Title: Copy number variations and cognitive phenotypes in unselected populations.

Authors: Männik K, Mägi R, Macé A, Cole B, Guyatt AL, Shihab HA, Maillard AM, Alavere H, Kolk A, Reigo A, Mihailov E, Leitsalu L, Ferreira AM, Nõukas M, Teumer A, Salvi E, Cusi D, McGue M, Iacono WG, Gaunt TR, Beckmann JS, Jacquemont S, Kutalik Z, Pankratz N, Timpson N, Metspalu A, Reymond A

Journal: JAMA

Year: 2015 May 26

Volume: 313

Issue: 20

Pages: 2044-54

DOI: 10.1001/jama.2015.4845

Original Investigation

Copy Number Variations and Cognitive Phenotypes in Unselected Populations

Katrin Männik, PhD; Reedik Mägi, PhD; Aurélien Macé, MSc; Ben Cole, BS; Anna L. Guyatt, MBChB; Hashem A. Shihab, PhD; Anne M. Maillard, PhD; Helene Alavere, MD, MSc; Anneli Kolk, MD, PhD; Anu Reigo, MD; Evelin Mihailov, MSc; Liis Leitsalu, MSc; Anne-Maud Ferreira, MSc; Margit Nõukas, MSc; Alexander Teumer, PhD; Erika Salvi, PhD; Daniele Cusi, PhD; Matt McGue, PhD; William G. Iacono, PhD; Tom R. Gaunt, PhD; Jacques S. Beckmann, PhD; Sébastien Jacquemont, MD; Zoltán Kutalik, PhD; Nathan Pankratz, PhD; Nicholas Timpson, PhD; Andres Metspalu, MD, PhD; Alexandre Reymond, PhD

Editorial page 2029

Supplemental content at jama.com

CME Quiz at jamanetworkcme.com and **CME Questions** page 2073

IMPORTANCE The association of copy number variations (CNVs), differing numbers of copies of genetic sequence at locations in the genome, with phenotypes such as intellectual disability has been almost exclusively evaluated using clinically ascertained cohorts. The contribution of these genetic variants to cognitive phenotypes in the general population remains unclear.

OBJECTIVE To investigate the clinical features conferred by CNVs associated with known syndromes in adult carriers without clinical preselection and to assess the genome-wide consequences of rare CNVs (frequency ≤0.05%; size ≥250 kilobase pairs [kb]) on carriers' educational attainment and intellectual disability prevalence in the general population.

DESIGN, SETTING, AND PARTICIPANTS The population biobank of Estonia contains 52 000 participants enrolled from 2002 through 2010. General practitioners examined participants and filled out a questionnaire of health- and lifestyle-related questions, as well as reported diagnoses. Copy number variant analysis was conducted on a random sample of 7877 individuals and genotype-phenotype associations with education and disease traits were evaluated. Our results were replicated on a high-functioning group of 993 Estonians and 3 geographically distinct populations in the United Kingdom, the United States, and Italy.

MAIN OUTCOMES AND MEASURES Phenotypes of genomic disorders in the general population, prevalence of autosomal CNVs, and association of these variants with educational attainment (from less than primary school through scientific degree) and prevalence of intellectual disability.

RESULTS Of the 7877 in the Estonian cohort, we identified 56 carriers of CNVs associated with known syndromes. Their phenotypes, including cognitive and psychiatric problems, epilepsy, neuropathies, obesity, and congenital malformations are similar to those described for carriers of identical rearrangements ascertained in clinical cohorts. A genome-wide evaluation of rare autosomal CNVs (frequency, ≤0.05%; ≥250 kb) identified 831 carriers (10.5%) of the screened general population. Eleven of 216 (5.1%) carriers of a deletion of at least 250 kb (odds ratio [OR], 3.16; 95% CI, 1.51-5.98; *P* = 1.5e-03) and 6 of 102 (5.9%) carriers of a duplication of at least 1 Mb (OR, 3.67; 95% CI, 1.29-8.54; *P* = .008) had an intellectual disability compared with 114 of 6819 (1.7%) in the Estonian cohort. The mean education attainment was 3.81 (*P* = 1.06e-04) among 248 (≥250 kb) deletion carriers and 3.69 (P = 5.024e-05) among 115 duplication carriers (≥1 Mb). Of the deletion carriers, 33.5% did not graduate from high school (OR, 1.48; 95% CI, 1.12-1.95; *P* = .005) and 39.1% of duplication carriers did not graduate high school (OR, 1.89; 95% CI, 1.27-2.8; *P* = 1.6e-03). Evidence for an association between rare CNVs and lower educational attainment was supported by analyses of cohorts of adults from Italy and the United States and adolescents from the United Kingdom.

CONCLUSIONS AND RELEVANCE Known pathogenic CNVs in unselected, but assumed to be healthy, adult populations may be associated with unrecognized clinical sequelae. Additionally, individually rare but collectively common intermediate-size CNVs may be negatively associated with educational attainment. Replication of these findings in additional population groups is warranted given the potential implications of this observation for genomics research, clinical care, and public health.

JAMA. 2015;313(20):2044-2054. doi:10.1001/jama.2015.4845

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Alexandre Reymond, PhD, Center for Integrative Genomics, University of Lausanne, Genopode Bldg, 1015 Lausanne, Switzerland (alexandre.reymond@unil.ch).

Recent studies showed that human individuals differ on approximately 0.8% of their genome.[1] The Database of Genomic Variants catalogs approximately 2.4 million DNA copy number variations (CNVs), genetic sequences that differ in numbers of copies in the human genome, mapping to approximately 200 000 unique loci that cover 72% of the human genome.[2] Copy number variations have been shown to contribute to interindividual variation in a wide variety of traits and conditions by globally influencing the transcriptome.[3-6] Large, defined herein as larger than 500 kb, recurrent CNVs have been associated with complex disorders, particularly developmental delay and intellectual disability[7,8] Intellectual disability is characterized by limited intellectual functioning and impaired adaptive behavior in everyday life. These CNVs are listed in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER)[9] and are often regrouped under the term *genomic disorder*.[8]

The effects of large CNVs in adult populations remain unclear because associations of large rare CNVs with pathologies were almost exclusively evaluated using clinically ascertained pediatric cohorts with known intellectual impairment. The purpose of this study was to investigate the characteristics of adult carriers of known pathological CNVs who were not clinically preselected and to assess the burden of rare intermediate-size autosomal CNVs (defined as 500 kb > CNV ≥250) on educational attainment and intellectual disability.

Methods

Estonian Genome Center, the University of Tartu, Cohort

The Estonian Genome Center, the University of Tartu (EGCUT), cohort is a population biobank containing 5% of the Estonian adult population.[10] Samples have been collected in all 15 Estonian counties and diverse social groups by 454 general practitioners (corresponding to 56% of practioners registered to the Estonian Health Board). The age, sex, and geographical distribution of the 52 000 participants closely reflect those of the Estonian adult population. The detailed description of the Estonian cohort was previously published.[10] At baseline, general practitioners performed a standardized objective examination of the participants and filled out a questionnaire that included more than 1000 health- and lifestyle-related questions, as well as provided the diagnoses of diseases present in the medical history of the participating individual using the format of the *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)*[10] (see details in eMethods of the Supplement). The data are continuously updated through periodic linking to national electronic health registries. The wide range of phenotypes, ages, and social groups makes the cohort ideally suited to population-based studies. For details on EGCUT cohort phenotype data (see eFigure 1 and the eMethods section in the Supplement). The Estonian Genome Center is conducted according to the Estonian Human Genes Research Act and is managed in conformity with the International Organization for Standardization ISO 9001:2008. The ethics review committee on Human Research of the University of Tartu approved the project. Written informed consent was obtained from all participants for the baseline and follow-up investigations.

