list of rare, high quality, coding variants in ID-associated genes from the merged VCF files. This list contained 9015 variants, of which 8476 were functional (8389 missense; 70 in-frame indels; and 17 variants resulting in loss of a stop codon) and 539 in total were LOF (221 nonsense; 189 frameshift; 77 essential splice donor; and 52 essential splice acceptor) (Figure 3-3). The average number of LOF variants per person was 0.54, while the average number of missense variants per person was 9.05.

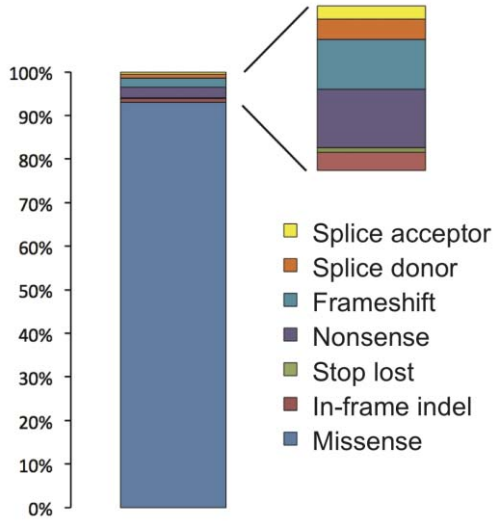


**Figure 3 - 3: Classes of variant identified through the R filtering pipeline.**
Total number of variants = 9015, total number of samples = 996.

Dr Lucy Raymond and Dr Detelina Grozeva imposed further stringent filters on this list of 9015 variants, to identify variants that are highly likely to be causative. These filters took into account factors such as the type of the variant, frequencies in the public and internal databases, presence in the human gene mutation database (http://www.hgmd.org/), consistency with the estimated mode of inheritance based on the Developmental Disorder Gene2Phenotype (DDG2P) gene list (https://decipher.sanger.ac.uk/), and the clinical phenotype of the affected individual. They validated a subset of the variants using either Sanger sequencing or exome sequencing of non-amplified DNA. Using this case-only diagnostic analytical approach, they found that 109 individuals (10.9%) had likely causative LOF variants, and 34

individuals (3.4%) had likely causative missense variants, giving a total estimated diagnostic yield of ~14%.

### 3.3.5 *SETD5* is a novel intellectual disability-associated gene

To identify novel ID-associated genes, Dr Lucy Raymond and Dr Detelina Grozeva focused on genes that had the highest number of LOF variants in the list that I generated. They found that seven individuals had a rare, high-quality, LOF variant in *SETD5* (0.7% of the cohort). They confirmed all the variants using Sanger sequencing, and confirmed that five are *de novo* by Sanger sequencing of parental DNA (paternal DNA was unavailable for two probands). I calculated that the probability of this occurring by chance in a cohort of this size is very low ($p = 5.25 \times 10^{-9}$). The mutations in *SETD5* were all different, and only one LOF mutation (which was more 3' than any identified in these ID patients) was listed in the NIH Heart, Lung, and Blood Institute's Exome Variant Server (NHLBI EVS). No other candidate gene was confirmed as being a novel ID-associated gene using this approach, because they either had a high number of LOF mutations listed in the NHLBI EVS database, or the variants were all the same, increasing the chances they are in fact a sequencing error (Table 3-4).

| Gene | Total number LOFs | Number Independent LOFs | Number NHLBI EVS LOF | Reason excluded from further analysis |
|---|---|---|---|---|
| *DCHS2* | 22 | 9 | 13 | High number LOFs in NHLBI EVS |
| *SETD5* | 7 | 7 | 1 | NA |
| *MIB1* | 9 | 7 | 13 | High number LOFs in NHLBI EVS |
| *STAB2* | 6 | 6 | 12 | High number LOFs in NHLBI EVS |
| *PCDH10* | 7 | 1 | 1 | Low number of independent LOFs |
| *UTP14A* | 6 | 1 | 0 | Low number of independent LOFs |

**Table 3 - 4: Candidate genes with the highest number of LOF variants.**
Table includes candidate genes not listed as ID-associated in OMIM. Table is sorted according to number of independent LOFs. Data courtesy of Dr Detelina Grozeva.

An international team of collaborating clinicians documented and compared the phenotypes of the seven patients with *SETD5* mutations. In addition to ID, there were several common and recurring features including ritualised behavior or ASD, abnormal ears, eyebrows, eyes, and nose, and skeletal and gastrointestinal abnormalities. They noticed that the facial appearance of the cases was, in some aspects, strikingly similar (Figure 3-4). Due to the phenotypic similarity of the cases, and the small probability of this many mutations occurring by chance, we concluded that these LOF mutations in *SETD5* are causative in these seven patients, and that LOF of *SETD5* causes a potentially recognisable syndrome. Indeed, LOF of *SETD5* may be a relatively common cause of ID (191).


