Contribution of rare and common variants to intellectual disability in a high-risk population sub-isolate of Northern Finland

Mitja I Kurki[1,2,3], Elmo Saarentaus[3], Olli Pietiläinen[2,4], Padhraig Gormley[1,2], Dennis Lal[1,2], Sini Kerminen[3], Minna Torniainen-Holm[3,5], Eija Hämäläinen[3], Elisa Rahikkala[6,7,8], Riikka Keski-Filppula[6,7,8], Merja Rauhala[9], Satu Korpi-Heikkilä[9], Jonna Komulainen–Ebrahim[10], Heli Helander[10], Päivi Vieira[10],Veikko Salomaa[5], ,Matti Pirinen[3], Jaana Suvisaari[5], , Jukka S Moilanen[6,7,8] ,Jarmo Körkkö[9], Outi Kuismin[3,6,7,8], Mark J Daly†[1,2,3,11,12], Aarno Palotie†*[1,2,3,11,12]

1 Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA;

2 The Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA;

3 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland;

4 Department of Stem Cell and Regenerative Biology, University of Harvard, Cambridge, Massachusetts, USA;

5 National Institute for Health and Welfare, Helsinki, Finland

6 PEDEGO Research Unit, University of Oulu, Oulu, Finland

7 Medical Research Center, Oulu University Hospital, University of Oulu, Oulu, Finland

8 Department of Clinical Genetics, Oulu University Hospital, Oulu, Finland

9 Northern Ostrobothnia Hospital District, Center for Intellectual Disability Care, 90220 Oulu, Finland

10 Department of Children and Adolescents, Oulu University Hospital, Medical Research Center Oulu, University of Oulu, Oulu, Finland;

11 Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.

12 Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

* Corresponding author

†These authors jointly supervised this study

Abstract

The contribution of *de novo* and ultra-rare genetic variants in severe and moderate intellectual disability (ID) has been extensively studied whereas the genetic architecture of mild ID has been less well characterized. To elucidate the genetic background of milder ID we studied a regional cohort of 442 ID patients enriched for mild ID (>50%) from a population isolate of Finland. We analyzed rare variants using exome sequencing and CNV genotyping and common variants using common variant polygenic risk scores. As controls we used a Finnish collection of exome sequenced (n=11311) and GWAS chip genotyped (n=11699) individuals.

We show that rare damaging variants in genes known to be associated with cognitive defects are observed more often in severe (27%) than in mild ID (13%) patients (p-value: 7.0e-4). We further observed a significant enrichment of protein truncating variants in loss-of-function intolerant genes, as well as damaging missense variants in genes not yet associated with cognitive defects (OR: 2.1, p-value: 3e-8). For the first time to our knowledge, we show that a common variant polygenic load significantly contributes to all severity forms of ID. The heritability explained was the highest for educational attainment (EDU) in mild ID explaining 2.2% of the heritability on liability scale. For more severe ID it was lower at 0.6%. Finally, we identified a homozygote variant in the CRADD gene to be a cause of a specific syndrome with ID and pachygyria. The frequency of this variant is 50x higher in the Finnish population than in non-Finnish Europeans, demonstrating the benefits of utilizing population isolates in rare variant analysis of diseases under negative selection.

Introduction

Intellectual disability (ID) is a relatively common disorder characterized by deficits in both intellectual and adaptive functioning in conceptual, social and practical domains. A diagnosis of ID requires deficits in a broad range of intellectual functions, deficits in adaptive functioning resulting in failure to meet developmental and sociocultural standards for personal independence and social

responsibility, and an onset during the developmental period[1]. The population prevalence estimates of ID varies between 1%-3% and is clearly lower (<0.5%) for more severe forms of ID (IQ<50) than for mild forms [2].

The genetic contribution to severe and moderate intellectual disability has been extensively studied. While genome-wide studies using microarrays and exome sequencing have identified a prominent role of *de novo* copy number variations (CNVs), INDELs and single nucleotide variants in mostly severe ID with reported diagnostic yields of 13%-42%, their role in mild ID is less studied but expected to have a less prominent role [3,4]. Intriguingly siblings of mild ID individuals have low IQ compared to the general population whereas the IQ of siblings of severe ID individuals do not differ from the general population[5]. This suggests that mild ID represents a low extreme in a normal distribution of IQ, while severe ID is a distinct condition with different etiology[5].

The observation that intellectual disability has a high co-morbidity with other neurodevelopmental and neuropsychiatric diseases such as autism, schizophrenia and epilepsy has stimulated the hypothesis that these diseases might, in part, have shared genetic backgrounds and thus alterations in the same pathways [6].

One strategy to shed light on the genetic background of diseases is to use populations where the incidence of the trait is higher, and/or where the population history provides benefits for variant identification. Finland is a well-characterized genetic isolate where the small size of the founder population, subsequent bottleneck effects, and genetic drift have caused an enrichment of some rare and low frequency variants as compared to other European populations[7,8]. In a population with a recent bottleneck, such as Finland, variants conferring a high risk for a disease with reduced fecundity can exist at markedly higher frequencies than in older populations because negative selection has not had time to drive down the allele frequencies, and therefore these variants are easier to associate to a disease [9].

Interestingly, ID (Figure 1) and other neurodevelopmental and neuropsychiatric diseases (NDD) like schizophrenia (Supplementary Figure 1) have a higher prevalence in North-Eastern Finland as compared to South-Western Finland[10,11]. It has been hypothesized that such a pattern is related to the recent bottlenecks

of these regions. The Eastern and Northern parts of Finland were inhabited more permanently only after the internal migration of small groups in the 16th century while Southern coastal regions were already more populous (Figure 1) [12]. The regional genetic differences between the early and late settlements (east-west and north-south) can be clearly recapitulated from genome-wide common SNP data [13,14].

The aforementioned Finnish population history and the observation of geographical differences in the prevalence of neurodevelopmental diseases in Finland motivated us to initiate the Northern Finland Intellectual Disability (NFID) study, a geographically based cohort of ID patients and their family members recruited from specialty clinics in the two most Northern provinces of Finland. The only study exclusion criterion was having a known or suspected genetic or environmental cause for the phenotype and therefore the majority of our patients have the most common mild form of ID (259/498). Here we describe a comprehensive genetic characterization of 442 independent NFID patients with unknown disease etiology, enriched for mild (51.4%) forms of ID (Table 1).

We then examined the genetic architecture of this ID cohort that has undergone a population bottleneck and has a high proportion of mild ID cases. We studied the contribution of rare variants using exome sequencing, common variant polygenic risk scores, and CNVs using genome-wide association study (GWAS) arrays, each for different ID severity categories. We then compared the common and rare variants observed in the ID cohort to a large collection of pre-existing Finnish exome (n=11311) and GWAS array data (n=11699).

We also analyzed the geographical distribution of the polygenic risk variant load of educational attainment, IQ and schizophrenia across different parts of Finland. Finally, to explore the broader phenotypic impact of identified variant categories and individual variants in the NFID cohort, we compared the identified variants to 640 exome-sequenced individuals with cognitive impairment, schizophrenia (SCZ) or autism spectrum disorder (ASD).

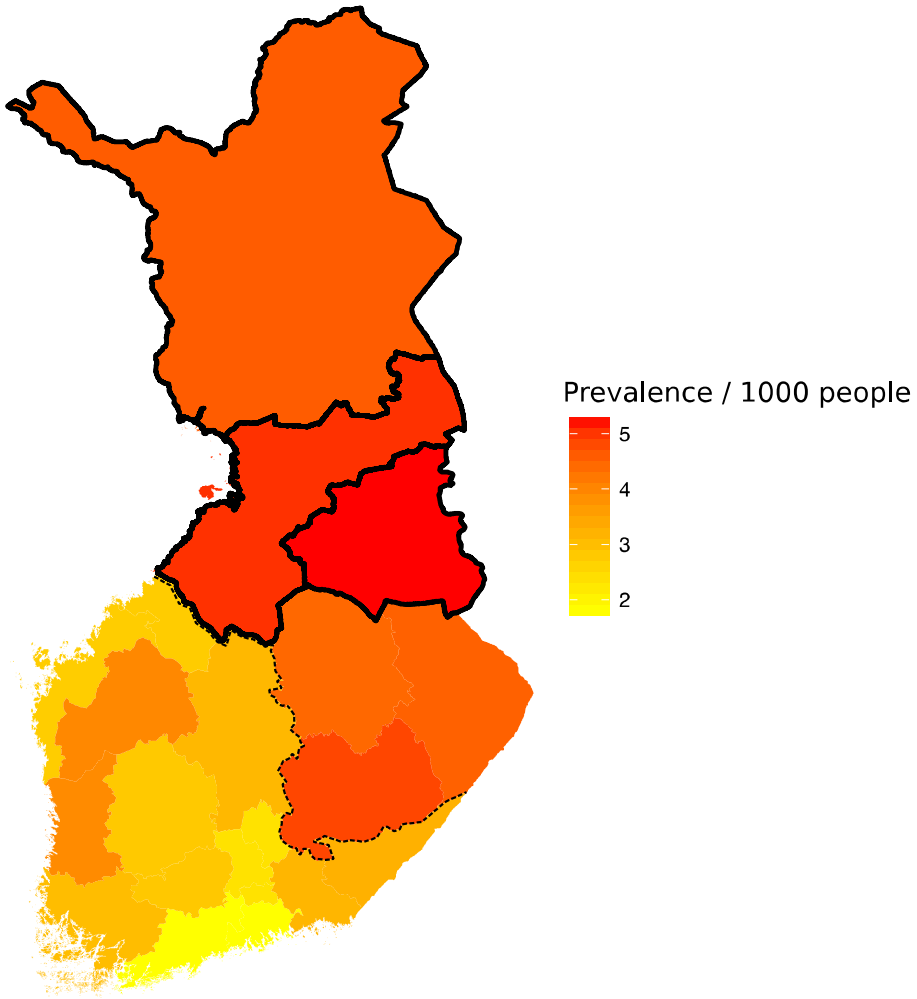

Figure 1. ID prevalence estimates in different municipalities in Finland. The primary NFID collection municipalities of Northern Ostrobothnia, Kainuu and Lapland are outlined in solid black. The approximate boundary between early and late settlements is shown with a dashed line.