The relevant phenotype traits of individual carriers of DECIPHER-listed syndromic CNVs in the EGCUT cohort (eTable 1 in the Supplement) were obtained from the baseline questionnaire and compared with the reviewed characteristics of corresponding syndromes (eTable 2 in the Supplement). To further investigate the clinical features of adult carriers not clinically preselected, all 16p11.2 600kb BP4-BP5 (breakpoint) deletion and reciprocal duplication carriers identified in the Estonian cohort were invited back for follow-up investigations. See the Results section for a detailed description and relevant references of this genomic disorder. These CNVs were selected because of their relatively high prevalence and variable phenotype. These carriers were phenotyped using the standardized clinical and neuropsychological protocol that had been developed previously to specifically study patients with the 16p11.2 syndromes who had been ascertained through clinical cohorts.[11,12] In agreement with the known population prevalence of 16p11.2 600kb BP4-BP5 CNVs,[11] 4 deletion carriers (0.05%) and 7 duplication carriers (0.09%) were identified in the Estonian set.

The EGCUT cohort (and Estonian population in general) is an outbred population with no substantial regional or ethnic differences. Single-nucleotide polymorphism (SNP) allele frequencies and linkage disequilibrium patterns are similar to those found in populations with European ancestry.[13] We did not find small series of nonrecurrent CNVs or inflation of recurrent rearrangements typical of founder effects[14,15] (eMethods in the Supplement). Accordingly, EGCUT samples have been successfully used to discover or replicate hundreds of SNP associations, which are vulnerable to population frequencies and stratification differences[16-18] (see eMethods and eFigure 2 in the Supplement for details on the Estonian population makeup and stratification).

CNV Calling

The genomic DNA of 8110 individuals (7020 for the discovery and 1090 for the replication cohorts; eTable 3 in the Supplement), randomly selected among the 52 000 EGCUT participants, was subjected to CNV analysis. A third cohort of 1066 individuals ("high-functioning replication cohort") was used to further assess the significance of the signal obtained regarding educational attainment. Due to the recruitment criteria that required participants to provide sufficient and consistent information in an advanced sleeping pattern-related questionnaire and a regular work schedule over a survey period of 6 months, the high-functioning replication cohort was biased toward higher than average sociocognitive functioning (eMethods in the Supplement). Single-nucleotide polymorphism genotyping and CNV calling were performed using Illumina platforms and the Hidden Markov Model-based software PennCNV according to the manufacturer's and developer's protocols,[19] respectively. The 6819 discovery, 1058 replication, and 993 high-functioning replication samples that passed the quality control parameters were retained (eMethods in the Supplement).

Genotype-Phenotype Correlations

The difference of studied phenotypes between CNV carriers and the general population was assessed. A 2-sided Fisher exact test and Welch 2-sample *t* test were used for statistical analysis in The R Project for Statistical Computing environment (http://www.r-project.org, R version 3.0.2). A threshold of $P \leq .05$ was set to indicate statistical significance. See eMethods in the Supplement for assessment of phenotype and determination of the prevalence of recurrent genomic syndromes. Briefly, intellectual disability is defined in the *Diagnostic and Statistical Manual of Mental Disorders* as a deficit in overall cognitive functioning along with limitations in adaptive behavior. It was diagnosed as described in F70-F79 of the *ICD-10*. All diagnoses, including intellectual disability, were diagnosed according to diagnostic standards throughout the participants' medical history and recorded before enrollment to the biobank. It is reported to the EGCUT database by the recruiting general practitioners. Intellectual disability prevalence is estimated at 1% to 3% in developed countries,[20] which is consistent with the prevalence found in the EGCUT discovery cohort (1.7%). Educational levels were uniformly coded at the time of enrollment according to the Estonian education curriculum from 1 to 7, ie, from less than primary school through scientific degree (eMethods in the Supplement). In both discovery and replication cohorts, the mean education attainment (MEA) corresponded to secondary education (MEA, 4.09; 95% CI, 4.07-4.12 for the discovery cohort and 4.0; 95% CI, 3.93-4.05 for the replication cohort) in agreement with the country's MEA. See eMethods in the Supplement for details on the Estonian population religiousness, school curriculum organization, and education system performance.

Function of CNV-Embedded Genes

Three previously published data sets were used to functionally annotate genes embedded in rare CNVs and assess if their characteristics could be used to predict CNV deleteriousness (eMethods in the Supplement): (1) the neurodevelopmental gene list[21,22]; (2) the haploinsufficiency scores (HiS), ie, the probability that a given gene does not maintain its normal function with only 1 functional copy[23]; and the list of ohnologs, ie, genes related by ancestral whole-genome duplication events.[24] Because a CNV may preserve a gene's integrity yet indirectly affect it through changes in the copy-number of its regulatory elements,[4,5,25] the potential contributions of the latter was also tested by stratifying CNVs using the number of encompassed regulatory elements identified in.[26] To further assess the functions of imbalanced genes, we used Thomson Reuters *MetaCore*, an integrated software suite for data-mining and pathway analysis based on a manually curated biological knowledge database (eMethods in the Supplement).

Avon Longitudinal Study of Parents and Children Cohort

The Avon Longitudinal Study of Parents and Children (ALSPAC) cohort is a birth cohort based in Bristol, United Kingdom,[27] which initially enrolled 14 541 pregnant women with expected delivery dates between April 1, 1991, and December 31, 1992. The 13 988 children who were alive at 1 year of age. Additional families were enrolled in later phases. Detailed phenotypic information on the children and their parents was collected during clinic visits and by the completion of questionnaires, as well as from linkage with external data sources (eMethods in the Supplement). Ethical approval for the study was obtained from ALSPAC Ethics and Law Committee and the local research ethics committees.

The Illumina HumanHap550 Quad platform was used to genotype 9912 children in ALSPAC. The CNVs were called with PennCNV.[19] The subset of 5218 unrelated individuals, who passed the array quality control, who gave consent, and for whom educational information was available were retained for analysis (eTable 3 and eMethods in the Supplement). Log R ratio (LRR) and B allele frequency metrics were derived from raw data using published guidelines.[28]

Within ALSPAC, educational attainment was assessed using data from the UK-based Key Stage 3 National Curriculum Tests in English and mathematics, taken at ages 13 and 14 years, also known as Standard Assessment Tests (SATs). A discrete level is awarded for these tests, but to further account for the exact score received and the fact that the maximum and minimum level achievable for mathematics was dependent on the tier of examination for which the child was entered, results were scaled and adjusted as described previously.[29,30] Due to nonnormal distribution of the data, these 2 variables were then inverse-rank normal transformed and then standardized. Furthermore, tertiles of the English and mathematics scores were created (eTable 4 in the Supplement). Differences in MEAs according to rare CNV carrier status (frequency ≤0.05%) were compared using a Welch 2-sided *t* test. This was performed separately for each of the inverse-rank transformed, standardized English and mathematics educational attainment scores. To obtain an interpretable estimate of effect, univariable logistic regression models were assessed separately for English and mathematics. The top tertile was coded as the reference group and the bottom tertile as the risk group. Separate odds ratios (ORs) were estimated for membership of the risk group, comparing CNV carriers corresponding to increasing size groups against baseline (no large CNVs at a frequency ≤0.05%). The binary educational outcome was then regressed against CNV carrier status as an ordered variable, including all 4 size categories, and the *P* value was reported as an assessment of trend.

Minnesota Center for Twin and Family Research Cohort

Participants from 2 studies conducted by the Minnesota Center for Twin and Family Research were used as replication samples: the Sibling Interaction and Behavior Study, and the Minnesota Twin Family Study, which is a longitudinal study of a community-based sample of same-sex twins born between 1972 and 1994 in Minnesota and their parents.[31] The Sibling Interaction Behavior Study is an adoption study of sibling pairs and their parents[32]; its community-based sample contains families in which both siblings are adopted, in which both are biologically related to the parents, or in which one is adopted and one is biologically related. In the current analyses, only a single random individual was selected for inclusion in analyses in order to create a data set of unrelated participants (n = 2390, eTable 3 in the Supplement). The collection, genotyping, and analysis of DNA samples for both studies were approved by the University of Minnesota Institutional Review Board's Human Subjects Committee. Written informed consent was obtained from all participants; parents provided written informed consent for their minor children.

Genotyping was performed using the Illumina 660W-Quad array. Whole blood extracted DNA samples were only analyzed if the participant was white non-Hispanic and the standard deviation of the GC-corrected[33] autosomal log R ratios was less than 0.20. The CNVs were called using PennCNV and then processed and filtered. Adjacent CNVs were merged if they had the same copy number and if the number of markers in the intervening gap was less than 20% of the number of total markers spanning the called CNVs. To replicate the results of the EGCUT discovery cohort the same parameters, ie, rare (frequency, ≤0.05%) deletions of 250 kb or longer and duplications of 1 Mb or longer were retained in the burden analysis.