**Figure 3 - 4: Facial appearance of individuals with *SETD5* mutations.**
Photographs of the seventh patient were unavailable. This figure is courtesy of Dr Lucy Raymond, and it has been published (191).

### 3.3.6 Individuals with intellectual disability have an enrichment of loss of function variants in sequenced ID-associated genes, compared to controls

I used the CAST method to assess the extent to which LOF variants in sequenced ID-associated genes are enriched in the ID cohort compared to the CHD cohort. I selected CAST rather than one of the other methods such as the weighted sum method or

SKAT, because according to the DDG2P list the mechanism of the vast majority of known ID-associated genes is loss of or reduction of protein function, so I think that the vast majority of causative variants in this cohort will have the same direction of effect.

First I excluded samples that had an excessive number of LOF variants (>4, which is >3.5 standard deviations from the mean number of variants per sample). Of the 986 ID samples remaining, 341 (34.6%) had at least one rare (internal frequency <1%) LOF variant in a sequenced ID-associated gene, compared to 225/903 (24.9%) in CHD. This represents a highly significant enrichment ($p = 2.8 \times 10^{-6}$) (Figure 3-5). This difference between the cohorts is most likely accounted for by the fraction of LOF variants that are causative of ID, suggesting that ~10% of ID cases in this cohort are caused by LOF variants in the sequenced genes. This is very consistent with the manual case-only diagnostic analysis, in which 109 (10.9%) cases were found to be caused by LOF variants.

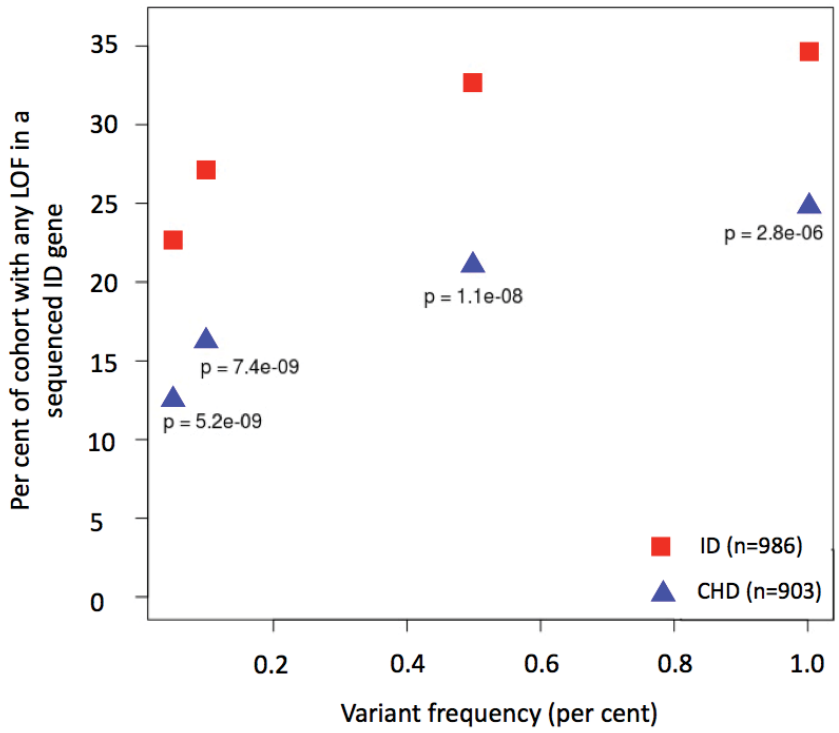


**Figure 3 - 5: Patients with intellectual disability have an enrichment of loss of function variants in sequenced intellectual disability-associated genes compared to controls.**
LOF = loss of function; ID = intellectual disability; CHD = congenital heart disease. Numbers in key show number of samples. P values were calculated by one-tailed Fisher's exact test.

I next applied more stringent internal variant frequency filters of 0.5%, 0.1% and 0.05%, the latter of which leaves unique variants only. Of the 986 ID samples, 223 (22.6%) had at least one unique LOF variant in a sequenced ID-associated gene, compared to 113/903 (12.5%) in CHD. Therefore, after application of this more stringent filter, the difference of around 10 percentage points between the cohorts is maintained, and the enrichment of LOF variants in ID becomes more significant ($p = 5.2 \times 10^{-9}$). This suggests that the vast majority of the LOF variants that cause ID in this cohort are unique within the cohort.