Table 1. Patient clinical characteristics in the NFID sample

ID SEVERITY	MILD	MODERATE	PROFOUND	UNDEFINED
TOTAL	259	126	72	41
FEMALE %	37 %	41 %	36 %	44 %
EPILEPSY	28 (11 %)	19 (15 %)	37 (51 %)	9 (22 %)
ASD	22 (8 %)	17 (13 %)	25 (35 %)	3 (7 %)
BEHAVIORAL IMPAIRMENT	38 (15 %)	28 (22 %)	15 (21 %)	0 (0 %)
PSYCHOTIC DISORDER	24 (9 %)	16 (13 %)	8 (11 %)	0 (0 %)
DYSMORPHISM	85 (33 %)	62 (49 %)	41 (57 %)	13 (32 %)

ASD=autism spectrum disorder

Materials and Methods

Samples

Starting in January 2013 subjects for the NFID (Northern Finland Intellectual Disability) project have been recruited from the Northern Ostrobothnia Hospital District Center for Intellectual Disability Care and from the Department of Clinical Genetics of Oulu University Hospital. In January 2016 the recruitment was expanded to include all pediatric neurology units and centers for intellectual disability care in the special responsibility area of Oulu University Hospital. Subjects of all ages with either intellectual disability or pervasive and specific developmental disorders (ICD-10 codes F70-79 and F80-89, respectively) of unknown etiology were included. Individuals with copy number variations of unknown clinical significance or highly variable phenotypes were also included in order to uncover other possible factors of genetic etiology. Subjects were identified through hospital records and invited via mail to take part in the study. In addition, they were recruited during routine visits to any of the study centers. All research subjects and/or their legal guardians provided a written informed consent to participate in the study. DNA samples from the participants were extracted primarily from peripheral blood. In a few sporadic cases where a blood sample could not be obtained, DNA was extracted from saliva. The ethical

committees of the Northern Ostrobothnia Hospital District and the Hospital District of Helsinki and Uusimaa reviewed and approved the study.

Clinically performed diagnostic examinations/tests varied considerably depending on the subject´s age, clinical diagnosis and phenotype. During the past 20 years, blood and urine metabolic screening tests, chromosome karyotyping, FMR1 CGG repeat analysis, electroencephalography (EEG) and brain computed tomography (CT) or magnetic resonance imaging (MRI) have been routinely performed on almost all individuals with remarkable developmental delay or intellectual disability. Array CGH and whole exome sequencing have been widely used for less than ten and three years, respectively.

Identification of sequenced neurodevelopmental disorder cases from population registries and disease collections

We identified individuals with neurodevelopmental disorder (NDD) phenotypes (intellectual disability, schizophrenia, autism and epilepsy) among 5904 individuals with exome sequence data in the FINRISK study. FINRISK is a series of population-based health examination surveys carried out every five years since 1972 to monitor the risk of chronic diseases (Vartiainen et al. 2010). The cohorts have been followed up for disease end-points using annual record linkage with the Finnish National Hospital Discharge Register and the National Causes-of-Death Register.

Additional Finnish NDD cases were included from cohorts of schizophrenia and autism patients sequenced as part of the UK10K-study (i.e. subcohorts UK10K_NK_SCZ, UK10K_KUUSAMO_SCZ and UK10K_ASDFI) and a collection of autism patients from Southern Finland (AUTISM_ASDFI) (see Supplemental Note 1 for cohort descriptions). We genetically matched each NDD case to five exome sequenced controls using the first 2 principal components (PCs). We further divided these cases and controls approximately to Northern Finnish NDD (NFNDD, Northern Finland NeuroDevelopmental Disorder) and Southern Finnish NDD (SFNDD, Southern Finland NeuroDevelopmental Disorder) cohorts based on principal component analysis (PCA).

Regional prevalence estimation of intellectual disability and schizophrenia in Finland

The Social Insurance Institution of Finland provides social security coverage for Finnish residents. The Social Insurance Institution of Finland centrally provides all disability pensions in Finland and maintains a database of all residents on a disability pension and the reason for the pension. We requested the number of individuals over 16 years of age receiving a disability pension for schizophrenia (SCZ) or ID at the end of year 2016 in each of the 19 high level administrative regions in Finland. We divided the number of beneficiaries by the population aged over 16 in each region to get a crude estimate of the relative prevalence of more severe SCZ and ID cases. The prevalence of schizophrenia particularly is higher in more detailed prevalence estimaties[11]. Schizophrenia tends to be underdiagnosed in the first years of illness[15], and only 50% of patients with schizophrenia receive a disability pension after five years of initial diagnosis[16].

GWAS genotyping and CNV calling

Copy Number Variant calling

To analyze the copy number variations (CNVs), we performed DNA Chip Array (Illumina HumanCoreExome v 12.0, Illumina PsychArray) based copy number analysis of 497 cases and 504 unaffected family members of the NFID cohort. To assess CNV frequencies in the general population we used as controls a

population-based cohort of 13 390 participants from the FINRISK study [17]. Similar to patient calls, CNV calls in controls were generated using genotyped DNA raw data from the Illumina HumanCoreExome v12.0 and v12.1.
CNVs were called using an automated CNV pipeline powered by PennCNV (Wang, et al., 2007) for sensitive CNV calling. Adjacent CNVs of similar copy number were called as one if the adjoining region between the two calls was ≤20 % of the joined CNV. To increase the confidence in the called CNVs we considered only CNVs supported by at least 10 consecutive probes and which covered a genomic region of at least 100 kb, omitting regions previously reported to have an elevated frequency for artifacts [18]. The large regional requirement was set to support analysis across the different DNA chips.

CNV Quality Control

Samples were excluded if they had: 1) a high variance (SD > 0.3) in intensity (1.5 % in NFID; 5.6 % in FINRISK), 2) a high (> 0.005) drift of B allele frequency (0 additional samples in NFID; 0.2 % in FINRISK), and 3) CNVs called in excess of 10 for one individual (10 samples in NFID; 8.9 % in FINRISK). All called CNVs for the NFID cohort, both for patients and for unaffected family members, were manually curated. For the FINRISK population cohort, CNVs were manually curated if large (> 500 kb) or if they fit into a category of interest relevant to study (see Identifying likely pathogenic mutations chapter below). Otherwise, CNVs of controls were rejected if at least 50 % of the CNV overlapped a known artifact region, or had a poor coverage (≥ 9250 base pairs per SNP).

GWAS data processing

Genotyping and Quality Control

All samples were genotyped in seven batches on either the Illumina CoreExome or Illumina PsychArray, which contains 480 000 common variants. The NFID samples were genotyped in three batches, one with Illumina CoreExome and two with PsychArray. FINRISK population controls were genotyped in five batches using Illumina CoreExome.
We excluded markers that exhibited high ‘missingness’ rates (>5%), low minor allele frequency (<1%), or failed a test of Hardy–Weinberg equilibrium (p< 1e-9). We also excluded individuals with high rates of heterozygosity (> 3sd from the

mean), or a high proportion of missing genotypes (>5%). To control for any possible population stratification, we merged the genotypes from individuals passing QC with HapMap III data from European (CEU), Asian (CHB+JPT), and African (YRI) populations. We then performed a PCA on this combined data and excluded any population outliers not clustering with the other Finnish samples
We then merged genotyping batches one-by-one and repeated the QC procedures described above on the merged dataset. To prevent any potential batch effects in the merged data, we also excluded any markers that failed a test of differential missingness (P<1e-5) between the merged batches. Furthermore, during each round of merging we performed a pseudo association analysis (using a logistic mixed-model for individuals) between samples from each batch to identify markers where the minor allele frequency deviated significantly between batches (P<1e-5). Finally, we removed related individuals (identity by descent > 0.185).
We used a custom Finnish imputation reference panel containing 1941 low-pass whole genomes (4.6x) and 1540 high coverage exomes. We used Shape-IT [19] to pre-phase the GWAS data and Impute-2 [20] to impute variants from the reference panel into the GWAS data.

Exome sequencing

NFID cases were exome sequenced at the Broad Institute using Illumina Nextera Rapid Capture Exome-capture kit and sequenced with Illumina HiSeq2000 or 2500. NFID cases were jointly called with a collection of
Finnish individuals collected as part of the Sequencing Initiative Suomi (SISU)-study (www.sisuproject.fi). The sequence data processing and variant calling has been described previously [21]. See Supplemental Note 1 for descriptions of cohorts used in the current study.

We filtered samples with estimated contamination > 3% (n= 590), chimeric reads > 3% (n= 51), as well as those samples significantly deviating from other samples within each project/batch on selected metrics (transition/transversion ratio, insertion/deletion ratio, heterozygous / homozygous variant ratio, number of singletons, n=243) and finally indluced only those with empirically confirmed >=99% Finnish ancestry (described in Rivas et al. [21]).

We first split the multiallelic variants in to bi-allelic variant records. For genotype QC we set the following genotypes to missing; genotype quality (GQ)<20, read depth (DP)<10, heterozygote allelic balance less than 20% or greater than 80%, homozygous reference alt reads >= 10%, alternate allele homozygous reference reads >=10%. Variants were filtered out if Variant Quality Score Recalibration (VQSR) did not indicate PASS, the p-value from a test of Hardy-Weinberg Equilibrium (pHWE) < 1e-9 in controls (in females only in the X chromosome), SNP quality-by-depth (QD) < 2, INDEL QD < 3 or more than 20% of heterozygote calls had allelic balance out of the 20% - 80% range. To account for the different batches of exome sequencing we required a stringent genotype call rate >= 0.95 in cases and controls separately after genotype QC. All variant and genotype QC was performed using Hail[22] and executed in the Google Cloud platform's dataproc cluster.

Finally, we ensured cases and controls were approximately independent by filtering such that all samples had a pairwise kinship coefficient < 0.0442 to every other sample. We estimated kinship coefficient using King [23] and when possible we always retained cases rather than a related control (N filtered= 1,531).

Variant annotation

We annotated variants using VEP v.85 and the LOFTEE VEP plugin (https://github.com/konradjk/loftee) to filter likely false positive protein truncating variants (PTV). We considered variant annotations of the canonical (as defined by ENSEMBL) transcript only. A variant was considered to be a protein truncating variant (PTV) if LOFTEE predicted it to be a high confidence loss of function variant (stop-gained, splice site disrupting or frameshift) without any warning flags.

Statistical analysis

Identifying likely pathogenic mutations

As a basis for identifying "**Likely pathogenic variants**" we used a gene list curated within the Deciphering Developmental Disorders study (DDD) and a gene list of 93 exome-wide significant genes from the latest DDD study meta-

analysis of *de novo* variants [4]. We downloaded a gene list curated within the DDD study (https://decipher.sanger.ac.uk/ddd#ddgenes) containing 1,897 genes with varying degrees of evidence of mutations in those genes causing developmental delays. We further subset the list to only confirmed or probable developmental delay genes contributing to a brain/cognition phenotype. This gene set was further extended by a set of 93 genes with a significant excess of damaging de novo variants in the latest DDD meta-analysis[4]. These two lists resulted in a total of 818 genes (Supplemental Table 1). In the exome of each ID patient we searched for PTV or damaging missense (MPC>=2 [24]) variants that were not observed (as homozygotes in recessive genes) in non-Finnish GnomAD individuals or in our control individuals. We used only non-Finnish GnomAD individuals, as all Finnish individuals in GnomAD are included in our control exome cohort. Variants were classified as "**Other high impact variants**" if the variant was a PTV (in PTV constrained gene, pLI[25]>0.95) or a damaging missense variant (MPC>=2) in a gene that was not in the list of known genes (as above) and not observed in non-Finnish GnomAD individuals or in our control individuals. For homozygotes we used CADD[26] score >20 to filter to putatively damaging variants, as MPC score is a measure of heterozygous constraint. In homozygote variant filtering we required that the variant was not seen as homozygous in non-Finnish GnomAD samples or in our internal Finnish controls.

The algorithm for identifying pathogenic mutations was implemented in Hail [22] and executed in a Google Cloud dataproc cluster.