The full-scale intelligence quotient (FSIQ) was estimated using an abbreviated form of either the Wechsler Intelligence Scale for Children-Revised (WISC-R; for children ≤16 years) or the Wechsler Adult Intelligence Scale-Revised (WAIS-R; for individuals ≥16 years). The short forms consisted of 2 performance subtests (Block Design and Picture Arrangement) and Verbal subtests (information and vocabulary) and were prorated to determine FSIQ. Estimates from this short form have been shown to correlate 0.94 with FSIQ from the complete test.[34] Samples with multiple FSIQ measurements were averaged together for analysis (mean, 104.52; SD, 4.27; range, 67-150).

HYPERGENES Italian Cohort

The Italian cohort follow-up is based on 451 individuals belonging to the cohort ascertained as controls for genome-wide association studies of hypertension (HYPERGENES)[35] (eTable 3 in the Supplement). *Years of schooling* was defined in accordance with the International Standard Classification of Education 1997 classification, leading to 7 categories of educational attainment that are internationally comparable (eMethods in the Supplement). Single-nucleotide polymorphisms were genotyped using Illumina Human 1M-Duo BeadChips and CNVs called with PennCNV as for the EGCUT discovery cohort. Differences in means of educational attainment were compared using a Welch 2-samples 1-tailed *t* test and Wilcoxon rank-sum test in R. Both tests returned comparable results.

Results

Prevalence and Phenotypes of Pathological CNVs in Estonian Cohort

To investigate the medical burden of rare CNVs in the general population, we opted for a genotype-first approach and analyzed a random sample from the EGCUT cohort. Within a combined discovery and replication sample of 7877 unrelated individuals, 56 carriers (0.7%) of known recurrent autosomal genomic disorders were identified (eTable 1 in the Supplement). Although the prevalence of each genomic disorder is lower than previously reported in clinical cohorts,[36,37] it is only slightly lower than the 67 individuals expected according to the reported population prevalence of the 57 autosomal syndromes listed in the DECIPHER database of genomic disorders[9] (eTable 2, eTable 5, and eMethods in the Supplement). The EGCUT cohort is depleted (6 observed carriers of 17 expected, OR, 0.35; 95% CI, 0.11-0.94; *P* = .03) of the most deleterious CNVs (graded 1-2 by DECIPHER), whereas the frequency of CNVs graded 3 and ungraded is as expected (50 of 50; OR, 1; CI 95%, 0.66-1.51; *P* > .99).

The clinical features of the EGCUT carriers of DECIPHER-listed CNVs are comparable with those reported in disease cohorts. Thirty-one of 56 (55%; including only formal diagnosis) and 39 of 56 (70%; including self-reported problems) carriers recruited from the general population with no prior awareness of their genetic disorder present phenotypes previously associated with their genomic lesion in the literature (see eTable 1 for the phenotypes identified in the 56 EGCUT carriers and eTable 2 for phenotypes associated with DECIPHER-listed CNVs in the Supplement). For example, carriers of the 16p11.2 600kb BP4-BP5 deletions and reciprocal duplications identified in clinical cohorts show opposite phenotypes on body weight, head size, and volume of specific corticostriatal structures. They exhibit reduced FSIQ, as well as neuropsychiatric problems and congenital abnormalities.[11,12,38-44] Correspondingly, the baseline questionnaires of the 4 deletion (case Nos. 41-44 in eTable 1 in the Supplement) and 7 duplication (Nos. 45-51) carriers identified in the EGCUT cohort indicated high and low body mass indexes, respectively, neuropsychiatric traits, and learning and developmental problems. The follow-up evaluation of these carriers uncovered additional similarities in the spectrum and severity distribution of phenotypic features found in 16p11.2 BP4-BP5 rearrangement carriers identified through pan-European recruitment via clinical genetics centers (eTable 6 in the Supplement).

Rare Intermediate-Size CNVs and Educational Attainment

A genome-wide map of rare autosomal CNVs in the discovery set of 6819 individuals was generated (eTable 3 in the Supplement) and a total of 216 deletion and 509 duplication carriers were identified (≥ 250 kb with carrier frequency of ≤0.05%; eTable 7 in the Supplement). The underrepresentation of those with deletions compared with those with duplications (*P* = 2.2e-16) is consistent with previous reports and concordant with the hypothesis that the former are more deleterious.[1,14] We found evidence for an association between carrier status and prevalence of intellectual disability. Twenty-three individuals, equal to a 3.2% prevalence, were diagnosed with intellectual disability in the rare CNV carriers group vs 114 intellectual disability diagnoses (1.7%) in the EGCUT cohort (OR, 1.93; 95% CI, 1.17-3.06; *P* = .007). This finding was associated with deletions, 11 individuals (5.1%) had intellectual disability (OR, 3.16; 95% CI, 1.51-5.98; *P* = 1.5e-03). The prevalence of intellectual disability was higher in the carriers of DECIPHER-listed CNVs with 4 diagnosed individuals out of 45 (8.9%; OR, 5.74; 95% CI, 1.47-16.22; *P* = 7.2e-03). The difference with EGCUT remained statistically different even after exclusion of this group with known disease causing CNVs; the remaining 19 individuals with an intellectual disability diagnosis correspond to a prevalence of 2.8% (OR, 1.64; 95% CI, 0.95-2.71; *P* = .05).

Table 1. Prevalence of Intellectual Disability Diagnosis in Estonian Genome Center, the University Tartu, Cohort[a]

Cohort	Sample Size	No. of Individuals With Intellectual Disability	Prevalence, %	OR (95% CI)	*P* Value[b]
EGCUT population	6819	114	1.7		
DECIPHER-listed CNV carriers	45	4	8.9	5.74 (1.47-16.22)	7.2e-03
Deletion carrier by CNV size					
≥1 Mb	36	3	8.3	5.34 (1.03-17.42)	2.3e-02
≥500 kb	77	5	6.5	4.08 (1.26-10.25)	1e-02
≥250 kb	216	11	5.1	3.16 (1.51-5.98)	1.5e-03
500 kb ≤ to <1Mb	41	2	4.9	3.02 (0.35-11.9)	1.5e-01
250 kb ≤ to <500 kb	139	6	4.3	2.65 (0.94-6.11)	3.2e-02
Duplication carrier by CNV size					
≥1 Mb	102	6	5.9	3.67 (1.29-8.54)	8.3e-03
≥500 kb	235	8	3.4	2.07 (0.86-4.29)	6.6e-02
≥250 kb	509	12	2.4	1.42 (0.71-2.6)	2.9e-01
500 kb ≤ to <1 Mb	133	2	1.5	0.87 (0.11-3.38)	>.99
250 kb ≤ to <500 kb	274	4	1.5	0.87 (0.23-2.32)	>.99

Abbreviations: CNV, copy number variation; DECIPHER-listed CNV correspond to all syndromic CNV listed within the DECIPHER database (see Methods section); EGCUT, Estonian Genome Center, the University of Tartu (EGCUT); kb, kilobase; OR, odds ratio.

[a] The results are presented as cumulative or as size separated groups.

[b] Statistical significance was determined by comparing the prevalence of intellectual disability of CNV carriers with those of the Estonian population sampled (see eMethods and Leitsalu et al[10]).

We next assessed the correlation between CNV size and intellectual disability. It was previously reported that cohorts of affected patients show an excess of CNVs compared with controls and that this excess is larger for longer CNVs.[7] The frequency of intellectual disability increases with CNV size: 6 individuals (4.3%) with deletion ranging from 250 kb to 500 kb (OR, 2.65; 95% CI, 0.94-6.11; *P* = .03) vs 36 (8.3%) with at least 1-Mb deletions (OR, 5.34; 95% CI, 1.03-17.42; *P* = .02), whereas associations with duplications are only detectable when rearrangements exceed 1 Mb in size (102 diagnosed individuals [5.9%]; OR, 3.67; 95% CI, 1.29-8.54; *P* = .008; **Table 1**). Among the 275 individuals with smaller deletions (125 kb ≤CNV <250 kb) no apparent association existed (7 diagnosed individuals [2.5%]; OR, 1.5; 95% CI, 0.59-3.28; *P* = .24).