The LOF variants can be categorised according to chromosome, variant type, whether the sequenced ID-associated gene is known or a candidate, and whether it causes disease according to a biallelic or a non-biallelic mode of inheritance. I performed the CAST test to evaluate the degree of enrichment of each of these categories of unique LOF variants in the ID cohort (Table 3-5).

| Gene category | | Variant type | Number LOFs ID | Number LOFs CHD | P-value |
|---|---|---|---|---|---|
| Autosome or PAR | Known non-biallelic *76* | SNV | 42/986 (4.26%) | 8/903 (0.89%) | $1.922 \times 10^{-6}$* |
| | | Indels | 14/986 (1.42%) | 7/903 (0.78%) | 0.132 |
| | Known biallelic *52* | SNV | 25/986 (2.54%) | 16/903 (1.77%) | 0.164 |
| | | Indels | 15/986 (1.52%) | 6/903 (0.66%) | 0.058 |
| | Candidate *212* | SNV | 67/986 (6.8%) | 32/903 (3.54%) | $9.795 \times 10^{-4}$* |
| | | Indels | 33/986 (3.35%) | 30/903 (3.32%) | 0.54 |
| X chromosome (males only) | Known *76* | SNV | 13/925 (1.41%) | 0/467 (0%) | 0.0048* |
| | | Indels | 11/925 (1.19%) | 0/467 (0%) | 0.011 |
| | Candidate *149* | SNV | 14/925 (1.51%) | 2/467 (0.43%) | 0.056 |
| | | Indels | 7/925 (0.76%) | 1/467 (0.21%) | 0.191 |

**Table 3 - 5: Enrichment of unique LOF variants in the ID cohort, split by category.**
The numerator in the 'Number LOFs ID' and 'Number LOFs CHD' columns show the number of samples in each cohort that have one of more unique LOF variant of the category indicated. The number of genes in each category is given in italics. PAR = pseudo-autosomal region; SNV = single nucleotide variant; LOF = loss of function variant, ID = intellectual disability cohort; CHD = congenital heart disease control cohort. P values calculated using Fisher's exact test. *Below Bonferroni-corrected threshold of 0.005.

LOF SNVs in autosomal, known ID-associated genes with non-biallelic mode of inheritance are significantly enriched in the ID cohort ($p = 1.922 \times 10^{-6}$). In contrast, I identified no significant enrichment in known ID-associated genes with biallelic mode of inheritance ($p = 0.164$). Given that the parents of the probands in this cohort are unaffected, this suggests that dominant, *de novo* mutations are an important cause of disease in our cohort. This is consistent with studies showing that *de novo* LOF mutations are a particularly important cause of ID (57, 84).

Furthermore, LOF SNVs in autosomal, candidate ID-associated genes are significantly enriched in the ID cohort ($p = 9.795 \times 10^{-4}$). This very strongly suggests that some of these candidate genes are real ID-associated genes, even though they have not yet been definitively proved as such. Unfortunately, I could not use the CAST test to identify the individual candidate genes that were driving this signal, because relatively small cohort sizes and effect sizes render the CAST test underpowered for this purpose. Additionally, LOF SNVs in X-linked, known ID-associated genes in males are significantly enriched in the ID cohort ($p = 0.0048$). Interestingly, I identified no significant enrichment in X-linked candidate genes ($p = 0.056$). This suggests that, compared to the autosomes, a higher proportion of ID-associated genes on the X chromosome have been identified. This is unsurprising, as the X chromosome has been disproportionately well studied in ID (13). I did not detect any significant enrichment of LOF indels, which is likely due to reduced sensitivity of indel calling programs compared to SNVs.

There is no significant enrichment of synonymous variants in sequenced ID-associated genes in the ID cohort compared to CHD ($p = 0.475$). Subcategorising the synonymous variants reveals no significant enrichment in any category (data not shown). This is important because if the enrichment of missense variants was a spurious result due to a difference in the cohorts such as population stratification, one would expect to see an equivalent enrichment in synonymous variants. This finding therefore increases the chance that the observed enrichment is real and biologically relevant.

### 3.3.7 In known ID-associated genes on the X chromosome, unique missense variants tend to be more damaging in ID patients than controls.