All CNVs passing QC criteria were classified as either 1) likely pathogenic, 2) other high impact variant, or 3) uncertain. A "likely pathogenic" classification was largely similar to the pathogenicity criteria according to the guidelines of the American College of Medical Genetics [27], with the distinction that the size criteria for pathogenic CNVs were lowered to 1 Mb for deletions, and 500 kb for de novo deletions. CNVs were additionally considered likely pathogenic (class 1) when overlapping at least 75 % with an established disease associated locus [28], or deleting an ID associated gene of interest (see above). CNVs were classified as "other high impact variant" (class 2) if both: A) they were never seen in

unaffected family members, population controls, or the high quality variant set of the Database of Genomic Variants; and B) they deleted a gene with a high probability of loss-of-function intolerance[25] (pLI>0.95). Otherwise, a CNV was classified as a variant of uncertain significance (class 3).

Polygenic risk scores

As SNP weights we used summary statistics from GWA studies of schizophrenia [29], IQ [30], and educational attainment [31]. To avoid potential biases caused by non-random regional sampling of individuals in the GWA studies the summary statistics were generated after excluding all Finnish cohorts.

For polygenic scoring we used only well-imputed and genotyped common SNPs (Impute 2 info >= 0.9, allele frequency > 0.05). We pruned the SNPs to a subset of uncorrelated SNPs (r2 < 0.1 within 500kb) and used the remaining SNPs for calculating a polygenic risk score (PRS) for each individual by summing the product of beta from the summary statistics and the number of effect alleles (genotype dosage for imputed SNPs) over all SNPs. Our primary hypothesis testing used a PRS constructed from nominally significant variants ($p<0.05$) in the original GWAS study. The genetic scores were standardized to z-scores using the mean and standard deviation from Finnish population controls.

For visualizing geographical differences in the PRSs within Finland, we subset the controls to those whose parents' birthplaces were within 100 km of each other. An individual's coordinates were set to the average of the parents' birthplaces' longitude and latitude. We smoothed the PRS across a map of Finland. At each map position we calculated weighted average by weighting each individual's PRS by the inverse of the squared distance between the map point and the individual's coordinate. Individuals within 50km from the map point contributed equally to the map point, i.e., the full weight was given to those individuals independent of their exact distance from the map point.

Association analysis

To control for population stratification we matched each case to its five genetically closest control individuals given by the first two PC's using the optmatch R package.

For replication and for studying the neurodevelopmental spectrum of candidate variants in the exome analysis, we additionally identified neurodevelopmental (NDD) cases (ID, SCZ and ASD) from the Finnish FINRISK population cohort as well as Finnish disease-specific collections sequenced in the UK10K study (SCZ and ASD) (Table 2). Each NDD case was genetically mapped to its five closest controls that were not matched to NFID patients.

For the dominant association analysis we used both Fisher's exact test and Firth bias corrected logistic regression using the four first PC's as covariates. We meta-analyzed the results across the three cohorts (NFID, North NDD and South NDD) using Mantel-Haenzel meta-analysis (rma.mh in metaphor[32] R package) for Fisher's analysis and a sample size weighted meta-analysis for Firth [33].

For the recessive analysis we used a recessive allele frequency test (RAFT) [34], which takes the population allele frequency of the variant tested into account to estimate the probability of observing as many cases and controls as homozygotes under the null. As we genetically matched all cases to controls we present the analysis results from Fisher's exact test and present Mantel-Haenzel meta-analysis and Firth results in the supplement.

Association analyses were performed using Hail [22] and executed in a Google Cloud dataproc cluster.

Enrichment analysis

For testing if different classes of variants were enriched in cases vs. controls we used Fisher's exact test and for significant variant classes we estimated the variance explained by Nagelkerke's pseudo r^2.

For the CNV analysis, we used the same cases and controls as in the exome analysis where GWAS data was available (433 cases and 1100 controls passing QC for CNV analysis). We tested for the degree of association between intellectual disability and different categories of pathogenic or likely pathogenic CNVs. Association analysis was performed testing carrier ratios using Fisher's exact test. The relevant categories were: 1) CNVs overlapping one of DECIPHER's syndromic regions 2) deletions overlapping a known developmental delay gene, and 3) deletions overlapping a gene with high probability of protein truncating variant intolerance (pLI>0.95) [25].

Heritability estimation

We estimated the variance explained by different variant categories by fitting a logistic model and computing Nagelkerke's pseudo r^2 from the fitted full and null models.

Case/control status was used as a dependent variable and as an explanatory variable we used either a binary indicator for presence/absence of variant in a given category (likely diagnostic or other high impact) or a continuous variable for PRS variance estimation. As we observed geographical differences in all evaluated PRSs we corrected for the first four PCs even after genetic matching of cases and controls to account for any residual stratification (i.e. the null model included the first four PCs). Confidence intervals for r^2 were estimated by drawing 5,000 bootstrap samples and computing the r^2 for each sample. We used the adjusted bootstrap percentile method[35] to obtain confidence intervals from the 5,000 bootstrapped samples.

We wanted to compare the variance explained not only for the whole ID cohort but also in mild and severe ID separately. As mild and severe ID have different population prevalence we transformed the observed scale variance explained to the liability scale [36]. We used the population prevalence from a cumulative normal distribution function with mean 100 and standard deviation 15. Prevalence of 1.94%, 1.91% and 0.034% were used for all ID (IQ<70), mild ID (50<=IQ<70) and other more severe ID combined (IQ<50) respectively.

Results

We first estimated the regional prevalence of ID in Finland using the social security disability benefits register. We observed a higher prevalence of individuals receiving disability benefits for ID in the Eastern and Northern parts of Finland as compared to Southern and Western Finland (**Figure 1**). The highest prevalence was observed in Kainuu and North-Ostrobothnia, two of the primary municipalities of the NFID patient collection (**Figure 1**).

We then aimed to genetically characterize the NFID patient cohort by exome sequencing and GWAS array genotyping. After joint genotype calling and quality control we analyzed the exomes of 442 independent ID patients (Table 2) and 2,206 genetically matched population controls. The CNV analysis was performed with 431 independent ID cases and 1,100 genetically matched controls. The polygenic risk score analysis was performed in 439 ID patients and 14,816 controls. For replication and for studying the neurodevelopmental spectrum of candidate variants in the exome analysis, we additionally identified neurodevelopmental disorder (NDD) cases (ID, SCZ and ASD) from the FINRISK population cohort as well as the Finnish disease-specific collections sequenced in the UK10K study (SCZ and ASD) (Table 2). To account for regional stratification, we genetically matched each NFID patient to their five closest Finnish control individuals. Likewise for all population NDD cases we matched each one to its five closest controls and divided the population NDD cases and controls to North (Northern Finland NeuroDevelopmental Disorder, NFNDD) and South Finland (Southern Finland NeuroDevelopmental Disorder, SFNDD) based on PCA (Table 2). The remaining 5,922 control exomes served as Finnish population specific controls, which we, in addition to GnomAD non-Finnish exomes, used for variant filtering.

Table 2 Cohorts used in the analyses. Sample sizes are the numbers used in the analysis after quality control. ASD: autism spectrum disorder. SCZ: schizophrenia. EPI: epilepsy. PRS: Polygenic Risk Score, CNV: copy number variant.

			Cases				Controls
Analysis cohort	**Analyses**	**N**	**ID**	**ASD**	**SCZ**	**EPI**	**N**
NFID (Primary cohort)	EXOME	442	442	62	47	92	2206†
	PRS	439	439	62	47	86	2195† (matched) 14816 (total)
	CNV	433	433	57	47	84	1100†
NFNDD (Northern Finland NeuroDevelopmental Disorder)	EXOME association analysis	314	17	40*	239	26	1548
SFNDD (Southern Finland NeuroDevelopmental Disorder)	EXOME association analysis	322	14	73*	211	33	1594
SISU controls	EXOME variant filtering	-	-	-	-	-	5922

* comorbidities were not available from ASD cohorts

† None of the controls from the FINRISK study have any NDD (intellectual disability, autism, schizophrenia) or epilepsy and all controls are genetically matched to cases.

Mutations in known genes causing cognitive impairment

To identify those individuals who had a potential causative variant in the exome analysis, we first searched for damaging missense or protein truncating variants (PTV) in 818 known, curated genes selected from the DDD-study (see Materials and Methods and full gene list in Supplementary Table 1). For genes where autosomal recessive inheritance has been reported, only homozygote variants were considered. Within these 818 genes we identified a likely pathogenic mutation in 64 patients (Supplementary Table 2). When we compared the rate of likely pathogenic variants to the 2,210 genetically matched controls from the

FINRISK population cohort, we observed the strongest enrichment in the PTV class of variants (OR: 12.04, 95% CI: 5.46–28.56, p: 1.4e-11) followed by dominant acting (OR: 6.03, 95% CI: 3.57–10.24, p: 7.5e-12) and recessive (OR: 1.83, 95% CI: 0.7–4.30, p: 1.4e-1) constrained/damaging missense variant classes (Figure 2).