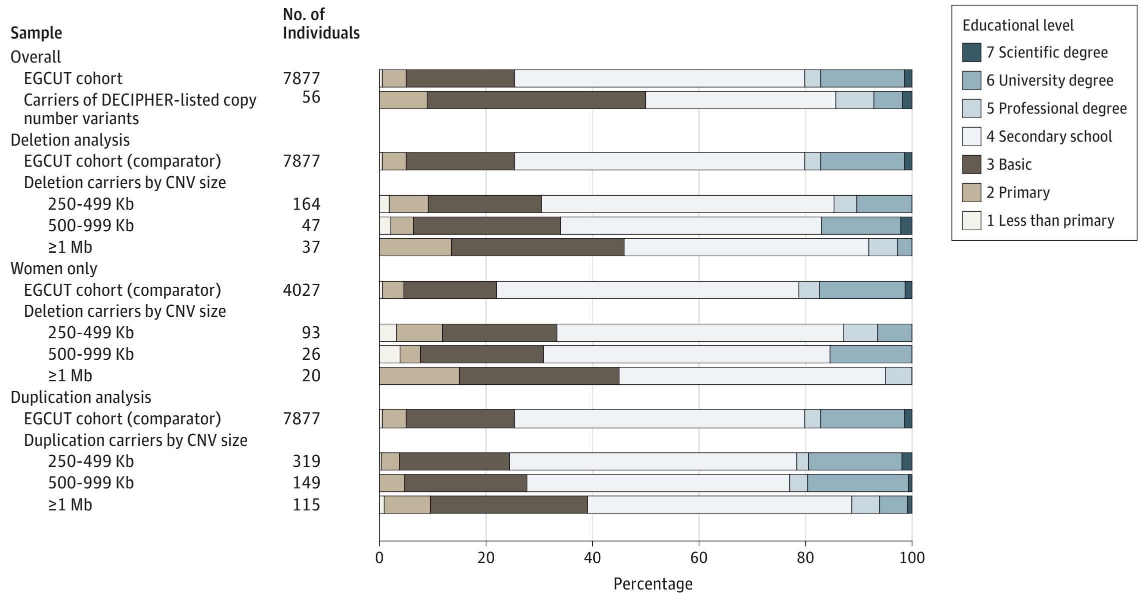
The diagnosis of intellectual disability is binary. Thus, to assess the effects of rare CNVs with greater granularity, we investigated whether their occurrence and size are related to achieved educational levels, a proxy for global cognition.[45,46] For this purpose, we used the scale of 7 sublevels of the Estonian education curriculum (eMethods in the Supplement). Although 1729 individuals (25.3%) sampled in the Estonian cohort did not complete secondary school (level 4; MEA, 4.09; 95% CI, 4.07-4.12), which was similar to the at-large Estonian population,[10] the proportion of those who did not complete secondary school is higher among carriers of DECIPHER-listed genomic disorders with 22 (48.9%) only reaching elementary or basic education (OR, 2.8; 95% CI, 1.49-5.3; *P* = 8.3e-04; MEA, 3.71; 95% CI, 3.38-4.04; *P* = .03; **Figure**). The fraction of carriers who failed to reach secondary education was associated with CNV size. For example, the carriers of CNVs of 1 Mb or larger have an MEA of 3.65 (95% CI, 3.49-3.81; *P* = 4.6e-07), and 56 (40.6%) of them did not complete secondary school (OR, 2.01; 95% CI, 1.40-2.87; *P* = 1e-04; Figure). Deletions are associated with most of the outcome, with MEAs decreasing to 3.5 (95% CI, 3.20-3.80; *P* = 4e-04); 17 carriers (47.2%) of those with a deletion of 1 Mb or larger did not complete their secondary education (OR, 2.63; 95% CI, 1.28-5.36; *P* = .006; Figure). A decrease is already seen in the group with deletions ranging from 250 to 500 kb who had an MEA of 3.86 (95% CI, 3.68-4.05; *P* = .1.7e-02) and 41 carriers (29.5%) not graduating from secondary school (OR, 1.23; 95% CI, 0.83-1.80; *P* = .28). Consistent with the intellectual disability results, 275 individuals with smaller deletions (125 kb ≤CNV <250 kb) were not associated with changes in educational attainment (MEA, 4.11; 95% CI, 3.97-4.25; *P* = .80), 72 (26.2%) of them had less than secondary education (OR, 1.04; 95% CI, 0.78-1.38; *P* = .78). Similarly, duplications were associated with an educational attainment decrease only when rearrangements were 1 Mb or larger (MEA, 3.71; 95% CI, 3.52-3.90; *P* = 1.5e-04) and 39 carriers (38.2%) did not complete secondary school (OR, 1.82; 95% CI, 1.19-2.77; *P* = 4.2e-03; Figure).

The EGCUT ancestry principal components are not associated with CNV burden (eFigure 2 in the Supplement), suggesting that genetic stratification is likely not confounding the association with educational attainment. Likewise, differences in educational achievement possibilities due to religion or ethnicity was not likely to account for the observed associations, as the surveys of the Organization for Economic Cooperation and Development (OECD) Program for International Student Assessment and Program for the International Assessment of Adult Competencies showed that the "free education for all" Estonian system is among the best in the world in terms of results and equal opportunity (eMethods in the Supplement).

Estonian Replication

A replication of the education analysis was conducted on a nonoverlapping random set of 1058 unrelated EGCUT individuals recruited similarly (eTable 3 in the Supplement) but sampled at a different time point and genotyped using a different array platform. Of those, 271 (25.6%) did not complete secondary school (MEA, 4.00; 95% CI, 3.93-4.05). However, carriers of deletion with sizes ranging from 250 kb to 500 kb, congruent with the discovery cohort, were associated with a nonsig-

Figure. Rare Intermediate-Size Copy Number Variations Associated With Lower Education Metrics



Comparison of educational achievement of participants in the Estonian Genome Center, University of Tartu (EGCUT) with carriers of Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER)–listed rearrangements or with carriers of deletions, female carriers of deletions, and duplications segregated by size (CNV frequencies ≤0.05%). The educational attainment decreases with copy number variation (CNV) size. See Table 2 for statistically significant differences between groups. Educational levels are coded according to the Estonian education curriculum (eMethods in the Supplement).

nificant trend suggesting a diminished educational attainment (MEA, 3.68; 95% CI, 3.39-3.97; *P* = .056), of whom 9 (36%) had only achieved a basic education or less (OR, 1.63; 95% CI, 0.63-3.98; *P* = .25). Similarly, duplication carriers with rearrangements of 1 Mb or larger had an MEA of 3.54 (95% CI, 2.97-4.11; *P* = .15) of whom 6 (46.2%) had achieved only basic education or less (OR, 2.49; 95% CI, 0.68-8.72; *P* = .11). The joint analyses of these 2 random cohorts confirmed the negative association on educational attainment among those with rare deletions of at least 250 kb (MEA, 3.81; 95% CI, 3.67-3.94; *P* = 1.06e-04) and 83 individuals (33.5%) of 248 in this group achieved less than secondary school (OR, 1.48; 95% CI, 1.12-1.95; *P* = .005). The same held true for duplications of 1 Mb or larger (MEA, 3.69; 95% CI, 3.51-3.87; *P* = 5.024e-05), 45 (39.1%) achieved less than secondary school (OR, 1.89; 95% CI, 1.27-2.8; *P* = 1.6e-03) (Figure and **Table 2**). We challenged these results further using a non-overlapping set of 993 unrelated EGCUT individuals biased toward higher than average sociocognitive functioning due to the different ascertainment criteria (eTable 3 and eMethods in the Supplement): MEA, 4.77; 95% CI, 4.69-4.84, lower than secondary education 9.4% [n = 93]). Even in a group that is probably partially depleted of severe CNVs, there was a trend toward a lower MEA among carriers of deletions ranging from 250 kb to 500 kb and duplications of 1Mb or larger of the same order of magnitude as in the discovery cohort (MEA, 4.36; 95% CI, 3.71-5.00; Δ, −0.41 and MEA, 4.44; 95% CI, 3.57-5.32; Δ, −0.33, respectively). Combining both independent replication cohorts confirmed the above results (MEA, 4.36 within the replication cohorts; MEA, 3.91 in the group of carriers of deletions ranging from 250-500 kb [95% CI, 3.63-4.20]; *P* = .004]; MEA, 3.79 in the group of carriers of duplication 1 Mb or larger [95% CI, 3.24-4.34; *P* = .057]). The same held true when all 3 Estonian cohorts were analyzed together (eTable 8 in the Supplement).