To test the hypothesis that unique, missense variants in sequenced ID-associated genes are more likely to be damaging in the ID cohort than the CHD cohort, I

compared the distribution of PolyPhen2, SIFT, and Condel scores using one-tailed, unpaired Mann-Whitney tests (65, 66, 196). I excluded samples with excessive numbers of missense variants (>25, which is >3.5 standard deviations from the mean number of variants per sample), and individuals in the ID cohort for whom a clearly causal LOF variant had been identified, from this analysis. For all scores, the only category of variant where there was a significant difference between the cohorts was missense variants in X-linked, known ID-associated genes. In this category, variants in ID cases were predicted to be significantly more damaging than those in controls ($p < 0.0001$) (Figure 3-6), suggesting that a proportion of this category of missense variant do indeed cause ID. In contrast, in known ID-associated genes on the autosomes, there is no difference in scores of predicted damage of unique missense variants between ID patients and controls (Figure 3-7).

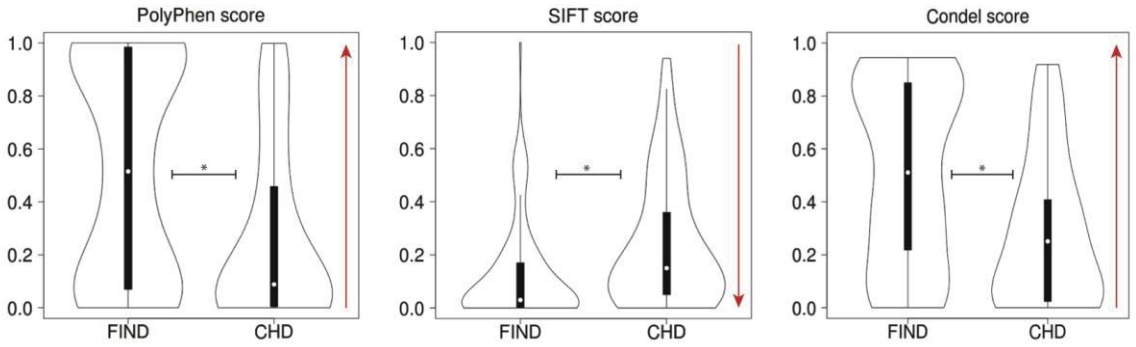


**Figure 3 - 6: In known ID-associated genes on the X chromosome, unique missense variants are predicted to be more damaging in ID patients than controls.**
The number of samples (ID = 825; CHD = 466) does not include those with excessive numbers of missense variants (>25), or ID samples with causative LOF variants identified. These plots consist of 154 missense variants in known ID-associated genes for the ID cohort, and 62 for the CHD cohort. * = p < 0.0001, calculated by Mann-Whitney tests. The red arrow on each plot indicates the direction of increase in predicted damage.

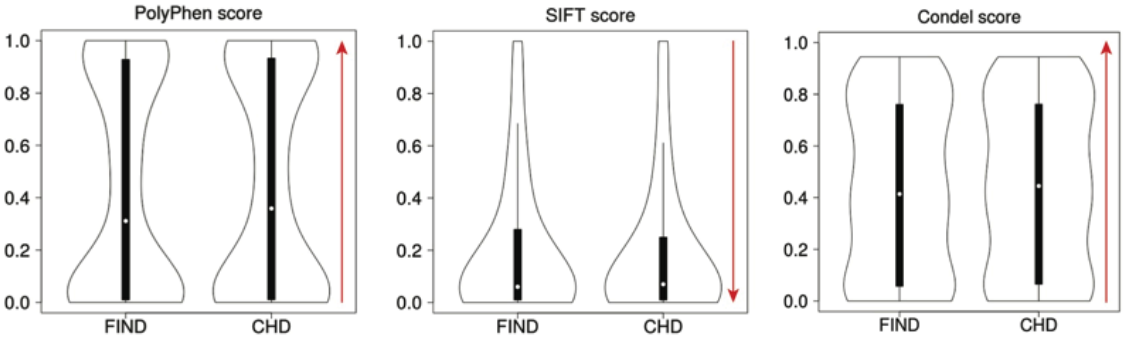


**Figure 3 - 7: In known ID-associated genes on the autosomes, unique missense variants are not predicted to be more damaging in ID patients than controls.**
The number of samples (ID = 877; CHD = 900) does not include those with excessive numbers of missense variants (>25), or ID samples with causative LOF variants identified. These plots consist of 1184 missense variants in known ID-associated genes for the ID cohort, and 1039 for the CHD cohort. There is no significant difference in scores of predicted damage between ID cases and controls (Mann-Whitney test, p > 0.66). The red arrow on each plot indicates the direction of increase in predicted damage.