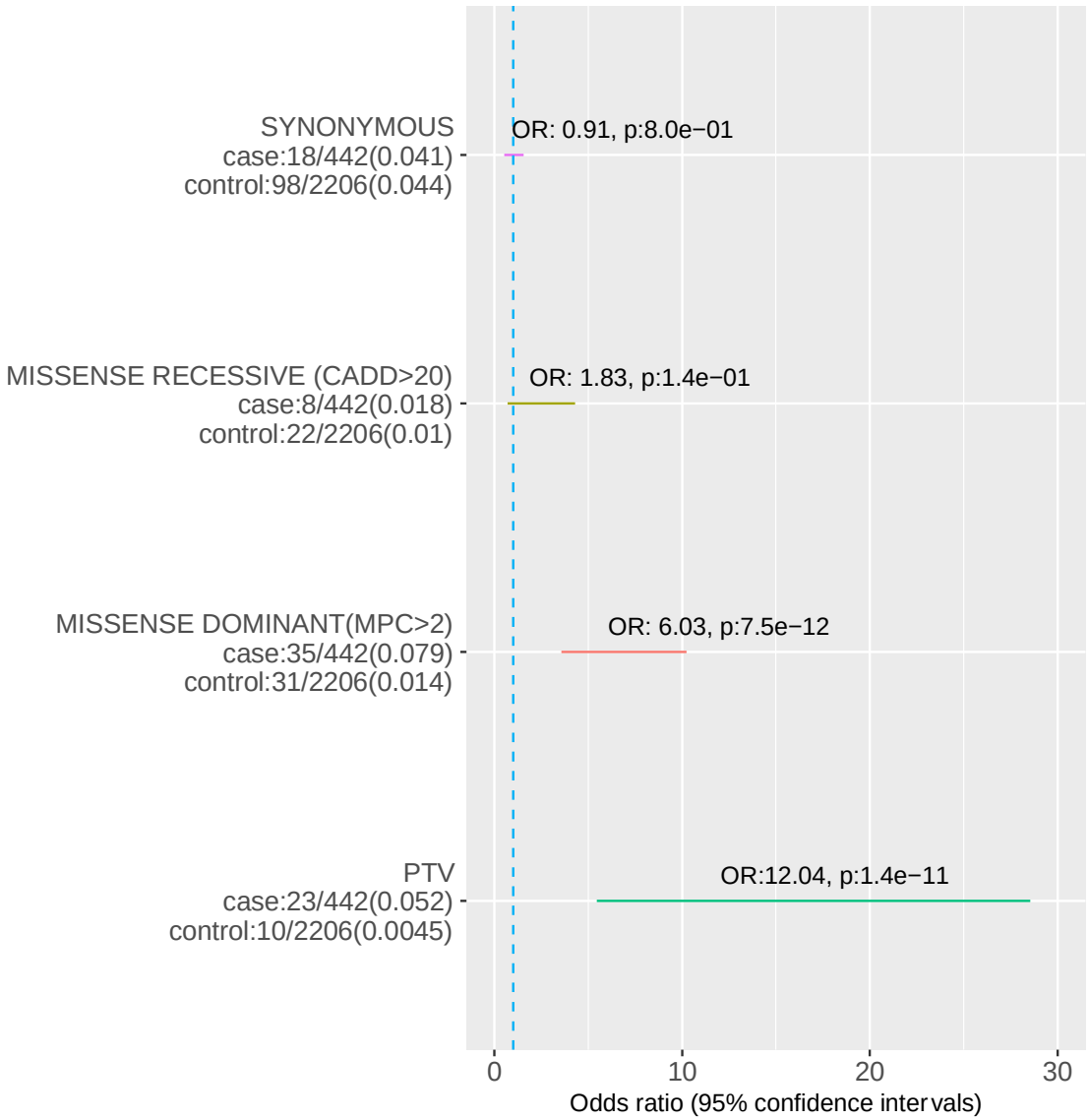

Figure 2. Enrichment of likely pathogenic and synonymous variants (from 818 known ID genes, see Materials and Methods) in NFID cases compared to genetically matched controls. Heterozygotes were counted only for those genes for which a dominant inheritance mode is reported. The number of carriers and total individuals are given on the left and in parenthesis the proportion of carriers. Circles indicate the odds ratio (OR) and lines indicate 95% confidence interval of the OR. The synonymous variant identification comparison was performed to assess if possible differences in the variant identification rate due to batch/capture differences were adequately controlled. PTV= protein truncating variant; CADD = Combined Annotation-Dependent Depletion pathogenicity score; pLI = probability of loss of function intolerance.

Burden of variants in genes not previously implicated in cognitive impairment

Given that ~85% of our cases did not have a variant affecting a known neurodevelopmental disorder gene, we next wanted to assess if there was a burden of rare variants outside of those known genes. We performed an enrichment analysis of variants that were either PTV or constrained damaging missense (MPC>2) variants and not observed in the non-Finnish GnomAD samples or in our internal Finnish controls (separate control cohorts were used for filtering and enrichment, see Table 2). First we wanted to verify that there was no spurious enrichment of variants caused by stratification (or other technical differences) by analyzing if there was an enrichment of synonymous variants (not observed in GnomAD or our Finnish controls) between cases and controls. No such enrichment was observed, suggesting that QC and case control matching were successful (Figure 3).

Dominant PTVs in high pLI genes (OR: 2.65, 95% CI: 2.02-3.47, p:2.2e-12) and constrained damaging missense variants not seen in GnomAD or Finnish controls within novel genes (OR: 1.64, 95% CI: 1.29-2.08, p: 5.9e-5) were significantly enriched in cases (Figure 3). The signal for PTV variants was almost exclusively in genes intolerant of PTV-mutations (pLI<0.95, OR: 1.16, 95% CI: 0.94-1.44, 1.7e-1; Figure 3). Homozygous PTVs (likely complete knockout of a gene) in novel genes were over twofold enriched in cases, but were not statistically significant due to low counts (OR 2.51, 95% CI: 0.40-11.78, p: 1.8e-1; Figure 3).

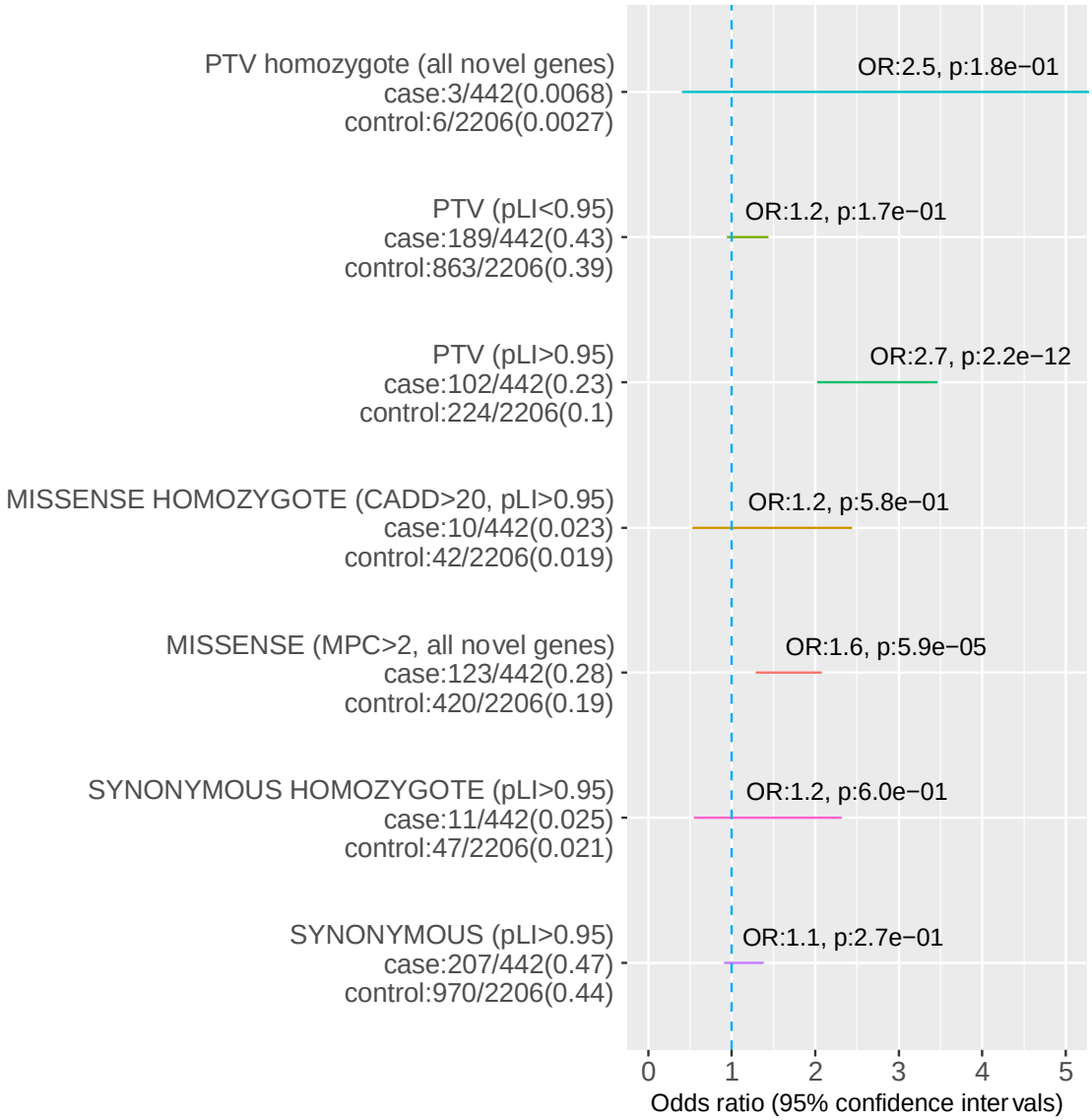

Figure 3 Enrichment of rare variants in genes not previously associated with NDDs (not observed in GnomAD or Finnish controls) in cases compared to genetically matched controls (see Materials and Methods). Rate in each variant category is first estimated for all novel genes and then after subsetting to only novel high pLI genes. All missense variants are predicted to be deleterious (MPC>2, see methods). On the left the number of carriers and total individuals are given and in parenthesis the proportion of carriers. Circles indicate the odds ratio (OR) and lines indicate the 95% confidence interval of the OR. The synonymous variant identification comparison was performed to assess if possible differences in the variant identification rate due to batches/capture differences were adequately controlled. PTV= protein truncating variant; CADD = Combined Annotation-Dependent Depletion pathogenicity score; pLI = probability of loss of function intolerance.

Copy Number Variants

After filtering CNVs according to our QC protocol (see Methods), we assessed the contribution of likely diagnostic CNVs in 433 NFID patients and 1100 genetically matched controls. We found that deletions of any type (>100kb in

size) were observed slightly more often in cases than in controls (OR: 1.25 (CI 0.92-1.69) p: 1.3-e1, Figure 4A: n_{cases}= 84 (19.4%), $n_{controls}$=177 (16.1%)). However, deletions that were >500kb were clearly more frequent in cases, regardless of their chromosomal location (OR 2.6 (CI 1.3 - 5.2) p: 3.7e-3, NFID: 30 individuals (6.8%), FINRISK: 20 individuals (1.8%); Figure 4 A). When the deletions were categorized based on their genomic position, the enrichment of deletions in cases became evident. Deletions that have previously been associated with syndromes, or spanned a previously identified ID gene, were more frequent in cases than controls (OR: 35.8 (CI 8.3 – 323.5), p: 4.4e-10) 15 patients (3.7 %) vs. two population controls (0.02 %); Fig 4A).

Using our classification algorithm, we identified a likely pathogenic CNV in a total of 25 cases, and a further 13 cases were found with a CNV classified as a “high impact variant” (Supplementary Table 3). This would associate a total of 38 cases to a CNV deletion. Most of the likely pathogenic CNVs met multiple pathogenicity criteria. For example, the 22q11.2 deletion, the most common syndromic CNV identified (eight cases, 1.8 %, and zero controls) was classified as pathogenic because it was large (2.56 Mb) and contains a deletion of a known ID-associated gene (*TBX1*). Additionally, the deletion also contains seven other high pLI genes in addition to *TBX1*. Deleterious variants in *TBX1* have previously been associated with many of the characteristics of the 22q11.2 syndrome [37]. The relatively high representation of 22q11.2 deletion syndrome underlines the diagnostic challenges of this relatively common CNV’s variable phenotypic consequences [38].

Large deletions (>500 kb for de novo deletions and >1 Mb for others) were the most commonly identified likely pathogenic CNVs, observed in 20 cases (4.6%) and seven controls (0.6%) (Fig 4B). A total of 11 cases (3.8 %) carried a CNV overlapping a region previously linked to syndromic ID, while the same was true for two controls (0.2%) (Figure 4 A). The syndromic CNVs identified in controls were non-ID associated (12p13.33 deletion) and a region with known variable phenotype (22q11 duplication syndrome [39]). Previously known ID-associated gene was deleted in 13 cases (3.0%) but in none of the controls.

As a pathogenic CNV was observed in only 25 cases (5.8%), we analyzed if there was an excess of smaller deletions in genes intolerant of PTV variations not previously associated with cognitive phenotype. After removing likely pathogenic CNV types, we observed such deletions in 13 cases (3.0%) and 10 controls (0.9%) (OR 6.2 (CI 2.5 - 15.8), p: 3.1e-5) (Supplementary Figure 5), totaling to 38 cases with a potential CNV deletion associated with ID.