ALSPAC, Italian, and Minnesota Cohort Follow-up

We sought to strengthen the inference from our results using the SATs scores of 5218 members of the ALSPAC birth cohort as an alternative measure of educational attainment (eTable 4 in the Supplement). When the MEA was studied using the transformed variables, mathematics scores were lower in carriers of rare intermediate-size deletions than in the controls (250 kb ≤CNV <500 kb: Welch 2-sided *t* test comparing means, *P* = .019), and English language scores were lower in carriers of large deletions (≥1 Mb, Welch 2-sided *t* test comparing means, *P* = .020; eTable 9 in the Supplement). Mean education attainment in English language and mathematics was lower in those who carried large duplications (≥1 Mb; *P* = .020 and *P* = .049, respectively, Welch 2-sided *t* test). These results support the association between educational attainment and rare CNVs using a different education metrics in a geographically distinct and differently ascertained cohort of adolescents. Larger CNV size was associated with the odds of individuals belonging to the lowest tertile of SATs score for both English language and mathematics (eTable 4 in the Supplement). This was apparent both for carriers of deletions. For English language test results, carriers of deletions of 250 kb ≤CNV <500 kb had an OR of 1.26 (95% CI, 0.81-1.95), deletions of 500 kb ≤CNV <1 Mb had an OR of 1.69 (95% CI, 0.88-3.30), and deletions of 1 Mb or larger had an OR of 4.18 (95% CI, 1.48, 14.87; *P* for

Table 2. Educational Attainment in Estonian Genome Center, the University of Tartu Cohort (Joint Analysis of Discovery and Replication Cohorts)[a]

Cohort	Sample Size	Mean Education Attainment (95% CI)[b]	*P* Value[c]	No. of Individuals Not Reaching Secondary Education	Prevalence, %	No. of Individuals Not Reaching Secondary Education, OR (95% CI)	*P* Value[c]
Estonian population	7877	4.08 (4.10-4.05)		2000	25.4		
DECIPHER-listed CNV carriers	56	3.64 (3.92-3.37)	3.0e-03	28	50	2.94 (1.67-5.16)	8.334e-05
Deletion carrier by CNV size							
≥1 Mb	37	3.51 (3.80-3.22)	4.0e-04	17	46	2.5 (1.23-5.03)	7.2e-03
≥500 kb	84	3.75 (3.98-3.52)	5.7e-03	33	39.3	1.9 (1.18-3.01)	5.4e-03
≥250 kb	248	3.81 (3.94-3.67)	1.06-e04	83	33.5	1.48 (1.12-1.95)	5.0e-03
500 kb ≤ to <1 Mb	47	3.93 (4.28-3.59)	3.83e-01	16	34.0	1.52 (0.77-2.87)	1.8e-01
250 kb≤ to <500 kb	164	3.84 (4.00-3.67)	4.1e-03	50	30.5	1.29 (0.9-1.82)	1.5e-01
Duplication carrier by CNV size							
≥1 Mb	115	3.69 (3.87-3.51)	5.024e-05	45	39.1	1.89 (1.27-2.8)	1.6e-03
≥500 kb	264	3.92 (4.05-3.79)	2.5e-02	86	32.6	1.42 (1.08-1.86)	1.0e-02
≥250 kb	583	4.04 (4.13-3.95)	4.93e-01	164	28.1	1.15 (0.95-1.39)	1.54e-01
500 kb ≤ to <1 Mb	149	4.10 (4.29-3.93)	8.19e-01	43	28.9	1.19 (0.81-1.72)	3.4e-01
250 kb≤ to <500 kb	319	4.14 (4.27-4.02)	2.95e-01	78	24.5	0.95 (0.72-1.24)	7.4e-01

Abbreviations: CNV, copy number variations; EGCUT, Estonian Genome Center, the University of Tartu; kb, kilobase; OR, odds ratio.

[a] The results are presented as cumulative or as size-separated groups. DECIPHER copy number variations (CNVs) correspond to all syndromic CNVs listed within the DECIPHER database (see Methods section).

[b] Academic levels are based on the Estonian education curriculum: less than primary, 1; primary, 2; basic, 3; secondary, 4; professional or college, 5; university or academic, 6; scientific degree, 7.

[c] Statistical significance was determined by comparing the educational achievements and fraction of individuals with lower educational attainment of CNV carriers with those of the Estonian population (eMethods in the Supplement and Leitsalu et al[10]).

trend = .002). For mathematics test results, carriers of deletions of 250 kb ≤CNV <500 kb had an OR of 1.42 (95% CI, 0.91-2.21); deletions of 500 kb ≤CNV <1Mb had an OR of 2.21 (95% CI, 1.01-5.06); and deletions of 1 Mb or larger had an OR of 3.69 (95% CI, 1.51-10.29; *P* for trend, < 2.0e-04). Substantive evidence for an association of duplications and educational attainment was only observed for English language test results, for which carriers with duplications of 250 kb ≤ CNV < 500 kb had an OR of 1.14 (95% CI, 0.81-1.61); carriers with duplications of 500 kb ≤CNV <1 Mb had an OR of 1.19 (95% CI, 0.76-1.87); and carriers with duplications of 1 Mb or larger had an OR of 2.22 (95% CI, 1.07-4.84; *P* for trend = .035). For mathematics, carriers with duplications of 250 kb ≤CNV <500 kb had an OR of 1.10 (95% CI, 0.78-1.54); carriers with duplications of 500 kb ≤CNV <1 Mb had an OR of 1.03 (95% CI, 0.68-1.55); carriers with duplications of 1 Mb or larger had an OR of 1.54 (95% CI, 0.80-3.01; *P* for trend = .27; **Table 3**).

These results were followed up in 2 separate cohorts of healthy individuals with normal cognitive functioning (eMethods in the Supplement). Consistent with this ascertainment, both the Italian and Minnesota cohorts suggested a paucity of DECIPHER-listed CNVs (1 observed vs 4 expected; *P* = .37; OR, 0.25; CI 95%, 0.005-2.53) among the Italian cohort and (14 vs 20; *P* = .39; OR, 0.7; 95%, CI; 0.32-1.46) among the Minnesota cohort (eTable 2). Of note, the analysis of the Italian cohort was restricted by a small sample size (n = 451; eTable 3) resulting in both a limited statistical power and limited CNV frequency calculation (≥0.25%). At this 5-fold higher level of prevalence, the MEA was lower in carriers of deletion 500 kb ≤CNV <1 Mb (Δ MEA = −0.26; *P* = .39, Wilcoxon test) and carriers of duplications of 1 Mb or larger (Δ MEA = −0.66; *P* = .11; Wilcoxon test; eTable 10 in the Supplement). A consistent, but similarly underpowered, association with lower FSIQ was found in carriers of rare deletions in the Minnesota cohort (500 kb ≤CNV <1 Mb, Δ = −4.23 IQ points, *P* = .43; ≥1 Mb, Δ = −13.82 IQ points, *P* = .09) and duplications (500 kb ≤ CNV <1 Mb, Δ = −5.56, *P* = .01; ≥ 1Mb, Δ = −6.03, *P* = .16; eTable 11 in the Supplement).