### 3.3.8 Evidence for an enrichment of unique, predicted damaging, missense variants in sequenced ID-associated genes in the ID cohort

One reason that detecting an enrichment of missense variants in case-control analyses is harder than for LOF variants is that a smaller proportion of missense than LOF variants cause disease. Therefore, any enrichment of damaging missense variants in

the ID cohort could be masked by the 'noise' of benign missense variants. I therefore applied the CAST test to unique missense variants that are predicted to be damaging by at least one of the three scores of predicted damage, in order to assess the possible contribution of causal missense variants in our cohort. I first excluded samples with excessive numbers of missense variants (>25). In the ID cohort, I also excluded the 109 samples for which a clearly causal LOF variant had been identified.

Of the ID samples, 691/877 (78.8%) had at least one unique, predicted damaging, missense variant in a sequenced ID-associated gene, compared to 688/900 (76.4%). This does not represent a significant enrichment (p = 0.129). However, two of the subcategories do have a significant enrichment (Table 3-6). Of the ID samples, 438/877 (49.9%) had at least one unique, predicted damaging, missense variant in a candidate autosomal gene compared to 393/900 (43.7%) CHD samples (p = 0.005), suggesting that around 6% of cases in our cohort might be caused by this category of variant. This suggests that variants in a subset of these candidate genes can indeed cause ID, which is consistent with the results of the CAST test on LOF variants. It is interesting that there is a more significant enrichment for candidate than known ID-associated genes. This could be a consequence of there being more candidate than known genes, or it could be that a higher proportion of candidate than known ID-associated genes operate by a non-LOF mechanism.

| Gene category | | Number missense ID | Number missense CHD | P-value |
|---|---|---|---|---|
| Autosome or PAR | Known non-biallelic *76* | 258/877 (29.4%) | 232/900 (25.8%) | 0.048 |
| | Known biallelic *52* | 233/877 (26.6%) | 213/900 (23.7%) | 0.088 |
| | Candidate *212* | 438/877 (49.9%) | 393/900 (43.7%) | 0.005* |
| X chromosome (males only) | Known *76* | 86/825 (10.4%) | 17/466 (3.6%) | $4.65 \times 10^{-6}$* |
| | Candidate *149* | 169/825 (20.5%) | 78/466 (16.7%) | 0.057 |

**Table 3 - 6: Enrichment of unique, predicted damaging, missense variants in the ID cohort, split by category.**
The numerator in the 'Number missense ID' and 'Number missense CHD' columns show the number of samples in each cohort that have one or more unique, predicted damaging missense variant of the category indicated. The number of genes in each category is given in italics. The number of total samples does not include those with excessive numbers of missense variants (>25), or ID samples with causative variants identified. PAR = pseudo-autosomal region; ID = intellectual disability cohort; CHD = congenital heart disease control cohort. P values calculated using Fisher's exact test. *Below Bonferroni-corrected threshold of 0.01.

Up to 7% of cases in our cohort might be caused by unique, predicted damaging, missense variants in known ID-associated genes on the X chromosome, because 86/825 (10.4%) males in the ID cohort have at least one, compared to 17/466 (3.6%) in the CHD cohort ($p = 4.65 \times 10^{-6}$).

# 3.4 Discussion

### 3.4.1 Summary

A targeted resequencing study was carried out as part of the UK10K project; 565 ID-associated genes were sequenced in 996 ID patients. I generated a list of rare, high quality, coding variants in the ID-associated genes in this cohort. From these data, causative variants were identified for ~14% of the cohort, and the novel ID-associated histone methyltransferase gene *SETD5* was identified. I next confirmed that there is no substantial difference in population structure between the ID cases and controls with CHD, and I used CAST to identify a highly significant enrichment of unique LOF variants in ID-associated genes in cases compared to controls. The size of the burden was consistent with the findings of the case-only diagnostic analysis. I subcategorised the LOF variants according to features of the variant itself, and features of the gene that it affects. From this, I found that the enrichment is greater in known than candidate genes, it is greater in genes with a non-biallelic rather than a biallelic mode of inheritance, and it is greater in SNVs than indels. I extended the analysis to missense variants. There was lower power to detect enrichment in missense variants, because a lower proportion of them are casual. Nevertheless, I found a moderately significant enrichment of missense variants in candidate autosomal genes, and a highly significant enrichment in known ID-associated genes on the X chromosome. This is consistent with the observation that missense variants in known ID-associated genes on the X chromosome are, on average, predicted to be significantly more damaging in ID cases than controls with CHD.

### 3.4.2 Loss-of-function of the histone methyltransferase gene *SETD5* is probably responsible for the cardinal features of 3p25 microdeletion syndrome

In this study, we showed for the first time that *de novo* LOF mutations in the histone methyltransferase gene *SETD5* cause ID, along with additional phenotypes such as ritualised behaviour, and dysmorphic facial features (191). In our cohort, this was a relatively frequent cause of disease, accounting for 0.7% of cases, which is similar to the frequency of *ARID1B* mutations, which are considered to be one of the more common causes of sporadic ID (173).