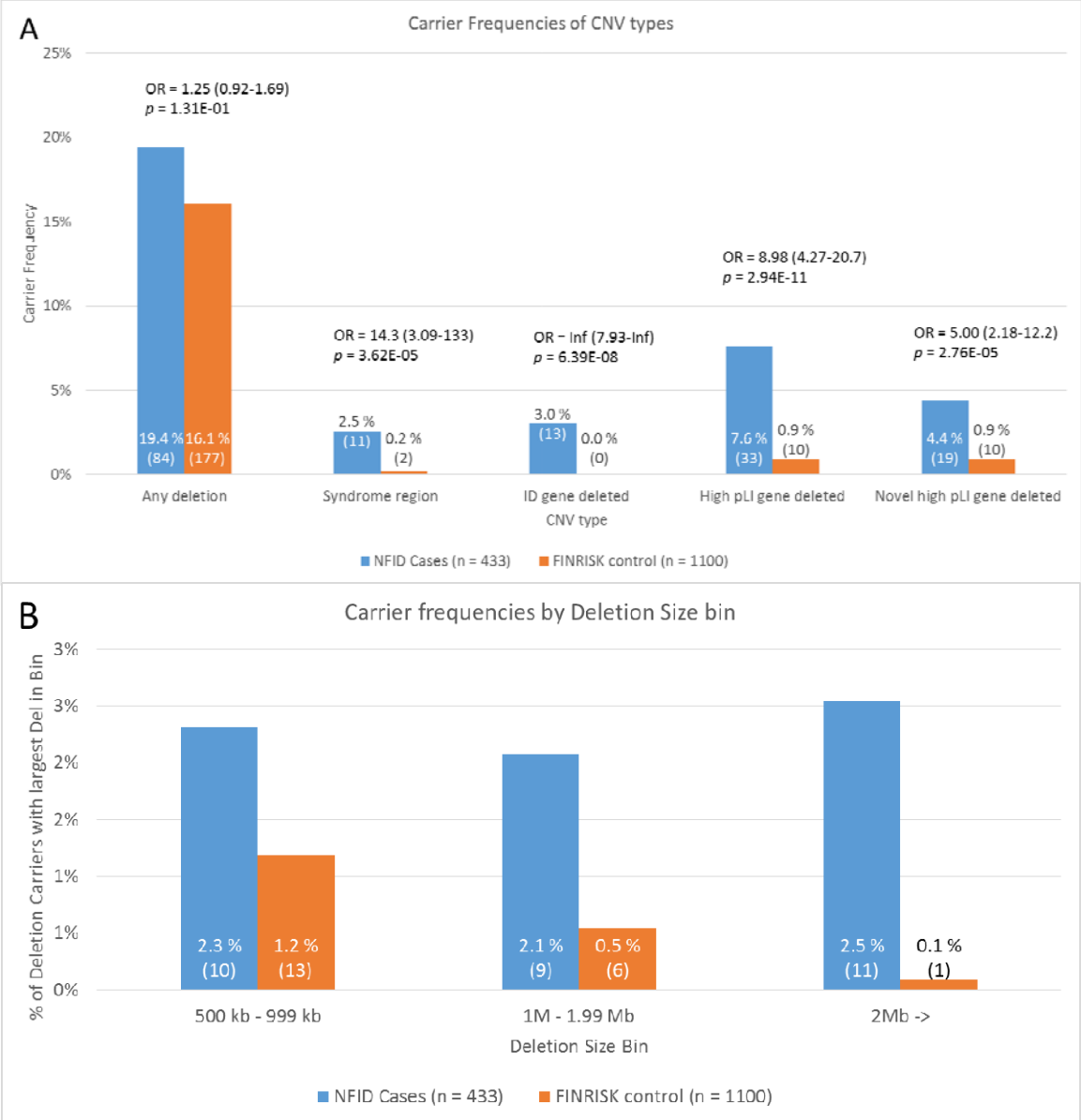

Figure 4 Distribution of different deletion categories. A) Deletion categories in ID patients showed enrichment for deletions (>100 kb) in general, and specifically in deletions covering syndrome regions (as defined by the DECIPHER database), deletions that are located in an ID-associated gene region (see above) or CNVs deleting a gene intolerant of protein truncating mutations (pLI > 0.95). B) Histogram of deletion carrier frequency by size of an individual's largest deletion.

Total genetic diagnosis rate

After combining exome and CNV data, we identified a likely diagnosis for 85 (19.6%) patients (Figure 5 A) of the 433 patients with both exome and CNV data available). The strongest risk factor was having a PTV in a known neurodevelopmental disorder gene (OR 12.3, 95% CI: 4.5-41.6, p: 3.3e-9) followed by a likely pathogenic deletion (OR 8.7, 95 % CI: 4.0-21.1, p: 6.5e-10) and then by a constrained missense variant in a known developmental disorder gene (OR 6.4, 95% CI: 3.3-12.7, p: 5.5e-10) (Figure 5B). We then analyzed if there was a signal from damaging variants (PTVs, missenses MPC>2 or CNVs) outside of known ID-associated genes (termed “Other high impact variants”). Indeed, we observed a significant enrichment of “Other high impact variants” in cases vs. controls (Figure 5 A). PTVs (OR 3.0, 95% CI: 2.2-4.2, p: 1.7e-12) and deletions (OR 3.5, 95% CI: 1.3-9.4, p:5.8e-3) in high pLI genes as well as constrained missense variants (OR 1.7, 95% CI: 1.3-2.2, p: 6.7e-5) were significantly enriched in the “Other high impact variants” category (Supplementary Figure 6 A).

As much less is known about the genetic architecture of mild ID as compared to the more severe ID diagnoses[3], we wanted to assess if rare variants in the same known genes contribute equally to mild and severe forms of ID in our cohort. For the analysis we combined the moderate and profound ID patients in a “severe category”. The overall rates of “likely pathogenic” (OR 5.6, p: 5.4e-10) and “other high impact” variants (OR 2.3, p: 6.5e-7) were significantly higher in the mild group than in controls (Supplementary Figure 7 A). However, when the variants were compared between mild and severe IDs, we observed a significantly higher (OR: 2.4, p:7.0e-4) proportion of “likely pathogenic variants” in known ID genes in the severe ID group (Figure 5 C). This is in line with the hypothesis that the etiology for mild ID is less driven by *de novo* and ultra-rare, high-impact variants than the etiology for severe ID. This is also consistent with previous epidemiological data suggesting mild ID is a distinct disease entity with different genetic architecture from more severe ID[5]. For CNVs, the diagnostic rate did not follow the same pattern of increased likely pathogenic CNV in more severe cases than in mild cases (Figure 5 D). This is likely because a large

fraction of ID patients who had a chromosomal abnormality had been identified in previous clinical cytogenetic analyses and thus excluded from this study.

The vast majority (87.5%) of the likely pathogenic variants identified in exome sequencing were heterozygote variants in dominant acting neurodevelopmental disease associated genes (Figure 2). As we observed a significant enrichment of damaging variants and CNVs on high pLI genes not previously linked to neurodevelopmental diseases, we categorized each patient into three diagnostic categories. These categories are 1) “Likely pathogenic variant” if the patient had any large deletion or damaging variant in known genes was identified (see Methods) 2) “Other high impact variant” if PTV or missense (MPC>2) variants (not seen in GnomAD or Finnish controls) or deletions were identified in a high pLI gene not previously associated with cognitive phenotype (i.e. no “Likely pathogenic” variant identified) or 3) “No diagnosis” category was given if no variant in 1 or 2 was identified.

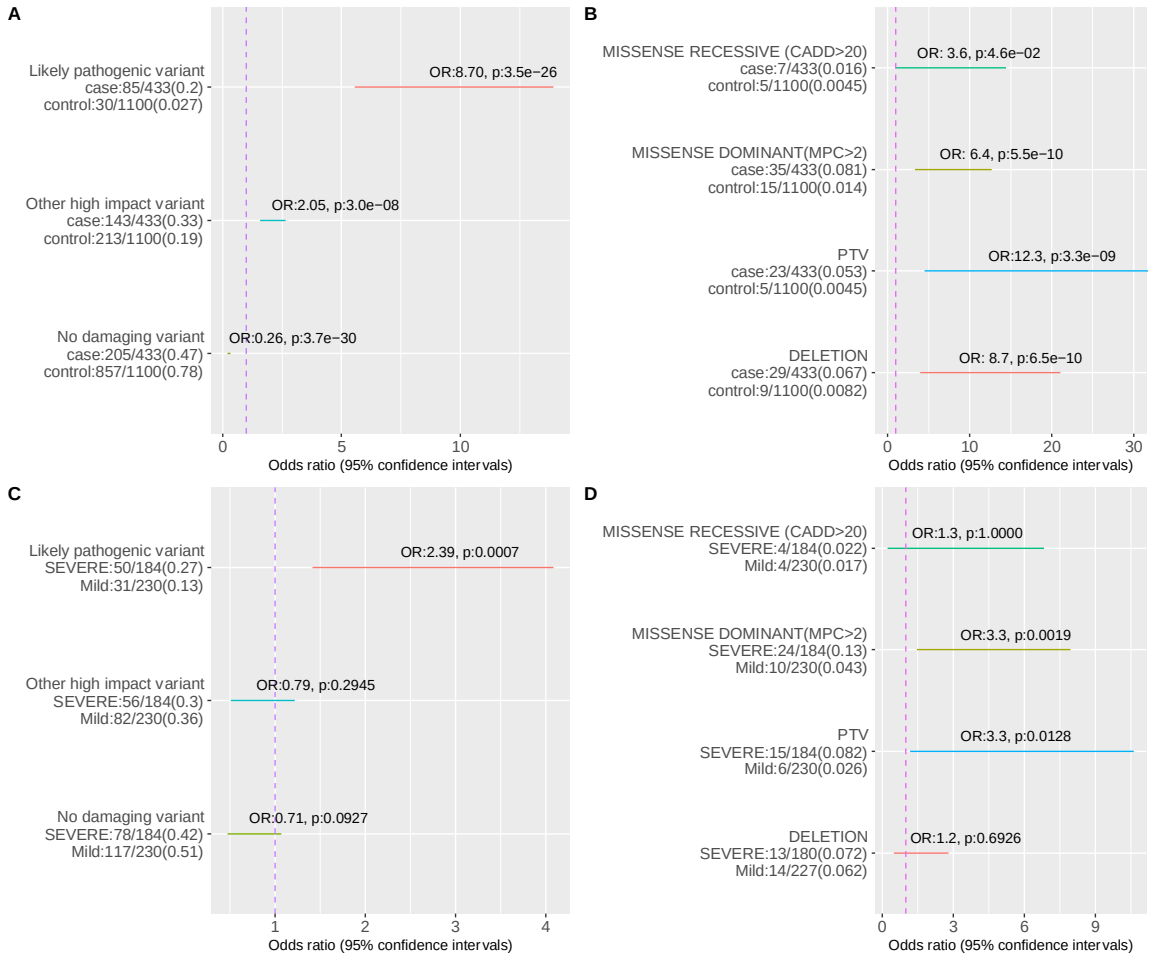

Figure 5 Comparison of the total rate of different classes of variants in cases vs. genetically matched controls and in mild vs. more severe ID individuals for which both exome and CNV data was available. On the left the number of carriers and total individuals are given and in parenthesis the proportion of carriers. Circles indicate odds ratio and lines indicate 95% confidence intervals of the odds ratio estimate. A) Total genetic diagnostic rate. B) Variant classes in "Likely pathogenic" variant categories. C) Comparison of the rate of identifying different classes of variants in mild vs. severe (moderate and profound ID combined) patients D) Comparison of the rate of variant types in "Likely pathogenic" category between mild and more severe forms of ID (moderate, severe and profound ID combined). Constrained missense (MPC>2) variants were analyzed in all genes instead of only high pLI genes in C and D as MPC score incorporates regional missense constraint.