Female Mutation Burden in EGCUT

In contrast to duplication carriers (male:female ratio, 1.06 (303: 285), an excess of female carriers was observed in every deletion size class of 250 kb or larger separately and together within the combined EGCUT discovery and replication cohort (male: female ratio, 0.78 (109:139); *P* = .14, OR, 1.22; 95% CI, 0.94-1.59). The reduction of MEA is greater in female carriers than in the male carriers (Δ MEA, −0.42 for females and −0.02 for males; Figure). Specifically, the female carriers of the 250 kb ≤ CNV <500 kb deletion had an MEA of 3.71 (95% CI, 3.50-3.92) compared with an MEA of 4.13 (95% CI, 4.09-4.16; *P* = 3e-04) in EGCUT females. The male carriers of similar size deletion had an MEA of 4.00 (95% CI, 3.76-4.24), whereas EGCUT males had an MEA of 4.02 (95% CI, 3.99-4.06; *P* = .85). Note that although 855 women (21.2%) in EGCUT earned college or academic degrees, the presence of a rare deletion is associated with a decreased fraction of women reaching the highest educational levels—levels 5 through 7. Only 12 women (12.9%) in the 250 kb ≤ deletion <500 kb group attained these highest educational levels (OR, 0.55; 95% CI, 0.27-1.02; *P* = .05; Figure). For example, only 1 of 20 women carrying deletions

Table 3. Univariable Logistic Regression Models for Performance in the SATs Assessment in Avon Longitudinal Study of Parents and Children Copy Number Variation Carriers[a]

	English Language				Mathematics			
			SATs Score, Median (IQR)				SATs Score, Median (IQR)	
Exposure	Score, OR (95% CI)	*P* Value for Trend	CNV Carriers	Controls	Score, OR (95% CI)	*P* Value for Trend	CNV Carriers	Controls
Deletion carrier by CNV size		2.0e-03				2.0e-04		
250 kb ≤ to <500 kb	1.26 (0.81-1.95)		6.18 (4.77-6.71)	6.24 (5.00-6.77)	1.42 (0.91-2.21)		5.67 (4.87-7.37)	7.08 (5.10-7.62)
500 kb ≤ to <1 Mb	1.69 (0.88-3.30)		5.24 (4.62-6.65)		2.21 (1.01-5.06)		5.57 (4.74-7.46)	
≥1 Mb	4.18 (1.48-14.87)		5.10 (4.46-5.33)		3.69 (1.51-10.29)		5.17 (4.18-6.72)	
Duplication carrier by CNV size		3.5e-02				2.73e-01		
250 kb ≤ to <500 kb	1.14 (0.81-1.61)		6.24 (5.00-6.75)	6.24 (5.00-6.77)	1.10 (0.78-1.54)		7.05 (5.03-7.52)	7.08 (5.10-7.62)
500 kb ≤ to <1 Mb	1.19 (0.76-1.87)		6.18 (5.10-6.62)		1.03 (0.68-1.55)		7.04 (5.32-7.51)	
≥1 Mb	2.22 (1.07-4.84)		5.23 (4.87-6.46)		1.54 (0.80-3.01)		5.46 (4.78-7.40)	

Abbreviations: CNV, copy number variation; IQR, interquartile range; kb, kilobase pair; OR, odds ratio; SATs, standard assessment tests.

[a] Univariable logistic regression models for performance in the SATs assessment according to CNV status (carrier frequency ≤0.05%). For each participant, the top tertile was coded as the reference tertile, and the bottom tertile as the risk tertile. The exposure was CNV carrier status, divided into 4 size groups. For each of the deletion and duplication analyses, ORs and 95% CIs are presented, for which each binary exposure is each CNV size group separately, in comparison with the control CNV group. Medians and IQRs for the derived SATs score variables are shown (see eMethods in the Supplement). The *P* values estimating trend were calculated by fitting an univariable logistic model, estimating OR of the binary educational outcome per increase in CNV size group. For both deletions and duplications, Controls are those individuals carrying neither a deletion nor a duplication of 250 kb or larger.

of 1 Mb or larger achieved more than a secondary school education. She carried the 17p12 deletion, which is causative for hereditary neuropathy with liability to pressure palsies (*HNPP* OMIM 162500). The joint analysis of the 3 Estonian cohorts confirmed that female deletion carriers are responsible for the majority of the decrease in educational attainment. All the EGCUT women combined had an MEA of 4.22 (95% CI, 4.19-4.25), whereas women with the 250 kb ≤ CNV <500 kb deletion had an MEA of 3.71 (95% CI, 3.54-3.88; *P* = 3.9e-08); 920 women (20.3%) achieved basic education or less, 49 deletion female carriers (33.6%; OR, 1.99; 95% CI, 1.37-2.85; *P* = 2.4e-04; eTable 8 in the Supplement). Consistent with the Estonian results, the Minnesota cohort female carriers of deletions of 500 kb or larger were associated with a stronger decrease of FSIQ (Δ = −13.73; *P* = .03) than were male carriers (Δ = −0.12; *P* = .98; eTable 11 in the Supplement).

Assessment of CNV Deleteriousness and Function

Investigating the functions of the 642 protein-coding genes encompassed in the identified rare deletions of 250 kb or longer, we found evidence of enrichment for genes with a role in neurogenesis, cognition, learning, memory, and behavior (29 of the top 50 gene ontology processes with strongest evidence; all with a false discovery rate of less than 2.45e-05; eTable 12 in the Supplement). We then assessed whether we could use gene characteristics to more accurately predict CNV deleteriousness. Copy number variations were stratified by the number of embedded protein-coding and noncoding genes, neurodevelopmental genes,[22] ohnologs,[24] the sum of imbalanced genes' probability score for haploinsufficiency (HiS),[23] and the highest HiS in the CNV. A decrease of cognitive abilities was present in carriers of deletions encompassing 2 or more genes (MEA, 3.82; 95% CI, 3.65-3.99; *P* = .003) and duplications including 11 or more genes (MEA, 3.74; 95% CI, 3.56-3.92; *P* = 3e-04; eFigure 3 in the Supplement). When genes were present in the rearranged interval, deleteriousness was associated with the presence of at least 1 protein-coding gene. Within the group of carriers of deletions with these characteristics, 8 individuals were diagnosed with an intellectual disability, a prevalence of 5.3% (OR, 3.31; 95% CI, 1.37-6.93; *P* = .4.6e-03). Together this group had an MEA of 3.79 (95% CI, 3.61-3.97; *P* = .1.4e-03), and 50 (33.3%) did not reach secondary education (OR, 1.47; 95% CI, 1.02-2.1; *P* = .029). These results are in agreement with the observation that the majority of mendelian pathogenic mutations disrupt coding sequences.[47] Prevalence of intellectual disability was best correlated with the presence of at least 1 neurodevelopmental gene with the deleted interval or with a high HiS sum. The group of carriers of deletion encompassing a neurodevelopmental gene has an MEA of 3.76 (95% CI, 3.48-4.05; *P* = .03). Within this group, 6 individuals (8.8%) were diagnosed with intellectual disability (OR, 5.69; 95% CI, 1.97-13.47; *P* = .001). The group of carriers of deletions with the highest quartile of HiS sums had an MEA of 3.91 (95% CI, 3.59-4.23; *P* = .27) with 4 individuals (8.9%) diagnosed with intellectual disability (OR, 5.74; 95% CI, 1.46-16.22; *P* = 7.2e-03). Presence of an ohnolog in the deletion is associated with a higher prevalence of intellectual disability, however to a lesser degree, affecting 6 individuals (5.9%; OR, 3.7; 95% CI, 1.29-8.54; *P* = .008). Neither separately nor together did the numbers of promoters, enhancers, transcriptional elements, and insulators within a CNV correlate with intellectual disability and educational attainment.

Discussion

Although various large pathogenic CNVs are known, the vast majority of rare CNVs of intermediate size (250-500 kb) are thought to be nondeleterious. In the current report we show that the presence of both recurrent syndromic and rare intermediate-size nonrecurrent CNVs, which are cumulatively frequent in the general population (10.5%), are associated with intellectual disability and negatively with educational attainment. For example, the frequency of intellectual disability increases to 4.3% among carriers of 250 kb ≤ deletions <500 kb compared with 1.7% in the Estonian general population. The MEA of carriers decreased from 4.09 to 3.86 with 29.5% not graduating from secondary school compared with 25.3% in the Estonian population. These results are likely to be underestimated through exclusion of the most severely affected patients, inclusion of patients with CNVs known to have no effect on cognition and incorrect inclusion of carriers of large somatic or tumorigenic genomic lesions.