There are three reasons why *SETD5* was selected as a candidate ID-associated gene to be sequenced in this study. First, a *de novo* LOF mutation in *SETD5* was reported in a single ID patient in a previous study (84). While intriguing, this was not sufficient for Rauch *et al.* to conclude that *SETD5* is definitely an ID-associated gene, and the authors did not extensively report the phenotype of this patient. Second, *de novo SETD5* mutations have been associated with ASD in several studies, and it is widely known that there is much overlap in the presentation and genetic aetiology of ID and ASD (197-199). Third, *SETD5* is one of only two protein-coding genes in the minimal critical region for 3p25 microdeletion syndrome (200).

The 3p25 microdeletion syndrome was first described in 1978 (201). Since then there were several other case reports of *de novo* deletions at this locus, resulting in phenotypes including ID, seizures, microcephaly, CHD, malformed ears and nose, and other dysmorphic craniofacial features (202-204). The sizes and breakpoints of the deletions in these cases varied, and so the minimum critical region was refined over time. Most recently, a case report refined it to only 124 kb, containing only three genes: *THUMPD3*, *SETD5*, and *LOC440944* (an RNA gene) (200).

The phenotypes of the patients with *SETD5* mutations described in this study are very similar to those of the patients with 3p25 microdeletion syndrome (191). Phenotypes that overlap in both groups include ID, abnormal eyebrows, a depressed nasal bridge, large or low-set ears, a long smooth philtrum, OCD or ritualised behaviour, skeletal abnormalities, and CHD. With the exception of ID, these phenotypes are variable, appearing in multiple, but not all, cases. The overlap between the two groups is not complete; for example, none of the patients in our study had seizures or microcephaly, which are features of some cases of 3p25 microdeletion syndrome. Therefore, while the possibility that haploinsufficiency of 3p25 genes other than *SETD5* might contribute to the clinical phenotype in some patients cannot be excluded, it appears highly likely that haploinsufficiency of *SETD5* is responsible for the cardinal features of 3p25 microdeletion syndrome.

One study of CNVs in patients with ASD came to a different conclusion. Pinto *et al.* identified a 24 kb deletion encompassing most of *SETD5* and no other genes in a single patient with ASD and borderline ID, but no other medical issues or dysmorphic features (199). They therefore suggest that while LOF of *SETD5* may be at least partially responsible for the intellectual and behavioural deficits of 3p25 microdeletion syndrome patients, it is probably not involved in the other features of the syndrome.

Pinto *et al.* was published before the *SETD5* study, so the authors were unaware of the seven patients described here (191). It is more likely that some form of genetic compensation explains the mild phenotype in their single patient, than that the overlapping phenotypes in the seven patients with *SETD5* mutations in this study are coincidental.

Interestingly, ID caused by *SETD5* mutations is another example of a clearly syndromic form of ID that is not recognised as such until a group of patients with shared aetiology are retrospectively examined together. This emphasises the importance of assembling groups of patients with shared aetiology. *SETD5* can also be added to the list of ID-associated genes that were discovered after being identified as candidates because they are in a CNV. Historically, this has been an important way to identify ID-associated genes, particularly in autosomes. Other ID-associated genes that were identified this way include *MBD5* and *KANSL1* (156, 157).

Before describing variants in a gene as causative of any rare disease, it is important to apply a high and consistent standard to the evidence assembled to support the assertion. For example, a rare variant that segregates with Mendelian disease in a single family is not necessarily causative (12). As sample sizes and the amount of sequencing data increases, the probability of finding recurrent similar variants in a given gene just by chance also increases. Therefore, it is also important to apply statistical tests to demonstrate that the variants in question are significantly enriched in patients. Furthermore, if LOF variants in a given gene are relatively common in the general population it is unlikely that LOF of that gene causes a rare disease. Several ID-associated genes have recently been called into question on this basis (64). Therefore in this study, my colleagues and I took care to apply a high standard of evidence to the data, before concluding that *SETD5* is a novel ID-associated gene. For example, we showed that LOF of *SETD5* in the general population is very rare, and we showed that the mutations identified were highly unlikely to have occurred by chance (191).