Polygenic common variant load

As it is evident that exome sequencing and CNV analysis identified a likely cause for ID in only 19.6% of the cases, we wanted to study the contribution of the polygenic load of common variants associated to intelligence quotient (IQ), educational attainment (EDU) and schizophrenia (SCZ) to Northern Finnish ID. There is a partial common variant genetic overlap between cognitive function and

schizophrenia [30,31,40] and therefore we also studied if the schizophrenia polygenic risk score would have any contribution to Northern Finnish ID. We estimated the regional prevalence of SCZ similarly as we did for ID and observed a similar regional enrichment in Northern and Eastern Finland (Supplementary Figure 1). First, to create a proper reference for NFID cases, we analyzed whether the geographical distribution of PRSs correspond to the population history of Finland. We genotyped and imputed 14,833 individuals from the population-based FINRISK collection and used all loci with a lead variant p-value <= 0.05 in the meta-analyses for schizophrenia, IQ and educational attainment. To visualize the geographical distribution of PRSs we used a distance weighted polygenic risk score in 2,186 Finnish individuals who did not have any neurodevelopmental disorders and whose parents were born within 100 km of each other (see methods). The PRSs for educational attainment and IQ were lower, and for SCZ higher in the Eastern and Northern part of Finland than in the Southern and Western Finland (Figure 6 A,C,E). Next we asked if the PRSs were associated with ID. When all of the ID cases were analyzed as one group, all PRSs were significantly associated with the ID phenotype as compared to genetically matched Finnish controls (Figure 6 A,C,E). The PRS for EDU, SCZ and IQ explained 0.94%, 0.55 and 0.48 of the heritability on the liability scale respectively (see Supplementary Figure 2 for heritability estimation using varying significance thresholds for locus inclusion).

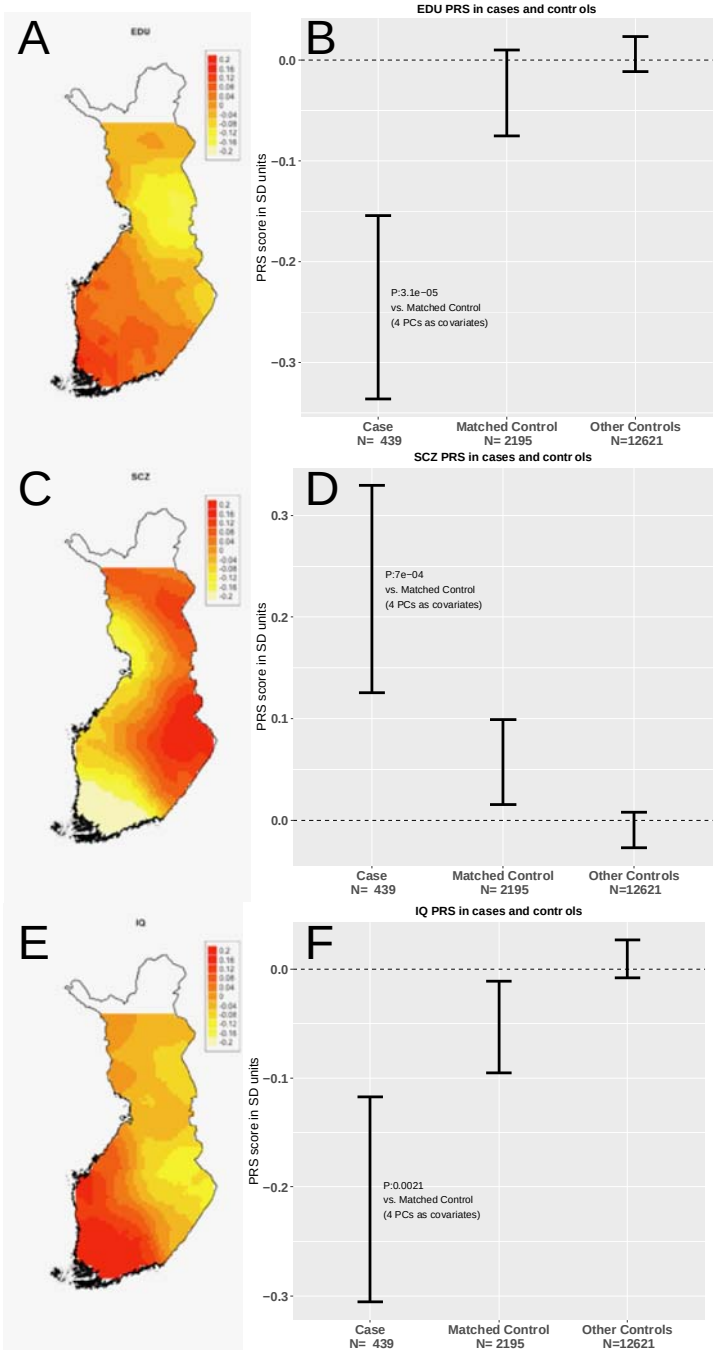

Figure 6 Regional distributions of PRSs within Finland and ID patients compared to genetically matched controls in SD units of PRS. A) Locally weighted educational attainment PRS distribution in Finnish population controls whose parents birthplace is within 100km of each other. B) Educational attainment PRS in cases and all population controls and genetically matched population controls C) Locally weighted schizophrenia PRS distribution in Finnish population controls whose parents birthplace is within 100km of each other. D) Schizophrenia PRS in cases and all population controls and genetically matched population controls. E) Locally weighted IQ PRS distribution in Finnish population controls whose parents birthplace is within 100km of each other F) IQ PRS in cases and all population controls and genetically matched population controls. Error bars indicate 95% confidence intervals around mean.

We next analyzed whether PRS values were different in the different ID groups: mild, moderate and profound. The PRS for EDU was lower and for SCZ higher in the mild ID cases compared to more severe forms, but the differences were not statistically significant (Supplementary Figure 3). Unexpectedly, the IQ PRS in the mild ID group was not significantly different from matched population controls. However the most severe ID form differed significantly from the population controls (Supplementary Figure 3). We also hypothesized that the EDU and IQ PRSs would be lower and SCZ PRS would be higher in those patients for which a likely causative mutation was not identified. Thus, we compared the PRSs between cases in different diagnostic categories but did not observe statistically significant differences between groups (Supplementary Figure 4). Assuming we had sufficient power to detect differences, this suggests that in addition to high penetrance variants, a more general polygenic component also contributes to the genetic background of ID.

Variants enriched in Finland

Finally, we asked if some variants enriched in Finland might contribute to the Northern Finnish ID phenotype as variants with reduced reproductive fitness can exist in markedly higher frequency in a population with a recent bottleneck [9]. We aimed to identify a subset of variants that are rare elsewhere but significantly more common in Finland using the largest exome variant database to date (GnomAD/ExAC v 2). We hypothesized that some of these variants would be associated with ID in the NFID cohort. To identify these variants, we compared the allele frequencies in Finnish samples to the allele frequency in non-Finnish Europeans. Loss-of-function and missense variants in the range of 0.1% - 5% (Supplementary Figure 9) were proportionally more enriched compared to other variants, suggesting that damaging variants that survived through the bottleneck are elevated in frequency. This is in line with our previous observation in smaller datasets [8].

Dominant variants enriched in Finland

We first analyzed low frequency and rare (MAF < 0.1% in GnomAD non-Finnish population maximum) single missense and loss-of-function variants enriched at least two-fold in Finland or absent in GnomAD non-Finnish individuals.

Singletons were excluded from the analysis (13,483 variants; 12,628 missense, 855 PTV). We identified 396 variants nominally ($p<0.05$) associated with ID (Supplementary Table 4). We then aimed to gain confidence or refute the identified associations by analyzing NDD cases and genetically matched controls from the Northern and Southern Finland (Table 2). After meta-analyzing all three cohorts, we identified 29 variants associated with a p-value < 0.001 (20 variants were found in cases only across the three cohorts). However, none of the meta-analyzed variants surpassed a multiple-testing correction for 13,483 tests.

Recessive variants enriched in Finland

Consistent with the bottleneck effect and the associated Finnish Disease Heritage, we expected to observe enrichment in recessive acting variants. We therefore asked if some of the enriched PTV or missense variants with low allele frequency in GnomAD (AF <0.01) were recessively associated with ID. We excluded singleton homozygotes (across all cases and controls) and variants observed as homozygous in GnomAD. After these filtering steps we performed a recessive analysis for 1408 variants (1379 missense variants and 29 PTVs). Eighteen variants were observed as homozygous more than once in cases across the three cohorts but not in controls (Table 3).

In the recessive analysis we identified a homozygous missense variant in the *CRADD* gene in three independent ID cases. Additionally, we identified one *CRADD* missense homozygote in the Northern Finland NDD case cohort (RAFT meta p: 5.75E-08). The variant is over 50 times more frequent in Finland than in non-Finnish Europeans. One other variant in the *HGF* gene achieved a p-value surviving Bonferroni correction for 1408 tests (3.6e-05), but it was observed only in two cases. Recessive variants in *HGF* have been identified in consanguineous families ascertained for non-syndromic deafness[41]. Our cases did not have history of any hearing problems. Among the 18 genes with case-only recessive candidate variants we observed significantly more genes that are intolerant of homozygote PTV variation (pRec>0.8[25]) than expected by chance. A pREC metric was available for 17 of the 18 of the candidate genes, of which eight were intolerant of homozygote PTV variation. In ExAC 4,508 out of 18,241 genes have pRec>0.8, and therefore we would expect 4.2 genes by chance (binomial test p-value 0.046). This suggests that some of the 18 candidate variants are true risk/causative variants for ID (see Supplementary Table 5 for all 59 nominally significant associations where at least 1 NFID homozygous case was observed).

Table 3 Homozygous Finnish enriched variants observed >= 2 times across NFID and the Southern and Northern Finnish NDD cases and not observed in any controls as homozygous. Variants in highlighted rows are significant after multiple testing correction.