The link between impaired cognitive functioning and lower academic achievement in CNV carriers parallels the recognized correlation between health and education.[48] This health-education gradient was postulated to result from the combination of heritable factors impacting both traits, poor early life health that affects learning, and health-related behaviors being modulated by education. Although recurrent CNVs conferring risk of autism spectrum disorders or schizophrenia were associated with a decrease in IQ of individuals from the general population[49] and phenotype mining of carriers of genomic variants in the Northern Finland 1966 Birth Cohort revealed an excess of lower IQ, school grade retention before age 14 years, and impaired hearing among individuals carrying deletions larger than 500 kb previously implicated in neurodevelopmental disorders,[14] both studies did not recognize that other CNVs, in particular nonrecurrent ones, were also associated with decreases in cognitive capabilities.

Although 40% to 80% of the variance in intelligence and 20% to 40% in educational attainment are explained by genetic factors,[50-53] studies failed to find major contributors to this heritability. For example, 3 individual SNPs each with an approximate effect size of 1 month of schooling per allele have been identified in a genome-wide association study involving more than 126 000 individuals (largest estimated effect, 0.02%)[17] and only a polygenic model including approximately 300 000 common SNPs genome-wide explained 28% to 29% of variation in general cognition.[54] Even though earlier studies failed to identify common CNVs as major contributors to the above heritabilities,[55-58] the results presented herein suggest that rare structural variants of 250 kb or larger for deletions and 1 Mb or larger for duplications are associated with complex traits such as educational attainment and variance in intelligence in population cohorts. About 2% of the analyzed biobank participants carry a rare CNV of 1 Mb or larger. Even without considering other health problems, a fifth of them appear to be linked with decreased quality of life, for the fraction reaching a secondary educational level is 15% lower when comparing CNV carriers to the general population. This reduction results in an MEA that is half a level lower. If we take into account also the carriers of the smaller intermediate-size CNVs associated with lower educational attainment identified in this report (at least 0.2% of the population) and the highly pathogenic anomalies absent from the EGCUT cohort (0.15%), the quality of life for 1 of 40 people might be negatively affected by rare CNVs. These variants may account for a sizable portion of the heritability of the complex "educational attainment" measure.[52]

The observed excess of females carrying rare genomic deletions supports the recently described female-biased mutational burden.[21,59] Females appear "protected" from neurodevelopmental disorders. This potentially allows females to be enrolled in general population cohorts despite the fact that they carry rare CNVs, whereas their male counterparts who likely present more severe phenotypes are excluded from such studies. Consequently and corroboratively, female deletion carriers are responsible for the majority of the signal on educational attainment.

Although intellectual disability prevalence was increased with presence of a neurodevelopmental or ohnolog gene in the deleted interval or a high haploinsufficiency score of imbalanced genes, none of the assessed evaluators correctly capture the variation in educational attainment, possibly because they are limited to protein-coding genes. Investigation of the function of the encompassed protein-coding genes revealed that they were enriched for genes involved in neurogenesis, cognition, learning, memory, and behavior. This is consistent with the hypothesis that these rearrangements are rare because they affect genes important for neurodevelopment and thus are rapidly purged from the population.

Although none of the carriers of known syndromic CNVs identified in the EGCUT cohort were previously diagnosed with a genetic disease, many had major clinical problems (eg, intellectual disability, congenital anomalies, neuropathies, neuropsychiatric disturbances, extreme obesity, and reproductive problems). Because the latter are most likely caused by the newly found genetic alterations, it suggests that these individuals have escaped the attention of the medical genetics system and thus far have not received proper examination and counseling.

We acknowledge several study limitations. Because this is an observational study, no causal inferences can be drawn and confounding bias due to another causal factor could not be excluded. Although caution is required in using educational attainment as a proxy for intellectual function, the confirmatory results obtained with SATs scores and FSIQ in geographically distinct cohorts mitigate this concern. Some of the results show borderline statistical significance, which can be explained by the fact that rare CNVs by definition translate to a small number of carriers. The investigation of the population variance of a complex trait such as educational attainment requires extremely large phenotyped data sets to reach sufficient power.

Conclusions

Known pathogenic CNVs in unselected, but assumed to be healthy, adult populations may be associated with unrecognized clinical sequelae. Additionally, individually rare but collectively common intermediate-size CNVs may be negatively associated with educational attainment. Replication of these findings in additional population groups is warranted given the potential implications of this observation for genomics research, clinical care, and public health.

ARTICLE INFORMATION

Author Affiliations: Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland (Männik, Ferreira, Reymond); Estonian Genome Center, University of Tartu, Tartu (Männik, Mägi, Alavere, Kolk, Reigo, Mihailov, Leitsalu, Nõukas, Metspalu); Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland (Macé, Maillard, Jacquemont, Kutalik); Swiss Institute of Bioinformatics, Lausanne, Switzerland (Macé, Ferreira, Beckmann, Kutalik); Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis (Cole, Pankratz); Bristol Genetic Epidemiology Laboratories, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (Guyatt, Shihab, Gaunt, Timpson); MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (Shihab, Gaunt, Timpson); Department of Neurology and Neurorehabilitation, Children's Clinic, Tartu University Hospital, Tartu, Estonia (Kolk); Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia (Leitsalu, Nõukas, Metspalu); Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany (Teumer); Deparment of Health Sciences, University of Milan, Milan, Italy (Salvi, Cusi); Institute of Biomedical Technologies, Italian National Research Council, Milan, Italy (Cusi); Department of Psychology, University of Minnesota, Minneapolis (McGue, Iacono); Institute of Social and Preventive Medicine, Lausanne University Hospital (CHUV), Lausanne, Switzerland (Kutalik).

Author Contributions: Dr Reymond had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Männik, Reymond.
Acquisition, analysis, or interpretation of data: All the authors.
Drafting of the manuscript: Männik, Reymond.
Critical revision of the manuscript for important intellectual content: All the authors.
Statistical analysis: Männik, Mägi, Macé, Cole, Guyatt, Shihab, Mihailov, Kutalik, Pankratz, Timpson.
Obtained funding: Metspalu, Reymond.
Administrative, technical, or material support: Mihailov, Alavere, Metspalu, Reymond.
Study supervision: Metspalu, Reymond.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Drs Timpson, Gaunt, and Shihab work within the MRC Integrative Epidemiology Unit, supported by the Medical Research Council [MC_UU_12013/1-9]. No other disclosures were reported.

Funding/Support: Dr Männik is a grantee of a scholarship from the Swiss Scientific Exchange New Member State of the European Union Program. Ms Guyatt is funded by a PhD studentship from the Wellcome Trust (grant 102433/Z/13/Z). Dr Jacquemont is a Bursary Professor of the Swiss National Science Foundation (SNSF). This study is supported by 2 SNSF grants (31003A_160203, Drs Reymond, and Kutalik), a specific 16p11.2 SNSF Sinergia grant (CRSII33-133044, Dr Reymond), the Simons Foundation Autism Research Initiative (SFARI274424, Dr Reymond), Leenaards Foundation Prizes (Drs Jacquemont, Reymond, and Kutalik), European Commission Framework Program 7 grants (278913, 306031, and 313010) (Dr Metspalu), Center of Excellence in Genomics (EXCEGEN) and University of Tartu (SP1GVARENG, Dr Metspalu), Estonian Research Council Grant (IUT20-60, Dr Metspalu), US Public Health Service grants from the National Institute on Alcohol Abuse and Alcoholism (AA09367 and AA11886, Dr Pankratz), the National Institute on Drug Abuse (DA05147, DA13240, and DA024417, Dr Pankratz), and the National Institute of Mental Health (MH066140, Dr Pankratz).

Role of the Funder/Sponsor: None of the funders or sponsors had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclosure: This publication is the work of the authors, who will serve as guarantors for the contents of this article.

Additional Data: ALSPAC data were generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. ALSPAC analyses were completed using computational facilities of the Advanced Computing Research Centre, University of Bristol (http://www.bris.ac.uk/acrc).