*SETD5* is predicted on the basis of sequence homology to encode a histone methyltransferase (205). As well as *SETD5* and *EHMT1* (pathogenic variants in which can cause Kleefstra syndrome as discussed) known ID-associated histone methyltransferases include *EZH2* and *MLL2* (also known as *KMT2D*). EZH2 is part of a complex that methylates a specific lysine residue on histone H3 (206). It has many important roles in development, including X chromosome inactivation, and stem cell

regulation (207, 208). *De novo* mutations in *EZH2* can cause Weaver syndrome, features of which include ID, overgrowth, and characteristic craniofacial dysmorphic features (209). *MLL2*, mutations in which can cause Kabuki syndrome, which also involves ID, and also catalyses methylation of histone lysine residues (210). Therefore, although little more is known about the function of *SETD5*, histone methyltransferases are clearly emerging as a very important class of ID-associated genes. *SETD5* fits well into the known pattern for ID-associated histone methyltransferases, because all known causative mutations are *de novo*, and the resulting phenotype is syndromic. These two features are consistent with all the other known examples of ID-associated histone methyltransferases discussed.

### 3.4.3 Insights from case-control enrichment analyses

The case-control enrichment analysis demonstrates that in this cohort, 10% of ID cases are caused by LOF variants in the sequenced genes. This is consistent with the results of the case-only diagnostic analysis, in which a causative LOF variant was identified for 10.9% of the cohort. Using case-control enrichment analysis I estimate that up to 13% of cases in this cohort are caused by unique, predicted damaging, missense variants (6% in candidate autosomal genes, plus 7% in known ID-associated genes on the X chromosome). This is much higher than the rate of causative missense variants found by manual case-only diagnostic analysis, which is only 3.6%. This suggests that the true proportion of the cohort where disease is caused by missense variants is higher than 3.6%. However, assigning pathogenicity to missense variants with a diagnostic level of confidence is more difficult than for LOF variants, and must be done conservatively.

Two previous exome sequencing studies of ID cohorts have estimated diagnostic yields of 16% and 31% respectively (57, 84). Another exome sequencing study of children with developmental disorders, many of whom had ID, had a diagnostic yield of 25% (11). Differences in ascertainment and methodology make direct comparisons of diagnostic yield between studies problematic. There are four reasons why our total estimated diagnostic yield of 14% is lower than that of the previous studies. First, we resequenced the exons of 565 known and candidate ID-associated genes only in a targeted approach, rather than sequencing all genes. Second, we sequenced probands only, not trios. Third, 94% of this UK10K ID cohort is male, whereas most other cohorts

are approximately 50% male, and it is possible that, on average, males with ID have a higher contribution from oligogenic causes. Finally, a proportion of cases in our cohort had been through extensive previous investigation, so the cohort is enriched for harder to solve cases.

Assigning causality to a novel candidate gene requires a high degree of evidence (12). In this study, the sizes of the cohorts were insufficient to have power to detect a significant enrichment of variants in individual novel candidate genes using case-control enrichment analyses. Nevertheless, the finding that there is a significant enrichment of both LOF and missense variants in candidate ID-associated genes shows that some of these variants must be causative. The enrichment of both LOF and missense variants in known genes with a non-biallelic mode of inheritance is greater than that in known genes with a biallelic mode of inheritance, which tells us that *de novo* mutations are probably an important cause of ID in our cohort, even though we did not sequence trios. These insights into the genetic architecture of the cohort highlight the utility of case-control enrichment analyses as a supplementary tool to manual case-only diagnostic analysis.

Fundamental differences between the X chromosome and autosomes may explain why the burden of missense variants is so much larger for known, ID-associated genes on the X chromosome, than for any other category of missense variants in this study. For example, a higher proportion of X chromosome genes are involved in brain development and function than autosomal genes (211-213). Given that this UK10K ID cohort is 94% male, one might therefore expect a disproportionate number of cases to be caused by pathogenic variants in the X chromosome because of this functional bias. Additionally, ID-associated genes have also been particularly well studied on the X chromosome, so a higher proportion of X-linked than autosomal ID-associated genes may have been identified (13).

Furthermore, differences between the X chromosome and autosomes may influence scores of predicted damage. Greater selection pressure acting upon the X chromosome results in less diversity on the X chromosome than autosomes (214). This also means that, in general, X chromosome genes are more conserved between species than autosomal genes (215). SIFT, for example, assesses how likely a variant is to be damaging, according to how conserved the affected locus is, with more conserved positions likely to be less tolerant to variation (66). As X chromosome genes

are generally more conserved than autosomal genes, scores of predicted damage might be, on average, higher on the X chromosome than autosomes.