												NFID		Population Replication	NDD
variant	gene	Previous evidence	type	pLI	pRec	GnomAD FI	GnomAD popmax	NORTH Finland AF	SOUTH Finland AF	NFID AF	RAFT meta p	case hom	RAFT p	case hom	RAFT p
12:94243956 G:A	**CRADD**	**AR Lissencephaly, ID** [42]	**missense**	**0.72**	**0.26**	**6.01E-03**	**9.15E-04**	**6.67E-03**	**3.48E-03**	**7.83E-03**	5.01E-08	3	1.86E-06	1	9.30E-03
7:81374424 G:C	**HGF**	**AR hearing loss(OMIM)**	**missense**	**1.00**	**0.00**	**1.12E-03**	**5.48E-04**	**1.30E-03**	**3.15E-04**	**2.52E-03**	1.34E-05	1	3.20E-03	1	2.54E-03
12:15784582 T:C	EPS8	AR deafness (OMIM) impaired cognitive function in mice [43]	missense	0.99	0.01	9.03E-03	1.30E-03	6.92E-03	7.36E-03	6.89E-03	1.28E-04	2	1.28E-04	0	NA
1:220236134 C:T	BPNT1	-	missense	0.00	0.91	1.07E-02	4.83E-03	5.55E-03	1.09E-02	6.89E-03	1.34E-04	2	1.34E-04	0	NA
7:1520077 T:C	INTS1	AR ID [44]	missense	0.14	0.86	1.30E-02	3.05E-03	1.45E-02	9.39E-03	1.71E-02	1.95E-04	3	1.95E-04	0	NA
2:95753239 A:G	MRPS5	-	missense	0.02	0.98	9.87E-03	4.56E-03	4.89E-03	1.16E-02	8.29E-03	2.88E-04	2	2.88E-04	0	NA
10:123844296 C:A	TACC2	-	missense	0.00	0.88	1.32E-02	1.10E-03	1.73E-02	1.38E-02	1.14E-02	1.05E-03	2	1.05E-03	0	NA
1:155028692 C:T	ADAM15	-	missense	0.00	0.95	8.69E-03	2.37E-03	1.23E-02	7.05E-03	1.18E-02	2.11E-03	1	4.84E-02	1	3.58E-02
18:14542688 G:A	POTEC	-	missense	0.11	0.88	1.86E-02	6.26E-03	2.47E-02	1.79E-02	1.98E-02	2.11E-03	2	1.01E-02	1	1.65E-01
21:19651329 G:C	TMPRSS15	Enterokinase deficiency (OMIM)	missense	0.00	0.03	1.73E-02	6.31E-03	2.35E-02	1.56E-02	1.93E-02	2.66E-03	1	1.42E-01	2	9.85E-03
15:60789800 T:C	RORA	AD ID [45]	missense	1.00	0.00	1.02E-02	9.14E-04	1.19E-02	9.59E-03	1.59E-02	4.03E-03	1	9.35E-02	1	3.38E-02
11:6023849 C:T	OR56A4	-	missense	0.30	0.61	1.69E-02	1.82E-03	1.81E-02	1.82E-02	1.61E-02	4.45E-03	2	4.45E-03	0	NA
11:3681309 G:A	ART1	-	missense	0.00	0.07	1.64E-02	3.68E-03	1.64E-02	1.67E-02	1.53E-02	5.67E-03	1	8.43E-02	1	5.76E-02
19:56424477 TC:T	NLRP13	-	frameshift	0.00	0.92	1.65E-02	1.28E-03	1.22E-02	1.75E-02	1.40E-02	6.07E-03	1	7.44E-02	1	8.54E-02
8:17612739 G:C	MTUS1	-	missense	0.00	1.00	1.46E-02	2.01E-03	1.67E-02	1.62E-02	1.54E-02	6.74E-03	1	9.22E-02	1	7.45E-02
1:183520048 A:T	SMG7	NMD-components linked to ID [46]	missense	1.00	0.00	2.49E-02	1.10E-03	2.37E-02	2.71E-02	2.62E-02	3.85E-02	1	2.85E-01	1	1.58E-01
X:23410887 C:T	PTCHD1	x-linked ID/AUTISM [47]	missense	0.94	0.06	2.43E-04	2.51E-05	0.00	0.00	2.27E-04	NA	1*	NA	1*	NA*

AR= Autosomal Recessive, AD = Autosomal Dominant, ID = Intellectual Disability

* RAFT statistic not valid for X-chromosome. Both carriers are hemizygote males

Variance explained by different variant categories

To put the relative contribution of different classes of genetic variation to a context, we estimated the variance explained by each category that was significant in the case-control comparisons. We added the 3 *CRADD* homozygotes to the likely pathogenic category as we clearly demonstrated the variant to be a causative recessive variant (see Table 4). For likely diagnostic and other high impact variants we used genetically matched cases and controls for which both exome and CNV data was available (433 cases and 1,100 controls). For the polygenic risk score we used 439 cases and 2,195 genetically matched controls. As we observed geographical differences in PRSs within Finland, we corrected the variance explained estimation using the first four PCs. All the PRSs were significantly associated with ID. The PRS's contribution to heritability on the liability scale is lower (IQ 0.48%, 95% CI: 0.067%-1.25%; SCZ 0.55%, 95% CI: 0.078-1.5%; EDU 0.94%, 95% CI: 0.31%-1.92%) than variance explained by likely pathogenic variants in known genes (4.54%, 95% CI: 3.04%-6.15%) or other high impact variants (in constrained genes not yet linked to ID 2.25%, 95% CI: 1.30%-3.52%) (Figure 7A). When comparing the different ID severities, the heritability explained on liability scale was the highest in mild ID for EDU (2.17%, 95% CI: 0.60%-4.77%) whereas it was smaller in more severe ID (0.55%, 95% CI: 0.04%-1.60%). The heritability estimation for all PRSs and ID categories is presented in Supplementary Figure 10.

The variance explained by likely pathogenic variants in known genes explained a slightly higher proportion of variance on the liability scale in more severe ID (7.01%, 95% CI: 4.49%-9.92%) than in mild ID (4.25%, 95% CI: 2.6%-9.5%). This is expected as we observed a significantly lower proportion of likely pathogenic variants in mild ID (13.7%) vs. more severe ID (27.2%) (Figure 7 B).

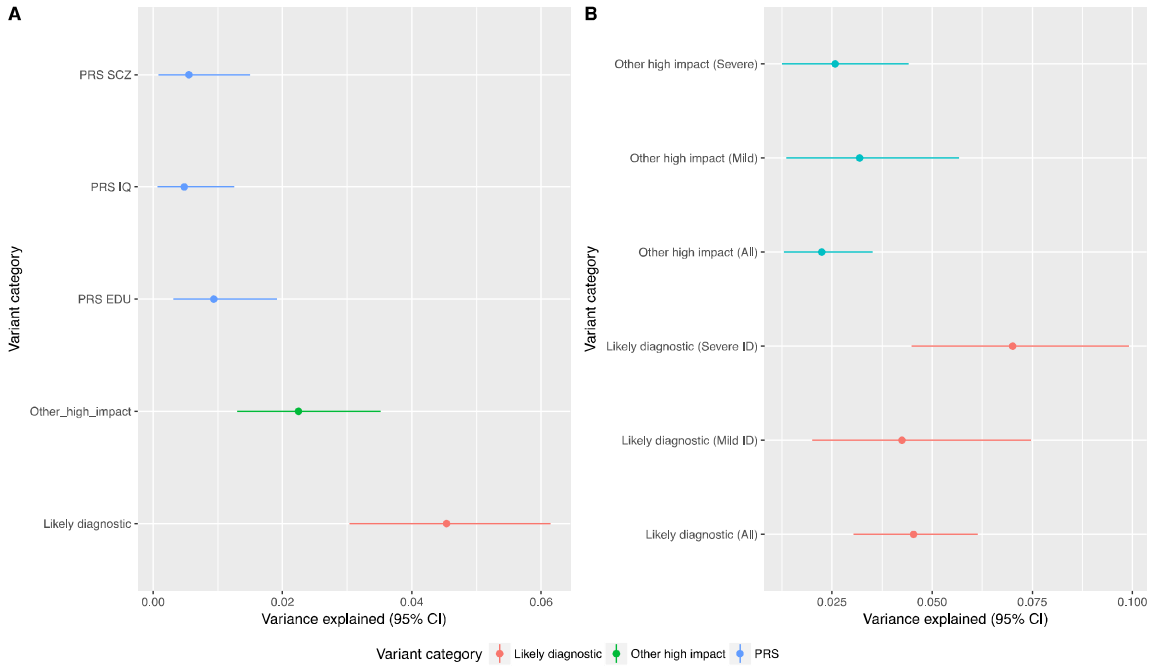

Figure 7 Estimate of heritability explained by different variant categories on liability scale. A) variance explained by genetic categories in all ID cases. B) Variance explained delineated by ID severity. Variance explained was estimated by Nagelkerke r2 while controlling for the first four PCs. 95% Confidence intervals of variance explained were estimated by 5000 bootstrap samples.

Discussion

Here we have described a comprehensive genetic analysis of an ID cohort from a population with a relatively high prevalence of ID. We studied the contribution of SNVs and INDELs, CNVs, and of a genome-wide common variant polygenic load. Unlike most published studies our ID cohort consists mostly of relatively mild ID cases. We identified a likely pathogenic variant in genes known to be associated with ID in 20% of the cases for which both exome and CNV data were available (Figure 5 A), explaining an estimated 4.5% of the heritability on the liability scale (Figure 7). Additionally, we observed a significant ~2-fold enrichment of damaging variants / CNVs in loss-of-function intolerant genes not yet linked to ID, which explained an additional 2.3% of the heritability on the liability scale (Figure 5 A and Figure 7). For the first time to our knowledge, we demonstrated that a common variant polygenic load is associated with ID. We observed educational attainment, IQ and schizophrenia polygenic risk scores to be associated with ID explaining an estimated 0.94%, 0.48% and 0.55% of the heritability on the liability scale, respectively.

We then focused on characterizing the genetic architecture of mild vs. more severe forms of ID and observed that a likely causative variant in known ID genes was significantly more often identified in more severe ID cases than in mild ID cases (Figure 5 C). This suggests that either mild ID has a more complex etiology or that variants in genes predisposing to mild ID are partly different than those predisposing to more severe forms of ID. Our observation is in agreement with epidemiological studies where mild ID has been suggested to represent a highly heritable low end of a normal distribution of IQ whereas severe ID is a distinct condition with different etiology[5]. Therefore mild ID should have less contribution from *de novo* and extremely rare variants, which have been the major focus of most genetic studies of ID.

To study the possibly more complex etiology of mild ID we first showed that the polygenic risk score of low educational attainment, low IQ and schizophrenia were all higher in the Eastern and Northern parts of Finland, coinciding with the more recent bottleneck and higher prevalence of intellectual disability and schizophrenia within those regions in Finland [10]. We then showed that the PRS for educational attainment, intelligence and SCZ all were significantly associated with ID in our cohort when compared to the genetically matched control population, thereby demonstrating the contribution of common low-risk variants to intellectual disability. This observation could be in part because most of our ID patients had mild ID. Indeed, the highest heritability explained (2.17%) was observed with EDU PRS in mild ID. The EDU PRS has been reported to explain 2.9% of the heritability of educational attainment in a population sample independent of the original GWAS[48]. Our results suggest that mild ID might be just a continuum of the population distribution of cognitive capacity and support the hypothesis of the polygenic background. The observation that the heritability explained by EDU PRS is clearly smaller in severe ID (0.55%) supports the earlier epidemiological findings that the genetic background of severe ID is different from mild ID[5], where penetrant mutations contribute more to the phenotype.