Additional Contributions: We thank the EGCUT participants, the EGCUT personnel for their compensated assistance in recruiting, phenotyping, routing samples, genotyping, and administrative responsibilities especially those of Mari Nelis, PhD, Lili Milani, PhD, Viljo Soo, Kairit Mikkel, Mari-Liis Tammesoo, MSc, and Steven Smit. We are grateful to Sven Bergmann, PhD, University of Lausanne for comments on the manuscript. We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. We are grateful to Mary Ward for her advice on the use of the ALSPAC education data. The UK Medical Research Council and the Wellcome Trust (grant 102215/2/13/2) and the University of Bristol provided core support for ALSPAC.

REFERENCES

1. Conrad DF, Pinto D, Redon R, et al; Wellcome Trust Case Control Consortium. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704-712.

2. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986-D992.

3. Chaignat E, Yahya-Graison EA, Henrichsen CN, et al. Copy number variation modifies expression time courses. *Genome Res*. 2011;21(1):106-113.

4. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. *Hum Mol Genet*. 2009;18(R1):R1-R8.

5. Henrichsen CN, Vinckenbosch N, Zöllner S, et al. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet*. 2009;41(4):424-429.

6. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848-853.

7. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nat Genet*. 2011;43(9):838-846.

8. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet*. 1998;14(10):417-422.

9. Swaminathan GJ, Bragin E, Chatzimichali EA, et al. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet*. 2012;21 (R1):R37-R44.

10. Leitsalu L, Haller T, Esko T, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu [published ahead of print February 11, 2014]. *Int J Epidemiol*. 2014. doi:10.1093/ije/dyt268.

11. Jacquemont S, Reymond A, Zufferey F, et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*. 2011;478(7367):97-102.

12. Zufferey F, Sherr EH, Beckmann ND, et al; Simons VIP Consortium; 16p11.2 European Consortium. A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *J Med Genet*. 2012;49(10):660-668.

13. Nelis M, Esko T, Mägi R, et al. Genetic structure of Europeans: a view from the North-East. *PLoS One*. 2009;4(5):e5472.

14. Pietiläinen OP, Rehnström K, Jakkula E, et al. Phenotype mining in CNV carriers from a population cohort. *Hum Mol Genet*. 2011;20(13): 2686-2695.

15. Walters RG, Coin LJ, Ruokonen A, et al. Rare genomic structural variants in complex disease:

lessons from the replication of associations with obesity. *PLoS One*. 2013;8(3):e58048.

16. Perry JR, Day F, Elks CE, et al; Australian Ovarian Cancer Study; GENICA Network; ConFab; LifeLines Cohort Study; InterAct Consortium; Early Growth Genetics (EGG) Consortium. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*. 2014;514(7520):92-97.

17. Rietveld CA, Medland SE, Derringer J, et al; LifeLines Cohort Study. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*. 2013;340 (6139):1467-1471.

18. Wood AR, Esko T, Yang J, et al; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46(11):1173-1186.

19. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665-1674.

20. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, Saxena S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res Dev Disabil*. 2011;32(2):419-436.

21. Jacquemont S, Coe BP, Hersch M, et al. A higher mutational burden in females supports a “female protective model” in neurodevelopmental disorders. *Am J Hum Genet*. 2014;94(3):415-425.

22. Krumm N, O’Roak BJ, Karakoc E, et al. Transmission disequilibrium of small CNVs in simplex autism. *Am J Hum Genet*. 2013;93(4):595-606.

23. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*. 2010;6(10): e1001154.

24. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 2010;107(20):9270-9274.

25. Reymond A, Henrichsen CN, Harewood L, Merla G. Side effects of genome structural changes. *Curr Opin Genet Dev*. 2007;17(5):381-386.

26. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473(7345): 43-49.

27. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the “children of the 90s”—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*. 2013;42(1):111-127.

28. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res*. 2006;16(9):1136-1148.

29. Ward ME, McMahon G, St Pourcain B, et al; Social Science Genetic Association Consortium. Genetic variation associated with differential educational attainment in adults has anticipated associations with school performance in children. *PLoS One*. 2014;9(7):e100248.

30. Levǎcić R, Jenkins A, Vignoles A, Steele F, Allen R. Estimating the relationship between school resources and pupil attainment at key stage 3. http://eprints.ioe.ac.uk/1319/1/Levacic2005estimatingfullreport.pdf. 2005. Accessed May 7, 2015.

31. Iacono WG, Carlson SR, Taylor J, Elkins IJ, McGue M. Behavioral disinhibition and the development of substance-use disorders: findings from the Minnesota Twin Family Study. *Dev Psychopathol*. 1999;11(4):869-900.

32. McGue M, Keyes M, Sharma A, et al. The environments of adopted and non-adopted youth: evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behav Genet*. 2007;37(3):449-462.

33. Diskin SJ, Li M, Hou C, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*. 2008;36(19):e126.

34. Sattler JM. *Assessment of Children (Revised).* Philadelphia,k PA: WB Saunders Co; 1974.

35. Salvi E, Kutalik Z, Glorioso N, et al. Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase. *Hypertension*. 2012;59(2):248-255.

36. Dittwald P, Gambin T, Szafranski P, et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res*. 2013;23(9):1395-1409.

37. Kaminsky EB, Kaul V, Paschall J, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med*. 2011;13(9):777-784.

38. Bochukova EG, Huang N, Keogh J, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*. 2010;463(7281):666-670.

39. Walters RG, Jacquemont S, Valsesia A, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*. 2010; 463(7281):671-675.

40. Shinawi M, Liu P, Kang SH, et al. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J Med Genet*. 2010;47(5):332-341.

41. Weiss LA, Shen Y, Korn JM, et al; Autism Consortium. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med*. 2008;358(7):667-675.

42. McCarthy SE, Makarov V, Kirov G, et al; Wellcome Trust Case Control Consortium. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet*. 2009;41(11):1223-1227.

43. Golzio C, Willer J, Talkowski ME, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature*. 2012;485(7398):363-367.

44. Maillard AM, Ruef A, Pizzagalli F, et al; 16p11.2 European Consortium. The 16p11.2 locus modulates brain structures common to autism, schizophrenia, and obesity. *Mol Psychiatry*. 2015;20(1):140-147.

45. Brody N. Intelligence, Schooling, and Society. *Am Psychol*. 1997;52(10):1046-1050.

46. Matarazzo JD, Herman DO. Relationship of Education and IQ in the WAIS-R Standardization Sample. *J Consult Clin Psychol*. 1984;52(4):631-634.

47. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*. 2011;12(11):745-755.

48. Deary IJ. Intelligence. *Annu Rev Psychol*. 2012; 63:453-482.

49. Stefansson H, Meyer-Lindenberg A, Steinberg S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*. 2014;505(7483):361-366.

50. Deary IJ, Penke L, Johnson W. The neuroscience of human intelligence differences. *Nat Rev Neurosci*. 2010;11(3):201-211.

51. Devlin B, Daniels M, Roeder K. The heritability of IQ. *Nature*. 1997;388(6641):468-471.

52. Flint J, Munafò M. Genetics. Herit-ability. *Science*. 2013;340(6139):1416-1417.

53. Vinkhuyzen AA, van der Sluis S, Maes HH, Posthuma D. Reconsidering the heritability of intelligence in adulthood: taking assortative mating and cultural transmission into account. *Behav Genet*. 2012;42(2):187-198.

54. Davies G, Armstrong N, Bis JC, et al; Generation Scotland. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). *Mol Psychiatry*. 2015;20(2): 183-192.

55. Kirkpatrick RM, McGue M, Iacono WG, Miller MB, Basu S, Pankratz N. Low-frequency copy-number variants and general cognitive ability: no evidence of association. *Intelligence*. 2014;42: 98-106.

56. McRae AF, Wright MJ, Hansell NK, Montgomery GW, Martin NG. No association between general cognitive ability and rare copy number variation. *Behav Genet*. 2013;43(3):202-207.

57. Need AC, Attix DK, McEvoy JM, et al. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum Mol Genet*. 2009;18(23):4650-4661.

58. Bagshaw AT, Horwood LJ, Liu Y, Fergusson DM, Sullivan PF, Kennedy MA. No effect of genome-wide copy number variation on measures of intelligence in a New Zealand birth cohort. *PLoS One*. 2013;8(1):e55208.

59. Desachy G, Croen LA, Torres AR, et al. Increased female autosomal burden of rare copy number variants in human populations and in autism families. *Mol Psychiatry*. 2015;20(2):170-175.