Unlike classic case-only diagnostic analysis, case-control enrichment analysis takes into account variants with incomplete penetrance, and variants that contribute to a phenotype in an oligogenic manner. However, when a burden of variants is identified, it is not currently feasible to distinguish how much of the burden is caused by causative variants with complete penetrance, and how much is caused by variants with incomplete penetrance, oligogenic variants, and secondary modifiers of phenotype. Purcell *et al.* described the burden that they identified in schizophrenia candidate genes as “polygenic”, but they use the term on the population level, and do not suggest that individuals necessarily have multiple causative alleles (14). Development of statistical methods that can distinguish between these scenarios would be a very welcome future development.

### 3.4.4 Limitations of this study

The major limitation of the study design is that we employed an inherently biased, targeted gene approach, in which only 565 known and candidate ID-associated genes were sequenced. This decision was taken for financial reasons, and it meant that any causative variants in other genes could not be identified, so the diagnostic yield is almost certainly lower than what it would have been had we done exome sequencing instead, for example. Similarly, only probands were included in this study, meaning that without performing additional sequencing, *de novo* mutations could not be distinguished from inherited variants, making it more difficult to interpret the results. The list of 565 known and candidate ID-associated genes was originally compiled in 2012, so now the list is quite out of date as many additional ID-associated genes have been identified since then (62).

Regarding the case-control enrichment analysis, the result that there is no enrichment of indels in ID cases compared to controls suggests that indels are called with low sensitivity by the UK10K pipeline. Another limitation to bear in mind is that categorisation of the sequenced ID-associated genes into known and candidate genes is to some extent a false dichotomy. This is actually a complex task, and the level of stringency required to distinguish between the two categories is not something on which the ID research community has reached a clear consensus. For example, some

think that variants in a certain minimum number of unrelated cases must be identified before a gene can be classified as “known”, as opposed to “candidate”, whereas others do not think this is always necessary ((62) and personal communication from Dr Matthew Hurles). Similarly, there are genes such as *NF1* in which variants are associated with ID in a proportion of cases, but not in a high enough proportion of cases to be definitively classified as ID-associated genes (personal communication from Dr. Lucy Raymond). I decided to use a recently generated, manually curated, stringent list of known genes for this study (62). It is likely that some researchers would argue that some of the genes I have categories as “known” are actually “candidate”, and vice versa.

Finally, it is disappointing that this study was underpowered to detect an enrichment of variants in individual genes. However, it is not at all surprising, as it has previously been shown that much larger samples sizes than ~1000 cases and ~1000 controls would be required to achieve this (190).

### 3.4.5 Further work

An ongoing project that will extend the work described in this chapter is a large exome sequencing project of 1151 individuals with ID or their relatives. Importantly, 541 (47%) of these individuals were also included in the targeted resequencing study described here. Therefore, the exome sequencing study will enable us to validate findings of the targeted resequencing study, and hopefully identify more causative variants and more novel ID-associated genes. This exome sequencing study includes several different family structures such as 49 trios and 121 affected sibling pairs, which will facilitate easier interpretation of variants than single probands too, because for example *de novo* or shared variants can be identified. At the time of writing, the sequencing, mapping, variant calling, and filtering for this study has been completed, and the data are being further analysed and interpreted.

Another exciting ongoing project is the development of a *SETD5* mouse model (https://www.komp.org/geneinfo.php?MGI_Number=1920145). Dr Jacqui White of the mouse genetics programme at WTSI has led this work. Homozygous null mice are unsurprisingly lethal, but early phenotyping on a small number of heterozygous mice so far suggests that they may have interesting features, such as dysmorphic craniofacial features, including a depressed nasal bone (personal communication from Dr Jacqui

White). This appears to confirm that *SETD5* has a role in development of the mid-face and the skull, as suggested by the seven patients in our study.

Plans are underway to assess the cognitive abilities of these mice. If the mice are indeed cognitively impaired, they could be valuable experimental tools with which to identify any downstream genes whose expression is altered as a result of LOF of *SETD5*. This might be achieved, for example, by performing RNA-seq on brain tissue from heterozygous *SETD5* knockout mice, along with their wildtype siblings as controls. This might really start to demonstrate how *SETD5* mutations cause ID, and could even ultimately lead to identification of therapeutic targets.

The most important outcomes of the work described in this chapter are as follows. We have identified a genetic diagnosis for ~14% of the ID patients in this UK10K cohort. We have identified *SETD5* as a novel ID-associated gene, supporting the importance of histone methyltransferases in the aetiology of ID. Additionally, we have demonstrated that LOF of *SETD5* is probably responsible for the cardinal features of 3p25 microdeletion syndrome. Finally, certain categories of variants are enriched in ID-associated genes in ID cases compared to controls, yielding insights into the genetic architecture of ID, and demonstrating the utility of case-control enrichment analysis as a supplementary analytical approach in large genomic studies of rare disease.