The PRS for IQ was only slightly below the matched controls in mild ID. This was unexpected. The reason remains speculative, but could potentially be contributed by the fact that the IQ PRS was generated from a smaller study samples (n=78,308) than the EDU score (n=293,723).

After observing a significant association between the common variant load and ID we hypothesized that PRSs would be different in those individuals in whom a likely pathogenic variant was identified and those where such variants were not identified. However such a difference was not observed, not even a suggestive trend (Supplementary Figure 4). This observation could be explained by assuming that rare high-risk variants and the common variant load act additively

to increase the risk of ID. Another explanation could be that there still might be other unidentified strong or moderate variants explaining the phenotype in many of the cases in which we were not able to conclusively identify a causative variant. We explored this hypothesis by grouping patients into the "Other high impact variant" category if they carried a PTV, CNV or damaging missense mutation in loss of function intolerant genes not previously linked to NDDs, but did not observe a difference in PRSs in that group either (Supplementary Figure 4). An additive effect of high impact rare variants and common variant polygenic load has recently been suggested in the genetic etiology of ASD[49], our data suggests a similar genetic architecture for ID.

Finally, we studied if some variants enriched in Finland in the relatively recent bottleneck would be associated with ID in our cohort. We conclusively identified a recessive variant in the *CRADD* gene enriched in Finland in three NFID patients and one NDD patient from the population NDD cohorts (Table 3). The allele frequency of this variant is 50x higher in the Finnish population than in non-Finnish Europeans. Recently recessive variants in *CRADD* have been reported in six patients from four families with megalencephaly, frontal predominant pachygyria, intellectual disability, and seizures [42]. All three of our patients had pachygyria, consistent with previously reported cases [42]. One of the patients identified in Di Donate *et al*. had Finnish origins and carries exactly the same homozygotic variant as our patients, clearly demonstrating that the variant is a causal for a specific syndrome. We also observed three cases that were homozygous missense variant carriers in the *INTS1* gene (Table 3). Recently a loss-of-function variants in *INTS1* have been identified in three unrelated moderate to severe ID patients [44]. One of our patients had mild ID and the two others had moderate/severe ID.

In the screen for dominant associations among Finnish enriched variants none of the variants surpassed stringent multiple testing correction (Supplementary Table 4). However one variant among the top 10 variants, a missense variant in the

DENR gene, was totally absent in non-Finnish GnomAD individuals, is very rare in the Finnish population but enriched in Northern Finland ($6.3x10^{-4}$ in GnomAD Finns; $9.7x10^{-4}$ in our Northern Controls and $3.1x10^{-4}$ in Southern controls). The variant replicated in the Northern NDD cohort and was extremely rare in Southern Finnish NDD cases and controls (1/322 in cases and 1/1,594 in controls) but had a high OR estimate consistent with associations in NFID and Northern NDD samples. Two *DENR* de novo missense variants have previously been identified in patients ascertained for autism spectrum disorder [50,51]. The case in Neale et al. had a full scale IQ of 67 and the case in Haas et al, had a language delay and poor comprehension. Two of the three *DENR* variant carriers in the NFID cohort had a suspected or confirmed ASD diagnosis. Eight individuals in the population NDD cases were schizophrenia patients. The SCZ cases had low scores on processing speed and verbal learning cognitive tests as compared to population controls (Supplementary Figure 10). ID or autism were not systematically diagnosed in the collection. Further studies are needed to conclusively determine if some of the other identified candidate genes are truly ID associated.

In conclusion, we demonstrate that a common variant polygenic load is a contributing factor in ID and more broadly characterized the genetic architecture of mild ID, which so far has been understudied. We also show that some damaging variants enriched in frequency in Finland contribute to intellectual disability and provide, yet another example of the power of utilizing population isolates such as Finland in disease gene mapping.

Author contributions

M.I.K performed the analyses; A.P, O.P and M.J.D conceived the study; E.S and D.L analyzed the CNV data; P.G. imputed the GWAS data; M.T-H and J.S provided schizophrenia cohort data; V.S provided population control data; S.K and M.P visualized the PRS geographical distributions; J.S.M, E.R, R.K-F, M.R,

S.K-H, J.K-E, H.H, P.V, J.K and O.K collected and performed the clinical examination of the NFID cohort. M.I.K and A.P wrote the manuscript. All authors critically revised the manuscript and read and approved the final version.

Data availability

All summary level data are available from the corresponding author on reasonable request.

The datasets generated during and/or analysed during the current study are not publicly available due to patient confidentiality and multiple different consents of population cohorts used but subset of the data are available from the corresponding author on reasonable request.

Code availability

All custom code used within the manuscript for all analyses is available from the corresponding author upon reasonable request.

Acknowledgements

We thank Social Science Genetic Association Consortium (SSGAC; www.thessgac.org), Psychiatric Genomics Consortium (PGC;Stephan Ripke; http://www.med.unc.edu/pgc) and Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research (http://ctg.cncr.nl/) for sharing summary statistics of their GWAS studies. We further thank SSGAC and PGC for kindly re-generating variant weights excluding Finnish cohorts.

1. American Psychiatric Association. *DSM 5*. *American Journal of Psychiatry* (2013). doi:10.1176/appi.books.9780890425596.744053
2. Ropers, H. H. Genetics of Early Onset Cognitive Impairment. *Annu. Rev. Genomics Hum. Genet* **11,** 161–87 (2010).
3. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17,** 9–18 (2016).
4. McRae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542,** 433–438 (2017).
5. Reichenberg, A. *et al.* Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proc. Natl. Acad. Sci.* **113,** 1098–1103 (2016).
6. Van Bokhoven, H. Genetic and Epigenetic Networks in Intellectual Disabilities. *Annu. Rev. Genet* **45,** 81–104 (2011).
7. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
8. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genet.* **10,** (2014).
9. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455-64 (2014).
10. Stoll, G. *et al.* Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. *Nat. Neurosci.* **16,** 1228–37 (2013).
11. Perälä, J., Saarni, S. I., Ostamo, A., Pirkola, S. & Haukka, J. Geographic variation and sociodemographic characteristics of psychotic disorders in Finland. *Schizophr. Res.* **106,** 337–347 (2008).
12. Peltonen, L., Jalanko, a & Varilo, T. Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.* **8,** 1913–23 (1999).
13. Jakkula, E. *et al.* The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population. *Am. J. Hum. Genet.* **83,** 787–794 (2008).
14. Kerminen, S. *et al.* Fine-Scale Genetic Structure in Finland. *G3 (Bethesda).* **7,** 3459–3468 (2017).
15. Isohanni, M. *et al.* A comparison of clinical and research DSM-III-R diagnoses of schizophrenia in a Finnish national birth cohort. Clinical and research diagnoses of schizophrenia. *Soc. Psychiatry Psychiatr. Epidemiol.* **32,** 303–8 (1997).
16. Kiviniemi, M. *et al.* Five-year follow-up study of disability pension rates in first-onset schizophrenia with special focus on regional differences and mortality. *Gen. Hosp. Psychiatry* **33,** 509–17 (2011).
17. Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland. *Eur. J. Public Health* **25,** 539–546 (2015).
18. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
19. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing

method for thousands of genomes. *Nat. Methods* **9,** 179–81 (2012).
20. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
21. Rivas, M. A. *et al.* A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. *Nat. Commun.* **7,** (2016).
22. Hail. Available at: https://github.com/hail-is/hail.
23. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
24. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017). doi:10.1101/148353
25. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** (2016).
26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
27. Kearney, H. M. *et al.* American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med.* **13,** 680–685 (2011).
28. 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).
29. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
30. 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).
31. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
32. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36,** (2010).
33. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case- control association testing of single low-count variants Clement. *Genet. Epidemiol.* **37,** 539–550 (2013).
34. Lim, E. T., Liu, Y. P., Chan, Y. & Tiinamaija, T. A Novel Test for Recessive Contributions to Complex Diseases Implicates Bardet-Biedl Syndrome Gene BBS10 in Idiopathic Type 2 Diabetes and Obesity. *Am. J. Hum. Genet.* **95,** 509–520 (2014).
35. Efron, B. Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* **82,** 171–185 (1987).
36. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).
37. Yagi, H. *et al.* Role of TBX1 in human del22q11.2 syndrome. *Lancet (London, England)* **362,** 1366–73 (2003).
38. Jonas, R. K., Montojo, C. A. & Bearden, C. E. The 22q11.2 Deletion Syndrome as a Window into Complex Neuropsychiatric Disorders Over the Lifespan. *Biol. Psychiatry* **75,** 351–360 (2014).

39. Yobb, T. M. *et al.* Microduplication and Triplication of 22q11.2: A Highly Variable Syndrome. *Am. J. Hum. Genet.* **76,** 865–876 (2005).
40. Riglin, L. *et al.* Schizophrenia risk alleles and neurodevelopmental outcomes in childhood: a population-based cohort study. *The Lancet Psychiatry* **366,** 1–6 (2016).
41. Schultz, J. M. *et al.* Noncoding Mutations of HGF Are Associated with Nonsyndromic Hearing Loss, DFNB39. *Am. J. Hum. Genet.* **85,** 25–39 (2009).
42. Di Donato, N. *et al.* Mutations in CRADD Result in Reduced Caspase-2-Mediated Neuronal Apoptosis and Cause Megalencephaly with a Rare Lissencephaly Variant. *Am. J. Hum. Genet.* 1–13 (2016). doi:10.1016/j.ajhg.2016.09.010
43. Menna, E. *et al.* Eps8 controls dendritic spine density and synaptic plasticity through its actin-capping activity. *EMBO J.* **32,** 1730–44 (2013).
44. Oegema, R. *et al.* Human mutations in integrator complex subunits link transcriptome integrity to brain development. *PLoS Genet.* **13,** 1–20 (2017).
45. Latypova *et al. Dominant RORA variants cause an intellectual disability syndrome associated with epilepsy, autistic features or cerebellar ataxia; (Abstract/Program #200). Presented at the 67th Annual Meeting of The American Society of Human Genetics, Date, Location (e.g*. (2017).
46. Tarpey, P. S. *et al.* Mutations in UPF3B, a member of the nonsense-mediated mRNA decay complex, cause syndromic and nonsyndromic mental retardation. *Nat. Genet.* **39,** 1127–33 (2007).
47. Chaudhry, A. *et al.* Phenotypic spectrum associated with PTCHD1 deletions and truncating mutations includes intellectual disability and autism spectrum disorder. *Clin. Genet.* **88,** 224–233 (2015).
48. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19,** 1563–1565 (2016).
49. Weiner, D. J. *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* (2017). doi:10.1038/ng.3863
50. Haas, M. A. *et al.* De Novo Mutations in DENR Disrupt Neuronal Development and Link Congenital Neurological Disorders to Faulty mRNA Translation Re-initiation Article De Novo Mutations in DENR Disrupt Neuronal Development and Link Congenital Neurological Disorders to Faulty . *CellReports* **15,** 2251–2265 (2016).
51. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–5 (2012).